Characterizing Gradient Descent Dynamics under Spurious Correlations

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 spurious features correlation affects the learning pace 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 the *bias-aligned* group and the *bias-conflicting* group converges at a rate proportional to $\frac{1}{\varepsilon h T}$, where ε is the proportion of the group, 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.

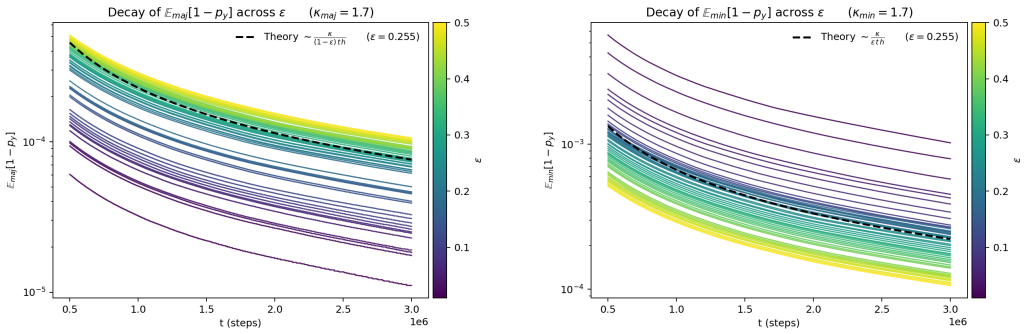

Figure 1: Classification Error of the majority group (left) and the minority group (right) during training and with diverse minority group ratio ε.

1 Introduction

The goal of Deep Learning using Empirical Risk Minimization (ERM) Vapnik (1999) is to tune a neural networks weights so that it can make the closest to ground truth predictions from some input data. This task being purely data-driven, it makes it vulnerable to spurious-correlation. *Spurious-correlation* is a situation where a strong statistical correlation between non-essential features and the ground truth is learned as a causation relationship. As a consequence, a neural network trained on data exhibiting such *spurious-correlation* may perform very well during training and give poor performances at evaluation when tested on data that do not exhibit the same correlation.

Many recent endeavor (Yao et al., 2022; Liu et al., 2021; Kirichenko et al., 2023; Idrissi et al., 2022; Kim et al., 2022; Zheng et al., 2025) have come to the same conclusion that retraining a model

Preprint.

preably trained on a dataset with *spurious-correlation* on a group-wise balanced dataset strongly mitigates the influence of *spurious-correlation*. Intuitively, the first training round consists in learning to extract some descriptive features, while the second round aims at adjusting the model prediction rules. Furthermore, Kirichenko et al. (2023) showed that just retraining the last layer was sufficient to reach state-of-the-art correction. This result motivates our choice on working on a one-layer neural network prepended by a frozen features extractor (Section 4). However, getting a group-wise balanced dataset in real-life scenarios requires being able to identify groups within the training dataset which can be very expensive or even non-feasible. While some work use some proxy as the misclassified datapoints (Liu et al., 2021) or the samples where the loss gradient is large (Kenfack et al., 2025), these methods usually lack strong theoretical evidence regarding their reliability.

In this work we propose a mathematical characterization of the learning speed of each groups. In particular this characterization gives a reliable foundation for group identification strategies such as Just Train Twice (Liu et al., 2021).

2 Problem Setup

Consider a training dataset $\mathcal{D}_{\text{train}}$ consisting of pairs $(\mathbf{x}_n, y_n)_{n=1,\dots|\mathcal{D}_{\text{train}}|}$, where $x_n \in \mathbb{R}^d$. Each input x_n is passed through a frozen feature extractor to obtain $\tilde{x}_n = (r_n, s_n) \in \mathbb{R}^{d_r+s_s}$, which is decomposed into two components: $r_n \in \mathbb{R}^{d_r}$ represents the *relevant features* that truly determine the label, i.e. $y_n = g(r_n)$, and $s_n \in \mathbb{R}^{d_s}$ represents a *spurious signal* which is not predictive of y_n. All pairs (r_n, s_n) are sampled from the joint distribution of the random variables **r** and **s**.

Definition 2.1 (Linear spurious correlation)**.** *A dataset is said to exhibit* spurious features *when* $d_s \geq 1$. *Furthermore, we say it exhibits* linear spurious correlation *when the joint distribution of the relevant features* $\boldsymbol{r}$ *and spurious features* $\boldsymbol{s}$ *in* $\mathcal{D}_{train}$ *satisfies :*

$$\boldsymbol{s} \mid \boldsymbol{r} \sim \begin{cases} A\boldsymbol{r} + \xi, & \textit{with probability } 1-\varepsilon, \\ B\boldsymbol{r} + \xi, & \textit{with probability } \varepsilon, \end{cases} \tag{1}$$

$$\tilde{A}A = \tilde{B}B = I, \tag{2}$$

where $\varepsilon \in [0,1]$. *The mappings* $A, B \in \mathbb{R}^{d_s \times d_r}$ *are some* spurious alignment operator *with left pseudo inverses* $\tilde{A}, \tilde{B} \in \mathbb{R}^{d_r \times d_s}$. *The random variable* ξ *represents a zero-mean additive noise, independent of* y*, with a symmetric conditional distribution given* $\boldsymbol{r}$*:*

$$\xi \perp\!\!\!\perp y, \qquad \textit{Law}(\xi \mid \boldsymbol{r}) = \textit{Law}(-\xi \mid \boldsymbol{r}), \qquad \mathbb{E}[\xi \mid \boldsymbol{r}] = 0. \tag{3}$$

Under this definition, $\mathcal{D}_{\text{train}}$ naturally decomposes into two disjoint groups:

$$\mathcal{G}_1 : \mathbf{s} = A\mathbf{r} + \xi, \qquad \mathcal{G}_2 : \mathbf{s} = B\mathbf{r} + \xi, \qquad \text{s.t.} \quad \varepsilon = \frac{|\mathcal{G}_2|}{|\mathcal{D}_{\text{train}}|}. \tag{4}$$

This partition of $\mathcal{D}_{\text{train}}$ motivates a formal definition of *group imbalance* in our setting.

Definition 2.2 (Group imbalance)**.** *When a dataset exhibits* linear spurious correlation *(see Definition 2.1), it exhibits a* group imbalance *when*

$$\varepsilon \neq 1/2,$$

that is, when one group is more prevalent than the other. The larger group is referred to as the majority group $\mathcal{G}_{maj}$*, and the smaller one as the* minority group $\mathcal{G}_{min}$.

Intuitively, this means that the spurious feature **s** is more often aligned with the relevant feature **r** through either A or B. As a result, the joint structure observed in the majority group dominates the training signal. We finally impose a mild separability condition on the spurious features in $\mathbb{R}^{d_s}$.

Assumption 2.3 (Group separability). *There exist two open sets* $\mathcal{S}_1, \mathcal{S}_2 \subset \mathbb{R}^{d_s}$ *such that*

$$\mathbb{P}\big(\boldsymbol{s} \in \mathcal{S}_k \mid \mathcal{G}_k\big) = 1 \qquad \forall k \in \{1,2\}, \tag{5}$$

$$d(\mathcal{S}_1, \mathcal{S}_2) := \inf_{(s,s') \in \mathcal{S}_1 \times \mathcal{S}_2} \|s - s'\| > 0. \tag{6}$$

This assumption ensures that the two groups are linearly separable *based solely on the spurious features* **s**, even though they remain indistinguishable when considering only the relevant features **r**.

3 Related Works

Group Robustness Metric Group-robustness Liu et al. (2021) refers to the ability of a neural network to have consistent performances across all groups. Given that the global empirical loss cannot fully capture group-robustness Kenfack et al. (2025), many works substitute ERM to the *Worst-Group Accuracy* (WGA) (Sagawa et al., 2020; Liu et al., 2021; Kenfack et al., 2025; Ye et al., 2025)

$$\mathrm{WGA}\big(f(\cdot, w), \mathcal{D}\big) = \max_{g \in \{\text{maj},\text{min}\}} \mathbb{E}_{\mathcal{G}_g}\big[\ell(f(\mathbf{x}, w^t), y)\big],$$

since WGA mitigates the dependancy to the underlying group-distribution in $\mathcal{D}$. However this objective loss is not applicable in real world scenarios where the group membership information is unknown.

Understanding Impact of Spurious Correlation on Learning Speed Very few works present a theoretical analysis of how *spurious-correlation* impacts neural networks learning speed, nor how their methodology improves this learning speed. Yao et al. (2022) shows an extensive analysis of how their method LISA guarantees a lower classification error than the standard ERM method. Idrissi et al. (2022) shows the WGA plot during training phase for various methods. In particular their plot for the ERM method strongly align with our prediction when the number of epochs grow larger than 200, exactly as our graphic displayed on Figure **??**.

4 Main Results

We study how spurious correlated training data affects the learning dynamics and generalization of neural networks. We consider a linear model defined by

$$f(x, w) = w^\top \cdot \tilde{x} = \begin{bmatrix} w_{-1,\mathbf{r}}^\top r + w_{-1,\mathbf{s}}^\top s \\ w_{1,\mathbf{r}}^\top r + w_{1,\mathbf{s}}^\top s \end{bmatrix},$$

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

$$\begin{bmatrix} p_{-1}(\tilde{x}, w) \\ p_1(\tilde{x}, w) \end{bmatrix} = \mathrm{Softmax}\big(f(x, w)\big). \tag{7}$$

We use the binary cross-entropy loss function

$$\ell\big(f(x, w), y\big) = -\log p_y(\tilde{x}, w),$$

and the parameters are updated via full-batch gradient descent with step size h:

$$w^{t+1} = w^t - h\nabla\, \mathcal{L}(\mathcal{D}_{\text{train}}, w^t). \tag{8}$$

We recall this fundamental optimization result

Result adapted from [Soudry et al. (2018, *Theorem 7*)] For almost all binary classification datasets (i.e., except for a measure zero) which are linearly separable (i.e., Eq 10 is feasible), any starting point w^0 and any small enough step size h, the iterates of gradient descent on Eq (8) will behave as:

$$w_k^t = \hat{w}_k \log(t) + \rho_k(t), \qquad \forall k \in \{-1, 1\} \tag{9}$$

where the residual w_{-1} and w_1 denotes the first and second rows of w, $\rho_k(t)$ is bounded and $(\hat{w}_k)_{k=-1,1}$ is the solution of the SVM:

$$\min_{w_{-1}, w_1} \|w_{-1}\|^2 + \|w_1\|^2 \quad \text{s.t.} \quad \forall n, :\ (w_{y_n} - w_{-y_n})^\top \tilde{x}_n \geq 1. \tag{10}$$

We can simply rewrite $(\hat{w}_k)_{k=-1,1}$ as:

$$\hat{w}_1 - \hat{w}_{-1} = \operatorname*{argmin}_{\theta \in \mathbb{R}^{d_r + d_s}} \|\theta\|^2 \quad \text{s.t.} \quad \forall n,\ y_n\, \theta^\top \tilde{x}_n \geq 1. \tag{11}$$

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

Theorem 4.1. *Assume without loss of generality that $\mathcal{G}_{maj} = \mathcal{G}_1$ and $\mathcal{G}_{min} = \mathcal{G}_2$. There exists two positive constants κ_{maj} and κ_{min} independent of ε such that for sufficiently large t,*

$$\mathbb{E}_{\mathcal{G}_{maj}}\big[\,1 - p_y(\boldsymbol{x}, w^t)\,\big] = \frac{\kappa_{maj}}{(1-\varepsilon)\,t\,h} + o\Big(\frac{1}{t\,h}\Big), \tag{12}$$

$$\mathbb{E}_{\mathcal{G}_{min}}\big[\,1 - p_y(\boldsymbol{x}, w^t)\,\big] = \frac{\kappa_{min}}{\varepsilon\,t\,h} + o\Big(\frac{1}{t\,h}\Big). \tag{13}$$

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 Appendix (Section D), we present a detailed proof for a simplified setting of Theorem 4.1.

Corollary 4.2 (Balancing is key)**.** *The optimal minority group proportion ε that maximizes the WGA is $\varepsilon^* = \frac{1}{2}$.*

This result reinforces a well-established principle: a balanced group-wise training distribution minimizes the model's reliance on spurious correlations ().

Proof. From Theorem 4.1, the WGA can be expressed as

$$\mathrm{WGA}(f) = \max_{g\in\{\mathrm{maj},\mathrm{min}\}} \mathbb{E}_{\mathcal{G}_g}\big[\ell(f(\mathbf{x}, w), y)\big] = \max\left(\frac{\kappa_{\mathrm{maj}}}{1-\varepsilon}, \frac{\kappa_{\mathrm{min}}}{\varepsilon}\right).$$

We then see that WGA is optimized when $\varepsilon = 1/2$. □

5 Impact of Spurious Correlation on Generalization Under Distribution Shift

Equations (12) and (13) show that the model learns faster on the majority subgroup $\mathcal{G}_{\mathrm{maj}}$ than on the minority subgroup $\mathcal{G}_{\mathrm{min}}$, since $\varepsilon < 1 - \varepsilon$. This asymmetry implies that features-alignment observed in the majority group dominate the learned representation. To quantify how this imbalance affects generalization, we evaluate the model on a *balanced* test set $\mathcal{D}_{\mathrm{bal}}$, where the *spurious correlation* (Definition 2.1) present in $\mathcal{D}_{\mathrm{train}}$ does not hold. In other words, $\varepsilon_{\mathrm{bal}}$=1/2, i.e $\left|\mathcal{G}^{\mathrm{bal}}_{\mathrm{maj}}\right| = \left|\mathcal{G}^{\mathrm{bal}}_{\mathrm{min}}\right|$. This can be achieved while maintaining the two classes balanced as well:

$$\Big|\{(\mathbf{x}, y) \in \mathcal{D}_{\mathrm{bal}}, y{=}1\}\Big| = \Big|\{(\mathbf{x}, y) \in \mathcal{D}_{\mathrm{bal}}, y{=}{-1}\}\Big|.$$

The generalization risk Sugiyama et al. (2007) on $\mathcal{D}_{\mathrm{bal}}$ is defined as

$$\begin{aligned}\mathcal{L}(\mathcal{D}_{\mathrm{bal}}, w^t) &= \frac{1}{|\mathcal{D}_{\mathrm{bal}}|} \sum_{(\mathbf{x},y)\in\mathcal{D}_{\mathrm{bal}}} \ell\big(f(\mathbf{x}, w^t), y\big) \\ &= \frac{1}{2}\,\mathbb{E}_{\mathcal{G}^{\mathrm{bal}}_{\mathrm{maj}}}\big[-\log p_y(\mathbf{x}, w^t)\big] + \frac{1}{2}\,\mathbb{E}_{\mathcal{G}^{\mathrm{bal}}_{\mathrm{min}}}\big[-\log p_y(\mathbf{x}, w^t)\big].\end{aligned}$$

Using the first-order Taylor expansion $\log(1{-}x) \approx -x$ for small x, we obtain

$$\begin{aligned}\mathcal{L}(\mathcal{D}_{\mathrm{bal}}, w^t) &\approx \frac{1}{2}\,\mathbb{E}_{\mathcal{G}^{\mathrm{bal}}_{\mathrm{maj}}}\big[1 - p_y(\mathbf{x}, w^t)\big] + \frac{1}{2}\,\mathbb{E}_{\mathcal{G}^{\mathrm{bal}}_{\mathrm{min}}}\big[1 - p_y(\mathbf{x}, w^t)\big] \\ &= \frac{\kappa_{\mathrm{maj}}}{2(1-\varepsilon)\,t\,h} + \frac{\kappa_{\mathrm{min}}}{2\varepsilon\,t\,h} + o\Big(\tfrac{1}{t\,h}\Big) \\ &= \frac{1}{2\,t\,h}\left(\frac{\kappa_{\mathrm{maj}}}{1-\varepsilon} + \frac{\kappa_{\mathrm{min}}}{\varepsilon}\right) + o\Big(\tfrac{1}{t\,h}\Big).\end{aligned}$$

The expression above makes explicit how imbalance amplifies generalization error: as ε decreases (i.e., as the dataset becomes more imbalanced), the term $\frac{\kappa_{\mathrm{maj}}}{1-\varepsilon} + \frac{\kappa_{\mathrm{min}}}{\varepsilon}$ grows rapidly, causing a slower

decay of the balanced loss. The above expression also implies that the optimal fraction of minority samples ε satisfies

$$\varepsilon^* = \left(1 + \sqrt{\frac{\kappa_{\text{maj}}}{\kappa_{\text{min}}}}\right)^{-1}.$$

Empirically (see Figure **??**), we observe that $\kappa_{\text{maj}} \approx \kappa_{\text{min}}$, which confirms Corollary 4.2 and yields the optimal value $\varepsilon^* = 1/2$.

6 Experiments

6.1 Setup

6.2 Results

7 Discussion

8 Conclusion

References

Henry Adams, Elin Farnell, and Brittany Story. Support vector machines and radon's theorem. *arXiv preprint arXiv:2011.00617*, 2020.

Patrick Haffner. Escaping the convex hull with extrapolated vector machines. *Advances in Neural Information Processing Systems*, 14, 2001.

Badr Youbi Idrissi, Martin Arjovsky, Mohammad Pezeshki, and David Lopez-Paz. Simple data balancing achieves competitive worst-group-accuracy. In Bernhard Schölkopf, Caroline Uhler, and Kun Zhang (eds.), *Proceedings of the First Conference on Causal Learning and Reasoning*, volume 177 of *Proceedings of Machine Learning Research*, pp. 336–351. PMLR, 11–13 Apr 2022. URL `https://proceedings.mlr.press/v177/idrissi22a.html`.

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*.

Patrik Kenfack, Ulrich Aïvodji, and Samira Ebrahimi Kahou. Gradtune: Last-layer fine-tuning for group robustness without group annotation. In *Workshop on Spurious Correlation and Shortcut Learning: Foundations and Solutions*, 2025.

Nayeong Kim, Sehyun Hwang, Sungsoo Ahn, Jaesik Park, and Suha Kwak. Learning debiased classifier with biased committee. In *ICML 2022: Workshop on Spurious Correlations, Invariance and Stability*, 2022. URL `https://openreview.net/forum?id=bcxmUnTPwny`.

Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. In *The Eleventh International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=Zb6c8A-Fghk`.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks. In *International Conference on Learning Representations*, 2020.

Timothy Sauer. *Numerical analysis*. Pearson, 2018.

Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57, 2018.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert Müller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(5), 2007.

Vladimir N Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

Matt Visser. Primes and the lambert w function. *Mathematics*, 6(4):56, 2018.

Roderick Wong. *Asymptotic approximations of integrals*. SIAM, 2001.

Huaxiu Yao, Yu Wang, Sai Li, Linjun Zhang, Weixin Liang, James Zou, and Chelsea Finn. Improving out-of-distribution robustness via selective augmentation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25407–25437. PMLR, 17–23 Jul 2022. URL `https://proceedings.mlr.press/v162/yao22b.html`.

Wenqian Ye, Guangtao Zheng, and Aidong Zhang. Improving group robustness on spurious correlation via evidential alignment. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 3610–3621, 2025.

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.

Guangtao Zheng, Wenqian Ye, and Aidong Zhang. Shortcutprobe: Probing prediction shortcuts for learning robust models. In James Kwok (ed.), *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pp. 7146–7154. International Joint Conferences on Artificial Intelligence Organization, 8 2025. doi: 10.24963/ijcai.2025/795. URL `https://doi.org/10.24963/ijcai.2025/795`. Main Track.

Appendix

A Table of Contents

B Notations

For any vector $u \in \mathbb{R}^{d_r} \setminus \{0\}$, we denote the orthogonal projections

$$\Pi_u : x \in \mathbb{R}^{d_r} \mapsto \frac{u^\top x}{\|u\|} \frac{u}{\|u\|} \in \text{Span}\{u\} \tag{14}$$

$$\bar{\Pi}_u : x \in \mathbb{R}^{d_r} \mapsto x - \frac{u^\top x}{\|u\|} \frac{u}{\|u\|} \in \text{Span}\{u\}^\perp \tag{15}$$

so that $\Pi_u + \bar{\Pi}_u = I_{d_r}$.

We define the four subgroup partitions:

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

We denote by $\mathbf{v}$ the direction learned by the hard-margin SVM trained on the relevant features:

$$\mathbf{v} = \underset{\theta \in \mathbb{R}^{d_r}}{\text{argmin}} \ \|\theta\|^2 \quad \text{s.t.} \quad \forall n,\ y_n\, \theta^\top \mathbf{r}_n \geq 1. \tag{16}$$

Then, Adams et al. (2020, Section 2.4) shows that

$$d \ := \ \inf \Big\{ \|r - r'\| \ \Big| \ (r, s) \in \mathcal{G}^+,\ (r', s') \in \mathcal{G}^- \Big\} \ = \ \frac{2}{\|\mathbf{v}\|}. \tag{17}$$

where d denotes the twice the SVM margin.

C Relaxed Version of Theorem 4.1

We make the following assumptions.

Assumption C.1 (Density of the projected margin). *Let $Z := y\,\boldsymbol{v}^\top \boldsymbol{r}$ denote the label-corrected margin along the relevant direction $\boldsymbol{v}$. Assume there exists a probability density function λ supported on $[1, \alpha]$, with $\alpha > 1$ and $\lambda(1) > 0$, such that*

$$\mathit{Law}(\boldsymbol{r} \mid g) = \mathit{Law}(\boldsymbol{r} \mid g') \qquad \text{for all } (g, g') \in \{\mathcal{G}^+_{min}, \mathcal{G}^+_{maj}\} \times \{\mathcal{G}^-_{min}, \mathcal{G}^-_{maj}\}, \tag{18}$$

$$p_Z(z \mid g) = \lambda(z)\, \mathbb{1}_{[1,\alpha]}(z) \qquad \text{for all } g \in \{\mathcal{G}^+_{min}, \mathcal{G}^+_{maj}, \mathcal{G}^-_{min}, \mathcal{G}^-_{maj}\}. \tag{19}$$

Assumption C.2 (Initial condition). *Writing the projected parameters*

$$\psi_t := w^t_{-1,\boldsymbol{r}} + \tilde{A} w^t_{-1,\boldsymbol{s}} \in \mathbb{R}^{d_r}, \qquad \phi_t := w^t_{-1,\boldsymbol{r}} + \tilde{B} w^t_{-1,\boldsymbol{s}} \in \mathbb{R}^{d_r}, \tag{20}$$

$$\hat{\psi} := \hat{w}_{-1,\boldsymbol{r}} + \tilde{A} \hat{w}_{-1,\boldsymbol{s}} \in \mathbb{R}^{d_r}, \qquad \hat{\phi} := \hat{w}_{-1,\boldsymbol{r}} + \tilde{B} \hat{w}_{-1,\boldsymbol{s}} \in \mathbb{R}^{d_r}, \tag{21}$$

the model initialization satisfies

$$\boldsymbol{v}^\top \psi_0 \le -1, \qquad \boldsymbol{v}^\top \phi_0 \le -1, \qquad w^0_{-1,\boldsymbol{r}} = -w^0_{1,\boldsymbol{r}}, \qquad w^0_{-1,\boldsymbol{s}} = -w^0_{1,\boldsymbol{s}}.$$

Assumption C.3. *Let $\|\cdot\|$ denote the operator norm on matrices:*

$$\forall M \in \mathbb{R}^{d_s \times d_r}, \qquad \|M\| := \sup_{x \neq 0} \frac{\|Mx\|}{\|x\|}.$$

We assume that at least one of the following holds:

$$\|A\| \le d(\mathcal{S}_1, \mathcal{S}_2)\, \frac{\|\boldsymbol{v}\|}{2} \qquad \textit{or} \qquad \|B\| \le d(\mathcal{S}_1, \mathcal{S}_2)\, \frac{\|\boldsymbol{v}\|}{2}, \tag{22}$$

where $d(\mathcal{S}_1, \mathcal{S}_2)$ is the group-separation distance in the spurious features space defined in Eq. (6)*, and $\boldsymbol{v}$ is the SVM direction introduced in Eq.* (16)*.*

Assumption C.4 (Noise-free case). *For clarity, we analyze the dynamics without perturbation, setting $\xi = 0$.*

Assumption C.1 simply requires that the distribution of the (label-corrected) margin Z be identical across all four subgroups and that its support be bounded by α. In addition, it states that the conditional distribution of the relevant features depends only on the label y and not on the group.

Assumption C.2 is very reasonable as discussed in Remark D.6 and guarantees numerical stability as required in the proof of Proposition D.7.

Assumption C.3 is a mild geometric constraint relating the group-separation scale $d(\mathcal{S}_1, \mathcal{S}_2)$ to the SVM margin $d/2 = 1/\|\mathbf{v}\|$ (Eq (17)).

Assumption C.4 is introduced solely to simplify the computations in the proof.

Theorem C.5. *Assume without loss of generality that $\mathcal{G}_{maj} = \mathcal{G}_1$ and $\mathcal{G}_{min} = \mathcal{G}_2$. For all $t \ge 0$ such that*

$$h \leqslant \eta \frac{\ln t}{t} \frac{\cosh(\alpha)^2}{\epsilon\, \alpha} \qquad \textit{with } 0 < \eta < 1, \tag{23}$$

there exists two positive constants κ_{maj} and κ_{maj} independent on ε such that :

$$\mathbb{E}_{\mathcal{G}_{maj}}\big[\, 1 - p_y(\boldsymbol{x}, w^t) \,\big] = \frac{\kappa_{maj}}{(1-\varepsilon)\, t\, h} + o\Big(\frac{1}{t\, h}\Big), \tag{24}$$

$$\mathbb{E}_{\mathcal{G}_{min}}\big[\, 1 - p_y(\boldsymbol{x}, w^t) \,\big] = \frac{\kappa_{min}}{\varepsilon\, t\, h} + o\Big(\frac{1}{t\, h}\Big), \tag{25}$$

Theorem C.5 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 4.1.

D Proof of Theorem C.5

D.1 Symmetry Properties

Lemma D.1. *We have the following symmetry properties, for all $t \geq 0$:*

(i) $w^t_{-1,\boldsymbol{r}} = -w^t_{1,\boldsymbol{r}}$, (ii) $w^t_{-1,\boldsymbol{s}} = -w^t_{1,\boldsymbol{s}}$

Proof. These follow from the gradient formula:

$$\nabla_w \ell\big(f(\mathbf{x}, w), y\big)^\top = y(1 - p_y(\mathbf{x}, w)) \cdot \Big[\mathbf{r}^\top, \mathbf{s}^\top, -\mathbf{r}^\top, -\mathbf{s}^\top\Big] \tag{26}$$

Summing over $\mathcal{D}_{\text{train}}$ yields: $\nabla_{w_{-1,\mathbf{r}}} \mathcal{L} = -\nabla_{w_{1,\mathbf{r}}} \mathcal{L}$ and $\nabla_{w_{-1,\mathbf{s}}} \mathcal{L} = -\nabla_{w_{1,\mathbf{s}}} \mathcal{L}$ (27)

The result follows by induction using Eq (8). □

D.2 SVM Solution

Lemma D.2. *The SVM solution $\hat{w}$ defined in Eq* (11) *satisfies one of the following equalities:*

$$(\hat{w}_{1,r} - \hat{w}_{-1,r}) + \tilde{A}(\hat{w}_{1,s} - \hat{w}_{-1,s}) = \boldsymbol{v} \tag{28}$$

or

$$(\hat{w}_{1,r} - \hat{w}_{-1,r}) + \tilde{B}(\hat{w}_{1,s} - \hat{w}_{-1,s}) = \boldsymbol{v} \tag{29}$$

where $\boldsymbol{v}$ is defined in Eq (16)*.*

Proof. This results mainly follows from the geometrical characterization of the solution of the SVM problem (11). In fact, as explained in Haffner (2001), $\hat{w}_1 - \hat{w}_{-1}$ is just colinear to $x_+ - x_-$ where the pair $(x_+, x_-) \in \big(\mathcal{D}_{\text{train}} \cap \{y = 1\}\big) \times \big(\mathcal{D}_{\text{train}} \cap \{y = -1\}\big)$ realises the minimum distance between the convex hulls $\text{Conv}\big(\mathcal{D}_{\text{train}} \cap \{y = 1\}\big)$ and $\text{Conv}\big(\mathcal{D}_{\text{train}} \cap \{y = -1\}\big)$.

Now we show that (x_+, x_-) belongs to the same group, i.e $(x_+, x_-) \in \mathcal{G}_{\text{maj}}$ or $(x_+, x_-) \in \mathcal{G}_{\text{min}}$. Consider two samples $x \in \mathcal{G}^+_{\text{maj}}$ and $x' \in \mathcal{G}^-_{\text{min}}$ with their corresponding feature extracted $\tilde{x} = (r, Ar)$ and $\tilde{x}' = (r', Br')$ respectively with $r, r' \in \mathbb{R}^{d_r}$. We have

$$\begin{aligned} \|\tilde{x} - \tilde{x}'\|^2 &= \|r - r'\|^2 + \|Ar - Br'\|^2 \\ &\geq \|r - r'\|^2 + d(\mathcal{S}_1, \mathcal{S}_2)^2 \qquad \big(\text{using Eq (6)}\big) \end{aligned} \tag{30}$$

Now based on Assumption C.3, suppose without loss of generality that $\|A\| \leq d(\mathcal{S}_1, \mathcal{S}_2)\, \frac{\|\mathbf{v}\|}{2}$. On another hand one can notice that except a measure zero category of datasets, there exists $x^\dagger \in \mathcal{G}^+_{\text{maj}}$ with feature extracted $\tilde{x}^\dagger = (r^\dagger, Ar^\dagger)$ and $\|r' - r^\dagger\| < \min\big(\|r - r'\|,\ 1/\|\mathbf{v}\|\big)$. This is simply due to the invariance of $\mathbf{r}$'s distribution with respect to groups (Eq (18)). Then

$$\begin{aligned} \|\tilde{x} - \tilde{x}^\dagger\|^2 &= \|r - r^\dagger\|^2 + \|Ar - Ar^\dagger\|^2 \\ &< \|r - r'\|^2 + \|A\|^2 \|r - r^\dagger\|^2 \\ &\leq \|r - r'\|^2 + \Big(d(\mathcal{S}_1, \mathcal{S}_2)\, \frac{\|\mathbf{v}\|}{2}\Big)^2 \Big(\frac{1}{\|\mathbf{v}\|}\Big)^2 \\ &\leq \|r - r'\|^2 + \frac{d(\mathcal{S}_1, \mathcal{S}_2)^2}{4} \\ &< \|\tilde{x} - \tilde{x}'\|^2 \qquad \text{using Eq (30).} \end{aligned}$$

This shows that for each pair of samples taken from the two groups, there exists a pair within a single group that realise a smaller distance, except measure zero datasets. Therefore, (x_+, x_-) belongs to the same group $\mathcal{G} \in \{\mathcal{G}_{\text{maj}}, \mathcal{G}_{\text{min}}\}$. Without loss of generality, if we assume $\mathcal{G} = \mathcal{G}_{\text{maj}}$, this means that the SVM solution is exactly the same as the one obtained by restricting ourselves to group $\mathcal{G}_{\text{maj}}$:

$$\begin{aligned}
\hat{w}_1 - \hat{w}_{-1} &= \underset{\theta \in \mathbb{R}^{d_r+d_s}}{\operatorname{argmin}} \|\theta\|^2 \quad \text{s.t.} \quad \forall\, \tilde{x}_n \in \mathcal{G}_{\text{maj}},\ y_n\, \theta^\top \tilde{x}_n \geq 1 \\
&= \underset{\theta_r, \theta_s \in \mathbb{R}^{d_r} \times \mathbb{R}^{d_s}}{\operatorname{argmin}} \|\theta_r\|^2 + \|\theta_s\|^2 \quad \text{s.t.} \quad \forall\, \tilde{x}_n \in \mathcal{G}_{\text{maj}},\ y_n\, \theta_r^\top r_n + y_n\, \theta_s^\top s_n \geq 1 \\
&= \underset{\theta_r, \theta_s \in \mathbb{R}^{d_r} \times \mathbb{R}^{d_s}}{\operatorname{argmin}} \|\theta_r\|^2 + \|\theta_s\|^2 \quad \text{s.t.} \quad \forall\, \tilde{x}_n \in \mathcal{G}_{\text{maj}},\ y_n (\theta_r^\top + \theta_s^\top A) r_n \geq 1
\end{aligned}$$

Using the **r**-only SVM solution in Eq (16) we can clearly conclude that □

D.3 Projected Parameter Update Equation

Lemma D.3.

Proof. We present a 5 steps proof. We denote $w_{1,c} = w^t_{1,c}$ for $c \in \{\mathbf{r}, \mathbf{s}\}$ (suppressing the superscript t for readability).

Step 1: Computing parameters updates. Let's rewrite the parameters update

$$\begin{aligned}
\nabla_{w_{1,\mathbf{r}}} \mathcal{L}^t &= \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(\mathbf{x},y) \in \mathcal{D}_{\text{train}}} \nabla_{w_{1,\mathbf{r}}} \ell\big(f(\mathbf{x}, w^t), y\big) \\
&= \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(\mathbf{x},y) \in \mathcal{D}_{\text{train}}} y\big(1 - p_y(\mathbf{x}, w^t)\big)\mathbf{r} \quad \text{(by Eq. (26))}
\end{aligned} \tag{31}$$

Similarly

$$\begin{aligned}
\nabla_{w_{1,\mathbf{s}}} \mathcal{L}^t &= \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(\mathbf{x},y) \in \mathcal{D}_{\text{train}}} \nabla_{w_{1,\mathbf{s}}} \ell\big(f(\mathbf{x}, w^t), y\big) \\
&= \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(\mathbf{x},y) \in \mathcal{D}_{\text{train}}} y\big(1 - p_y(\mathbf{x}, w^t)\big)\mathbf{s}
\end{aligned} \tag{32}$$

Then combining them, we get ψ_t and ϕ_t updates $\big($introduced in Eq (20)$\big)$

$$\begin{aligned}
\psi_{t+1} - \psi_t &= -h\big(\nabla_{w_{1,\mathbf{r}}} \mathcal{L}^t - M^\top \nabla_{w_{1,\mathbf{s}}} \mathcal{L}^t\big) \\
&= \frac{-h}{|\mathcal{D}_{\text{train}}|} \sum_{n=1}^{|\mathcal{D}_{\text{train}}|} y_n\big(1 - p_{y_n}(\mathbf{x}_n, w^t)\big)\big(\mathbf{r}_n - M^\top \mathbf{s}_n\big) \\
&= \frac{-h\,|\mathcal{G}_{\text{maj}}|}{|\mathcal{D}_{\text{train}}|} \frac{1}{|\mathcal{G}_{\text{maj}}|} \sum_{n \in I_{\mathcal{G}_{\text{maj}}}} y_n(1 - p^t_{y_n})\big(\mathbf{r}_n - M^\top M \mathbf{r}_n\big) \\
&\quad \frac{-h\,|\mathcal{G}_{\text{min}}|}{|\mathcal{D}_{\text{train}}|} \frac{1}{|\mathcal{G}_{\text{min}}|} \sum_{n \in I_{\mathcal{G}_{\text{min}}}} y_n(1 - p^t_{y_n})\big(\mathbf{r}_n + M^\top M \mathbf{r}_n\big) \quad \big(\text{using Eq (4)}\big) \\
&= \frac{-2\varepsilon h}{|\mathcal{G}_{\text{min}}|} \sum_{n \in I_{\mathcal{G}_{\text{min}}}} y_n(1 - p^t_{y_n})\mathbf{r}_n \qquad (\text{since } M \text{ is an isometry})
\end{aligned} \tag{33}$$

where $I_g = \Big\{1 \leq n \leq |\mathcal{D}_{\text{train}}| \;\Big|\; (x_n, y_n) \in g\Big\}$ for $g \in \{\mathcal{G}_{\text{maj}}, \mathcal{G}_{\text{min}}\}$. And likewise

$$\phi_{t+1} - \phi_t = \frac{-2(1-\varepsilon)h}{|\mathcal{G}_{\text{maj}}|} \sum_{n \in I_{\mathcal{G}_{\text{maj}}}} y_n(1 - p^t_{y_n})\mathbf{r}_n \tag{34}$$

Step 2: Rewriting $y(1-p_y(\mathbf{x},w))$. On $\mathcal{G}_{\text{min}}$, the first coordinate of the model output expands as:

$$\begin{aligned} w_{1,\mathbf{r}}^\top \mathbf{r} + w_{1,\mathbf{s}}^\top \mathbf{s} &= (\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r} + \bar{\Pi}_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})^\top w_{1,\mathbf{r}} + \big(-M(\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r} + \bar{\Pi}_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})\big)^\top w_{1,\mathbf{s}} \quad \text{(by Eqs. (4)–(??))} \\ &= (\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})^\top \big(w_{1,\mathbf{r}} - M^\top w_{1,\mathbf{s}}\big) + (\bar{\Pi}_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})^\top \big(w_{1,\mathbf{r}} - M^\top w_{1,\mathbf{s}}\big) \\ &= (\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})^\top \psi_t + (\bar{\Pi}_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})^\top \psi_t, \end{aligned}$$

where $\hat{w}_1$ is defined in Eq. (10).

Define the sigmoid function

$$\sigma(x) := \frac{1}{1+e^{-2x}}, \qquad \sigma'(x) = \frac{1}{2\cosh^2(x)} > 0, \tag{35}$$

Then, using Eq. (7) and Lemma D.1, we obtain:

$$\text{On } \mathcal{G}_{\text{min}}^+ : \; y(1-p_y^t) = 1 - p_1^t = \sigma\Big((\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})^\top \psi_t + (\bar{\Pi}_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})^\top \psi_t\Big), \tag{36}$$

$$\text{On } \mathcal{G}_{\text{min}}^- : \; y(1-p_y^t) = -(1-p_{-1}^t) = -\sigma\Big(-(\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})^\top \psi_t - (\bar{\Pi}_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})^\top \psi_t\Big). \tag{37}$$

Following identical reasoning for $\mathcal{G}_{\text{maj}}$, we get:

$$\text{On } \mathcal{G}_{\text{maj}}^+ : \; y(1-p_y^t) = \sigma\Big((\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})^\top \phi_t + (\bar{\Pi}_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})^\top \phi_t\Big), \tag{38}$$

$$\text{On } \mathcal{G}_{\text{maj}}^- : \; y(1-p_y^t) = -\sigma\Big(-(\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})^\top \phi_t - (\bar{\Pi}_{\hat{w}_{1,\mathbf{r}}}\mathbf{r})^\top \phi_t\Big). \tag{39}$$

Now, to show Lemma **??**, we write

$$\begin{aligned} \hat{w}_{1,\mathbf{r}}^\top(\psi_{t+1} - \psi_t) &= \frac{-2\varepsilon h}{|\mathcal{G}_{\text{min}}|} \sum_{n \in I_{\mathcal{G}_{\text{min}}}} y_n(1-p_{y_n}^t)\, \hat{w}_{1,\mathbf{r}}^\top \mathbf{r}_n \qquad \big(\text{see Eq (33)}\big) \\ &= \frac{-2\varepsilon h}{|\mathcal{G}_{\text{min}}|} \sum_{n \in I_{\mathcal{G}_{\text{min}}^+}} \sigma\Big((\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r}_n)^\top \psi_t + (\bar{\Pi}_{\hat{w}_{1,\mathbf{r}}}\mathbf{r}_n)^\top \psi_t\Big)\, \hat{w}_{1,\mathbf{r}}^\top \mathbf{r}_n \\ &+ \frac{2\varepsilon h}{|\mathcal{G}_{\text{min}}|} \sum_{n \in I_{\mathcal{G}_{\text{min}}^-}} \sigma\Big(-(\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r}_n)^\top \psi_t - (\bar{\Pi}_{\hat{w}_{1,\mathbf{r}}}\mathbf{r}_n)^\top \psi_t\Big)\, \hat{w}_{1,\mathbf{r}}^\top \mathbf{r}_n \quad \big(\text{using Eqs (36)-(37)}\big) \\ &= \underbrace{\frac{-2\varepsilon h}{|\mathcal{G}_{\text{min}}|} \sum_{n \in I_{\mathcal{G}_{\text{min}}^+}} \Big(\sigma\Big((\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r}_n)^\top \psi_t + (\bar{\Pi}_{\hat{w}_{1,\mathbf{r}}}\mathbf{r}_n)^\top \psi_t\Big) - \sigma\Big((\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r}_n)^\top \psi_t\Big)\Big)\, \hat{w}_{1,\mathbf{r}}^\top \mathbf{r}_n}_{\text{First term}} \\ &- \frac{2\varepsilon h}{|\mathcal{G}_{\text{min}}|} \sum_{n \in I_{\mathcal{G}_{\text{min}}^+}} \sigma\Big((\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r}_n)^\top \psi_t\Big)\, \hat{w}_{1,\mathbf{r}}^\top \mathbf{r}_n \\ &+ \underbrace{\frac{-2\varepsilon h}{|\mathcal{G}_{\text{min}}|} \sum_{n \in I_{\mathcal{G}_{\text{min}}^-}} \Big(\sigma\Big(-(\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r}_n)^\top \psi_t - (\bar{\Pi}_{\hat{w}_{1,\mathbf{r}}}\mathbf{r}_n)^\top \psi_t\Big) - \sigma\Big(-(\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r}_n)^\top \psi_t\Big)\Big)\, \hat{w}_{1,\mathbf{r}}^\top \mathbf{r}_n}_{\text{Second term}} \\ &+ \frac{2\varepsilon h}{|\mathcal{G}_{\text{min}}|} \sum_{n \in I_{\mathcal{G}_{\text{min}}^-}} \sigma\Big(-(\Pi_{\hat{w}_{1,\mathbf{r}}}\mathbf{r}_n)^\top \psi_t\Big)\, \hat{w}_{1,\mathbf{r}}^\top \mathbf{r}_n \end{aligned} \tag{40}$$

where $I_g = \Big\{1 \le n \le |\mathcal{D}_{\text{train}}| \;\Big|\; (x_n, y_n) \in g\Big\}$ for $g \in \{\mathcal{G}_{\text{min}}^+, \mathcal{G}_{\text{min}}^-\}$. We need to bound the two selected terms.

Step 3: Bounding the first term. We start by making this remark. Eq (10) assures that for $(\mathbf{x}, y) \in \mathcal{G}_{\text{min}}$,

$$
\begin{aligned}
\big(\bar{\Pi}\,\psi_{t+1} - \bar{\Pi}\,\psi_t\big)^\top \bar{\Pi}\,\psi_t &= \big(\bar{\Pi}(\psi_{t+1} - \psi_t)\big)^\top \bar{\Pi}\,\psi_t \\
&= -2\varepsilon h\,\mathbb{E}_{\mathcal{G}_{\text{min}}}\Big[y(1-p_y^t)\big(\bar{\Pi}\,(Z\mathbf{v}+\mathbf{r}_\perp)\big)^\top \bar{\Pi}\,\psi_t\Big] \qquad \big(\text{using Eq (33)}\big) \\
&= -2\varepsilon h \sum_{g\in\{\mathcal{G}_{\text{min}}^+,\,\mathcal{G}_{\text{min}}^-\}} \mathbb{E}_g\Big[y(1-p_y^t)\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t\Big]\, p(g\mid \mathcal{G}_{\text{min}}) \quad \big(\text{since } \bar{\Pi}\,(Z\mathbf{v}+\mathbf{r}_\perp)=\mathbf{r}_\perp\big)
\end{aligned}
\tag{41}
$$

We consider each of the two terms in the above sum separately. Starting with the $\mathcal{G}_{\text{min}}^+$ term,

$$
\begin{aligned}
-2\varepsilon h\,\mathbb{E}_{\mathcal{G}_{\text{min}}^+}\Big[y(1-p_y^t)\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t\Big] &= -2\varepsilon h\,\mathbb{E}_{\mathcal{G}_{\text{min}}^+}\Big[\sigma\big(\mathbf{v}^\top\psi_t\,Z+\mathbf{r}_\perp^\top\psi_t\big)\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t\Big] \qquad \text{(using Eq (36))} \\
&\le -2\varepsilon h\,\mathbb{E}_{\mathcal{G}_{\text{min}}^+}\Big[e^{2\left(\mathbf{v}^\top\psi_t\,Z+\mathbf{r}_\perp^\top\psi_t\right)}\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t \mathbb{1}_{\left\{\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t<0\right\}}\Big] \\
&\quad -2\varepsilon h\,\mathbb{E}_{\mathcal{G}_{\text{min}}^+}\Big[e^{2\left(\mathbf{v}^\top\psi_t\,Z+\mathbf{r}_\perp^\top\psi_t\right)}\big(1-e^{2\left(\mathbf{v}^\top\psi_t\,Z+\mathbf{r}_\perp^\top\psi_t\right)}\big)\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t \mathbb{1}_{\left\{\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t\ge 0\right\}}\Big]
\end{aligned}
$$

where we used the inequality $e^{2x}\big(1-e^{2x}\big) \le \sigma(x) \le e^{2x}$. By rewriting the exponential argument

$$
\begin{aligned}
\mathbf{v}^\top\psi_t\,Z+\mathbf{r}_\perp^\top\psi_t &= \mathbf{v}^\top\Pi\,\psi_t\,Z+\mathbf{r}_\perp^\top\bar{\Pi}\psi_t \\
&= -\big\|\Pi\,\psi_t\big\|Z+\mathbf{r}_\perp^\top\bar{\Pi}\psi_t \qquad \text{(using Eq (50))}
\end{aligned}
\tag{42}
$$

we can inject it in Eq (42) to obtain

$$
\begin{aligned}
-2\varepsilon h\,\mathbb{E}_{\mathcal{G}_{\text{min}}^+}\Big[y(1-p_y^t)\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t\Big] &\le -2\varepsilon h\,\mathbb{E}_{\mathcal{G}_{\text{min}}^+}\Big[e^{-2\|\Pi\,\psi_t\|}e^{2\mathbf{r}_\perp^\top\bar{\Pi}\psi_t}\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t \mathbb{1}_{\left\{\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t<0\right\}}\Big] \\
&\quad -2\varepsilon h\,\mathbb{E}_{\mathcal{G}_{\text{min}}^+}\Big[e^{-2\|\Pi\,\psi_t\|}e^{2\mathbf{r}_\perp^\top\bar{\Pi}\psi_t}\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t \mathbb{1}_{\left\{\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t\ge 0\right\}}\Big] \\
&\quad +2\varepsilon h\,\mathbb{E}_{\mathcal{G}_{\text{min}}^+}\Big[e^{-4\|\Pi\,\psi_t\|}e^{4\mathbf{r}_\perp^\top\bar{\Pi}\psi_t}\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t \mathbb{1}_{\left\{\mathbf{r}_\perp^\top \bar{\Pi}\,\psi_t\ge 0\right\}}\Big] \\
&\le 2\varepsilon h\,e^{-2\|\Pi\,\psi_t\|}+
\end{aligned}
$$

where we used the inequality $\forall x:\ -xe^{2x}<1$ to bound the first expectation using $x=\mathbf{r}_\perp^\top\bar{\Pi}\psi_t$,

Step 4: Bounding the second term. We expand:

$$
\mathbb{E}\big[y(1-p_y^t)\mathbf{r}_\perp\big] = \sum_{g\in\{\mathcal{G}_{\text{maj}}^\pm,\,\mathcal{G}_{\text{min}}^\pm\}} \mathbb{E}_g\big[y(1-p_y^t)\mathbf{r}_\perp\big]\,p(g).
$$

From Eq. (37), we have:

$$
\mathbb{E}_{\mathcal{G}_{\text{min}}^-}\big[y(1-p_y^t)\mathbf{r}_\perp\big] = -\mathbb{E}_{\mathcal{G}_{\text{min}}^-}\big[\sigma(-\mathbf{v}^\top\psi_t\,Z-\mathbf{r}_\perp^\top\psi_t)\,\mathbf{r}_\perp\big] = \mathbb{E}_{\mathcal{G}_{\text{min}}^+}\big[\sigma(\mathbf{v}^\top\psi_t\,Z+\mathbf{r}_\perp^\top\psi_t)\,\mathbf{r}_\perp\big],
$$

where the last equality follows from Assumption **??*(1)***:

$$
\text{Law}\big(-Z\mid\mathcal{G}_{\text{min}}^-\big)=\text{Law}\big(Z\mid\mathcal{G}_{\text{min}}^+\big),\quad \text{Law}\big(-\mathbf{r}_\perp\mid\mathcal{G}_{\text{min}}^-\big)=\text{Law}\big(\mathbf{r}_\perp\mid\mathcal{G}_{\text{min}}^+\big),\quad Z\perp\!\!\!\perp\mathbf{r}_\perp.
$$

Thus,

$$
\sum_{g\in\{\mathcal{G}_{\text{min}}^\pm\}} \mathbb{E}_g\big[y(1-p_y^t)\mathbf{r}_\perp\big]\,p(g)=0, \quad \text{so that} \quad \mathbb{E}_{\mathcal{G}_{\text{min}}}\big[y(1-p_y^t)\mathbf{r}_\perp\big]=0.
$$

Repeating the argument for $\mathcal{G}_{\text{maj}}$ gives:

$$
\mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[y(1-p_y^t)\mathbf{r}_\perp\big]=0, \quad \text{hence} \quad \mathbb{E}\big[y(1-p_y^t)\mathbf{r}_\perp\big]=0. \qquad \square
$$

D.4 Parameter Projections on Orthogonal Subspaces

Proposition D.4. *Define the random variables:*

$$\gamma \overset{d}{:=} \boldsymbol{r}_\perp^\top w_{1,\boldsymbol{r}}^0, \qquad \rho \overset{d}{:=} (M\boldsymbol{r}_\perp)^\top w_{1,\boldsymbol{s}}^0. \tag{43}$$

Then, for all $t \geq 0$,

$$\boldsymbol{r}_\perp^\top w_{1,\boldsymbol{r}}^t = \gamma, \qquad (M\boldsymbol{r}_\perp)^\top w_{1,\boldsymbol{s}}^t = \rho. \tag{44}$$

This proposition shows that the updates to $w_{1,\mathbf{r}}^t$ and $w_{1,\mathbf{s}}^t$ occur along Span$\{\mathbf{v}\}$ and Span$\{M\mathbf{v}\}$, respectively. Hence, the learned representation focuses on the decision boundary orthogonal to $\mathbf{v}$, i.e., Span$\{\mathbf{v}\}^\perp$, which contains $\mathbf{r}_\perp$.

Proof. We expand the parameter updates in two parts.

Update of $w_{1,\mathbf{r}}^t$. We have

$$\begin{aligned}
\nabla_{w_{1,\mathbf{r}}} \mathcal{L}^t &= \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{train}}} \nabla_{w_{1,\mathbf{r}}} \ell\big(f(\mathbf{x}, w^t), y\big) \\
&= \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{train}}} y\big(1 - p_y(\mathbf{x}, w^t)\big)\mathbf{r} \quad \text{(by Eq. (26))} \\
&= \mathbb{E}\Big[y(1-p_y^t)\big(Z\mathbf{v} + \mathbf{r}_\perp\big)\Big] \\
&= \mathbb{E}\big[Zy(1-p_y^t)\big]\mathbf{v} + \mathbb{E}\big[y(1-p_y^t)\mathbf{r}_\perp\big] \\
&= \mathbb{E}\big[Zy(1-p_y^t)\big]\mathbf{v} \quad \text{(by Lemma \textbf{??}).}
\end{aligned} \tag{45}$$

Since $Zy > 0$ almost surely, the update in Eq. (8) implies that $w_{1,\mathbf{r}}^t$ is always updated along Span$\{\mathbf{v}\}$. Hence, $\mathbf{r}_\perp^\top \nabla_{w_{1,\mathbf{r}}} \mathcal{L}^t = 0$, and by induction on t,

$$\mathbf{r}_\perp^\top w_{1,\mathbf{r}}^t = \gamma.$$

Update of $w_{1,\mathbf{s}}^t$. Similarly,

$$\begin{aligned}
\nabla_{w_{1,\mathbf{s}}} \mathcal{L}^t &= \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{(\mathbf{x},y)\in\mathcal{D}_{\text{train}}} y\big(1 - p_y(\mathbf{x}, w^t)\big)\mathbf{s} \\
&= \mathbb{E}\big[y(1-p_y^t)\mathbf{s}\big] \\
&= \mathbb{E}_{\mathcal{G}_{\text{maj}}}\Big[y(1-p_y^t)M\mathbf{r}\Big]p(\mathcal{G}_{\text{maj}}) + \mathbb{E}_{\mathcal{G}_{\text{min}}}\Big[-y(1-p_y^t)M\mathbf{r}\Big]p(\mathcal{G}_{\text{min}}) \quad \text{(from Eq. (4))} \\
&= M\Big((1-\varepsilon)\,\mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[y(1-p_y^t)\mathbf{r}\big] - \varepsilon\,\mathbb{E}_{\mathcal{G}_{\text{min}}}\big[y(1-p_y^t)\mathbf{r}\big]\Big) \\
&= \Big((1-\varepsilon)\,\mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[Zy(1-p_y^t)\big] - \varepsilon\,\mathbb{E}_{\mathcal{G}_{\text{min}}}\big[Zy(1-p_y^t)\big]\Big)M\mathbf{v} \quad \text{(using Lemma \textbf{??}).}
\end{aligned} \tag{46}$$

Thus, $w_{1,\mathbf{s}}^t$ evolves only along Span$\{M\mathbf{v}\}$. Moreover, since $(M\mathbf{r}_\perp)^\top M\mathbf{v} = 0$ because M is an isometry, it follows that

$$(M\mathbf{r}_\perp)^\top \nabla_{w_{1,\mathbf{s}}} \mathcal{L}^t = 0, \quad \text{hence} \quad (M\mathbf{r}_\perp)^\top w_{1,\mathbf{s}}^t = \rho.$$

□

D.5 Learning Dynamics

Proposition D.5. *The parameter gap evolves according to*

$$w_{1,\boldsymbol{r}}^{t+1} - w_{1,\boldsymbol{s}}^{t+1} = w_{1,\boldsymbol{r}}^t - w_{1,\boldsymbol{s}}^t + h\,\mathcal{I}(w_{1,\boldsymbol{r}} - w_{1,\boldsymbol{s}}), \tag{47}$$

where

$$\phi(x) := \frac{1}{1+e^{-2x}}, \qquad \phi'(x) = \frac{1}{2\cosh^2(x)} > 0, \tag{48}$$

$$\mathcal{I}(x) = -2\varepsilon \int_\alpha^\beta z\,\phi(xz)\lambda(z)\,\mathrm{d}z \qquad \big(\textit{with } \lambda \textit{ defined in Assumption } \textbf{??}\textit{-(1)}\big) \tag{49}$$

Proof. We proceed in three steps.

Step 1: Writing the one-step update of $\mathbf{v}^\top\psi_t$. Using Eq (8) and projecting onto $\mathbf{v}$ and $M\mathbf{v}$ gives

$$\mathbf{v}^\top\psi_{t+1} = \mathbf{v}^\top\psi_t - h\Big(\mathbf{v}^\top\nabla_{w_{1,\mathbf{r}}}\mathcal{L}^t - (M\mathbf{v})^\top\nabla_{w_{1,\mathbf{s}}}\mathcal{L}^t\Big) = \mathbf{v}^\top\psi_t - h\,\Delta_t$$

We then expand the gradient difference:

$$\begin{aligned}
\Delta_t &= \mathbf{v}^\top\nabla_{w_{1,\mathbf{r}}}\mathcal{L}^t - (M\mathbf{v})^\top\nabla_{w_{1,\mathbf{s}}}\mathcal{L}^t \\
&= \mathbb{E}\big[Zy(1-p_y^t)\big]\mathbf{v}^\top\mathbf{v} - \Big((1-\varepsilon)\,\mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[Zy(1-p_y^t)\big] - \varepsilon\,\mathbb{E}_{\mathcal{G}_{\text{min}}}\big[Zy(1-p_y^t)\big]\Big)(M\mathbf{v})^\top M\mathbf{v} \quad \big(\text{Eqs (45)-(46)}\big) \\
&= \mathbb{E}\big[Zy(1-p_y^t)\big] - \Big((1-\varepsilon)\,\mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[Zy(1-p_y^t)\big] - \varepsilon\,\mathbb{E}_{\mathcal{G}_{\text{min}}}\big[Zy(1-p_y^t)\big]\Big) \quad (\text{since } M \text{ is an isometry}) \\
&= 2\varepsilon\,\mathbb{E}_{\mathcal{G}_{\text{min}}}\big[Zy(1-p_y^t)\big] \qquad \Big(\text{by expanding} \quad \mathbb{E}\big[Zy(1-p_y^t)\big]\Big) \\
&= 2\varepsilon\Big(\mathbb{E}_{\mathcal{G}_{\text{min}}^+}\big[Zy(1-p_y^t)\big]p(y{=}1\,|\,\mathcal{G}_{\text{min}}) + \mathbb{E}_{\mathcal{G}_{\text{min}}^-}\big[Zy(1-p_y^t)\big]p(y{=}{-}1\,|\,\mathcal{G}_{\text{min}})\Big)
\end{aligned}$$

Since $Zy > 0$ almost surely, the above quantity Δ_t is positive, implying that $\big(\mathbf{v}^\top\psi_t\big)_{t\geq 0}$ decreases over training. Given the initialization $\mathbf{v}^\top\psi_0 \leq -1$ (Assumption **??**-*(2)*), we have:

$$\forall t \geq 0, \quad \mathbf{v}^\top\psi_t \leq -1. \tag{50}$$

Step 2: Computing the expectations. We first notice that

$$\begin{aligned}
\mathbb{E}_{\mathcal{G}_{\text{min}}^-}\big[Zy(1-p_y^t)\big] &= \mathbb{E}_{\mathcal{G}_{\text{min}}^-}\big[Z\,\sigma(-\mathbf{v}^\top\psi_t\,Z - \mathbf{r}_\perp^\top\psi_t)\big] \qquad \big(\text{using Eq (37)}\big) \\
&= -\,\mathbb{E}_{\mathcal{G}_{\text{min}}^+}\big[Z\,\sigma(\mathbf{v}^\top\psi_t\,Z + \mathbf{r}_\perp^\top\psi_t)\big],
\end{aligned}$$

where the last equality follows from Assumption **??***(1)*:

$$\text{Law}\big(-Z \mid \mathcal{G}_{\text{min}}^-\big) = \text{Law}\big(Z \mid \mathcal{G}_{\text{min}}^+\big), \quad \text{Law}\big(-\mathbf{r}_\perp \mid \mathcal{G}_{\text{min}}^-\big) = \text{Law}\big(\mathbf{r}_\perp \mid \mathcal{G}_{\text{min}}^+\big), \quad Z \perp\!\!\!\perp \mathbf{r}_\perp.$$

On $\mathcal{G}_{\text{min}}$, we can rewrite the model's output first coordinate:

$$\begin{aligned}
w_{1,\mathbf{r}}^\top\mathbf{r} + w_{1,\mathbf{s}}^\top\mathbf{s} &= w_{1,\mathbf{r}}^\top\big(Z\mathbf{v} + \mathbf{r}_\perp\big) + w_{1,\mathbf{s}}^\top\big(-M\mathbf{r} + \xi\big) \\
&= \mathbf{v}^\top w_{1,\mathbf{r}}\,Z - (M\mathbf{v})^\top w_{1,\mathbf{s}}\,Z + w_{1,\mathbf{r}}^\top\mathbf{r}_\perp - w_{1,\mathbf{s}}^\top M\mathbf{r}_\perp + w_{1,\mathbf{s}}^\top\xi \\
&= \psi_1^t\,Z + \gamma_t \qquad \big(\text{using Eq (}\textbf{??}\text{)}\big)
\end{aligned}$$

Then, using Eq (7) and Lemma D.1, we get on $\mathcal{G}_{\text{min}}^+$

$$1 - p_1 = 1 - \frac{e^{-w_{1,\mathbf{r}}^\top\mathbf{r} - w_{1,\mathbf{s}}^\top\mathbf{s}}}{e^{w_{1,\mathbf{r}}^\top\mathbf{r} + w_{1,\mathbf{s}}^\top\mathbf{s}} + e^{-w_{1,\mathbf{r}}^\top\mathbf{r} - w_{1,\mathbf{s}}^\top\mathbf{s}}} = \sigma\big(\psi_1^t\,Z + \gamma_t\big)$$

For $1 - p_{-1}$, we similarly have on $\mathcal{G}_{\text{min}}^-$

$$1 - p_{-1} = 1 - \frac{e^{w_{1,\mathbf{r}}^\top\mathbf{r} + w_{1,\mathbf{s}}^\top\mathbf{s}}}{e^{w_{1,\mathbf{r}}^\top\mathbf{r} + w_{1,\mathbf{s}}^\top\mathbf{s}} + e^{-w_{1,\mathbf{r}}^\top\mathbf{r} - w_{1,\mathbf{s}}^\top\mathbf{s}}} = \sigma\big(-\psi_1^t\,Z - \gamma_t\big)$$

Using Equations (**??**)-(**??**), we can then write these expectations as the parametrized integrals:

$$\begin{aligned}
\mathbb{E}_{\mathcal{G}_{\text{min}}^+}\big[(1-p_1)\mathbf{r}\big] &= -\frac{1}{2\varepsilon}\,\mathcal{I}(w_{1,\mathbf{r}} - w_{1,\mathbf{s}}) \\
\mathbb{E}_{\mathcal{G}_{\text{min}}^-}\big[(1-p_{-1})\mathbf{r}\big] &= \frac{1}{2\varepsilon}\,\mathcal{I}(w_{1,\mathbf{r}} - w_{1,\mathbf{s}})
\end{aligned}$$

where $\mathcal{I}$ is defined in Eq (49).

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

$$\nabla_{w_{1,\mathbf{r}}}\mathcal{L}^t - \nabla_{w_{1,\mathbf{s}}}\mathcal{L}^t = -\,\mathcal{I}(w_{1,\mathbf{r}} - w_{1,\mathbf{s}})$$

Using the gradient descent rule Eq. (8), we finally obtain:

$$\begin{aligned} w_{1,\mathbf{r}}^{t+1} - w_{1,\mathbf{s}}^{t+1} &= w_{1,\mathbf{r}}^t - w_{1,\mathbf{s}}^t - h(\nabla_{w_{1,\mathbf{r}}}\mathcal{L}^t - \nabla_{w_{1,\mathbf{s}}}\mathcal{L}^t) \\ &= w_{1,\mathbf{r}}^t - w_{1,\mathbf{s}}^t + h\,\mathcal{I}(w_{1,\mathbf{r}} - w_{1,\mathbf{s}}) \end{aligned}$$

□

Remark D.6. We highlight two useful observations:

(i) The sequence $(w_{1,\mathbf{r}}^t - w_{1,\mathbf{s}}^t)_{t\geq 0}$ is decreasing and remains upper bounded by -1. Moreover, the convergence condition $\nabla_{w_{1,\mathbf{r}}}\mathcal{L}^t - \nabla_{w_{1,\mathbf{s}}}\mathcal{L}^t = 0$ implies that this sequence diverges toward $-\infty$.

(ii) Equation (47) corresponds to the explicit Euler scheme with step size h for the differential equation

$$v'(z) = \mathcal{I}\big(v(z)\big), \qquad v(0) = w_{1,\mathbf{r}}^0 - w_{1,\mathbf{s}}^0 \tag{51}$$

Hence, the sequence $(w_{1,\mathbf{r}}^t - w_{1,\mathbf{s}}^t)_{t\geq 0}$ 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.

D.6 Error Bound Between the Discrete and Continuous Dynamics

Proposition D.7. *Let $(w_{1,\boldsymbol{r}}^t - w_{1,\boldsymbol{s}}^t)$ evolve according to Eq.* (47)*, and let v denote the solution of the continuous ODE* (51)*. Then, for all $t \geq 0$ satisfying condition* (23)*,*

$$w_{1,\boldsymbol{r}}^t - w_{1,\boldsymbol{s}}^t = v(t\,h) + o(1).$$

Proof. We bound the error $|v(t\,h) - (w_{1,\mathbf{r}}^t - w_{1,\mathbf{s}}^t)|$ due to the Euler discretization. By Sauer (2018, *Corollary 6.5*), the global error depends on the Lipschitz constants of $(\mathcal{I})$ and its derivative. Writing

$$\mathcal{I}'(x) = -2\varepsilon \int_\alpha^\beta \frac{xz}{2\cosh(xz)^2}\lambda(z)\,\mathrm{d}z$$

and using that $\alpha \geq 1$, we have that both $|\mathcal{I}|$ and $|\mathcal{I}'|$ are increasing on $]-\infty, -1]$,

$$\sup_{x\leq -1}|\mathcal{I}| = |\mathcal{I}(-1)|, \qquad \sup_{x\leq -1}|\mathcal{I}'| = |\mathcal{I}'(-1)|.$$

Following the notation of Sauer (2018), set

$$L := \sup_{x\leq -1}|\mathcal{I}'|, \qquad M := \sup_{x\leq -1}|\mathcal{I}'\,\mathcal{I}|.$$

Then,

$$L = \epsilon \int_\alpha^\beta \frac{z\lambda(z)}{\cosh(z)^2}\,\mathrm{d}z \leq \frac{\varepsilon\,\alpha}{\cosh(\alpha)^2}\int_\alpha^\beta \lambda(z)\,\mathrm{d}z = \frac{\varepsilon\,\alpha}{\cosh(\alpha)^2}, \tag{52}$$

$$\frac{M}{L^2} = \frac{|\mathcal{I}(-1)|}{|\mathcal{I}'(-1)|} < 0.6, \tag{53}$$

since $\quad \frac{2\cosh(z)^2}{1+e^{2z}} < 0.6 \quad$ for $z \geq \alpha > 1$.

Under condition (23), Eq (52) gives $h \leq \frac{\eta \ln t}{Lt}$, and thus by Sauer (2018, *Corollary 6.5*), for all $z \leq t$:

$$\big|w_{1,\mathbf{r}}^z - w_{1,\mathbf{s}}^z - v(z\,h)\big| \leq \frac{M\,h}{2L}\big(e^{Lzh} - 1\big) \leq \frac{M\,t\,h^2}{2}\,e^{Lth} \leq \frac{M(\eta\ln t)^2}{2L^2\,t^{1-\eta}} \leq 0.3\frac{(\ln t)^2}{t^{1-\eta}},$$

where the last inequality follows from (53).

We conclude the proof by noticing that $\frac{(\ln t)^2}{t^{1-\eta}} = o(1)$ □

Remark D.8 (On the Practical Validity of Condition (23)). The constraint imposed by condition (23) is mild in practice. Using standard parameter values:

$$\alpha = 1, \qquad \varepsilon = 0.1, \qquad \eta = 0.5, \qquad h = 10^{-4},$$

we find that the condition remains valid up to $t \approx 1.71 \times 10^6$, 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 (23).
Indeed, letting T denote the total training horizon and setting

$$h_T = \eta \, \frac{\ln T}{T} \, \frac{\cosh(\alpha)^2}{\varepsilon \, \alpha},$$

we obtain: *(i)* condition (23) holds for all $2 \le t \le T$, and *(ii)* $h_T T \to \infty$ as $T \to +\infty$. Hence, both conditions can be satisfied simultaneously under realistic training regimes.

D.7 Conclusion of the Proof

Proof. We now complete the proof of Theorem C.5.

Using Eq. (48), we have

$$\mathbb{E}_{\mathcal{G}^+_{\text{min}}}\big[(1-p_1)(\mathbf{x}, w^t)\big] = \mathbb{E}_{\mathcal{G}^+_{\text{min}}}\Big[\phi\big((w^t_{1,\mathbf{r}} - w^t_{1,\mathbf{s}})\mathbf{r}\big)\Big] = \mathcal{J}\big(w^t_{1,\mathbf{r}} - w^t_{1,\mathbf{s}}\big),$$

where

$$\mathcal{J}(x) = \int_\alpha^\beta \frac{\lambda(z)}{1 + e^{-2xz}} \, \mathrm{d}z. \tag{54}$$

By Lemma E.1,

$$\mathbb{E}_{\mathcal{G}^+_{\text{min}}}\big[(1-p_1)(\mathbf{x}, w^t)\big] \sim -\frac{\lambda(\alpha)\, e^{2\alpha(w^t_{1,\mathbf{r}} - w^t_{1,\mathbf{s}})}}{2(w^t_{1,\mathbf{r}} - w^t_{1,\mathbf{s}})}.$$

But,
$$\frac{e^{2\alpha(w^t_{1,\mathbf{r}} - w^t_{1,\mathbf{s}})}}{2(w^t_{1,\mathbf{r}} - w^t_{1,\mathbf{s}})} = \frac{e^{2\alpha\, v(th) + o(1)}}{2\, v(th) + o(1)} \qquad \text{(using Proposition D.7)}$$

$$= \frac{e^{-\ln th + \ln\ln th - \ln\left(4\alpha^3\lambda(\alpha)\epsilon\right) + o(1)}}{-\ln th + o(\ln th)} \qquad \text{(using Proposition F.1)}$$

$$= e^{o(1)} \cdot \frac{\big(4\alpha^3\lambda(\alpha)\big)^{-1}}{-\,\epsilon\, t\, h}\big(1 + o(1)\big)$$

Therefore,

$$\mathbb{E}_{\mathcal{G}^+_{\text{min}}}\big[(1-p_1)(\mathbf{x}, w^t)\big] \sim -\frac{\lambda(\alpha)\, e^{2\alpha(w^t_{1,\mathbf{r}} - w^t_{1,\mathbf{s}})}}{2\,(w^t_{1,\mathbf{r}} - w^t_{1,\mathbf{s}})} \sim \frac{(4\alpha^3)^{-1}}{\epsilon\, t\, h}.$$

By model's weights symmetry (Proposition D.5) and features distribution symmetry (Assumption **??**),

$$\mathbb{E}_{\mathcal{G}^-_{\text{min}}}\big[(1-p_{-1})(\mathbf{x}, w^t)\big] = \mathbb{E}_{\mathcal{G}^+_{\text{min}}}\big[(1-p_1)(\mathbf{x}, w^t)\big] \sim \frac{(4\alpha^3)^{-1}}{\epsilon\, t\, h}.$$

Combining the two symmetric subgroups, we obtain:

$$\mathbb{E}_{\mathcal{G}_{\text{min}}}\big[(1-p_y)(\mathbf{x}, w^t)\big] = \mathbb{E}_{\mathcal{G}^+_{\text{min}}}\big[(1-p_1)(\mathbf{x}, w^t)\big]p(y{=}1 \mid \mathcal{G}_{\text{min}}) + \mathbb{E}_{\mathcal{G}^-_{\text{min}}}\big[(1-p_{-1})(\mathbf{x}, w^t)\big]p(y{=}{-1} \mid \mathcal{G}_{\text{min}})$$
$$\sim \frac{(4\alpha^3)^{-1}}{\varepsilon\, t\, h} \qquad \text{as } t\, h \text{ grows large enough.}$$

By symmetry, replacing ε with $(1 - \varepsilon)$ and $w_{1,\mathbf{r}} - w_{1,\mathbf{s}}$ with $w_{1,\mathbf{r}} + w_{1,\mathbf{s}}$ yields the analogous result for the majority group $\mathcal{G}_{\text{maj}}$. □

D.8 Generalizing the proof

E Asymptotic Behavior of $\mathcal{J}$

Lemma E.1. *Let $\mathcal{J}$ be defined as in Eq.* (54)*. Then,*

$$\mathcal{J}(x) \sim -\frac{\lambda(\alpha)\, e^{2\alpha x}}{2x} \qquad \text{as } x \to -\infty.$$

Proof. We proceed in two steps.

Step1: Approximation of $\mathcal{J}(x)$. For $x < 0$, we estimate the deviation between $\mathcal{J}(x)$ and the simplified integral:

$$\left| \mathcal{J}(x) - \int_\alpha^\beta e^{2xz}\lambda(z)\,\mathrm{d}z \right| = \left| \int_\alpha^\beta e^{2xz}\lambda(z)\,\frac{-e^{2xz}}{1+e^{2xz}}\mathrm{d}z \right|$$
$$\leq e^{2x\alpha}\int_\alpha^\beta e^{2xz}\lambda(z)\,\mathrm{d}z = o\left(\int_\alpha^\beta e^{2xz}\lambda(z)\,\mathrm{d}z\right) \quad \text{as } x \to -\infty.$$

Hence, $\mathcal{J}(x) \sim \int_\alpha^\beta e^{2xz}\lambda(z)\,\mathrm{d}z$ as $x \to -\infty$.

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

$$\int_\alpha^\beta e^{2xz}\lambda(z)\,\mathrm{d}z = \int_\alpha^\beta e^{-2uz}\lambda(z)\,\mathrm{d}z,$$

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

$$\int_\alpha^\beta e^{-2uz}\lambda(z)\,\mathrm{d}z \sim \frac{\lambda(\alpha)\, e^{-2\alpha u}}{2u} = -\frac{\lambda(\alpha)\, e^{2\alpha x}}{2x} \qquad \text{as } x \to -\infty.$$

Combining Steps 1 and 2 yields

$$\mathcal{J}(x) \sim -\frac{\lambda(\alpha)\, e^{2\alpha x}}{2x}.$$

□

Corollary E.2. *Let $\mathcal{I}$ be defined as in Eq.* (49)*. Then,*

$$-\mathcal{I}(x) \sim -\frac{\varepsilon\,\alpha\lambda(\alpha)\, e^{2\alpha x}}{x} \qquad \text{as } x \to -\infty.$$

Proof. The proof follows the same reasoning as in Lemma E.1, with the only difference being that λ is replaced by the function $z \mapsto 2\varepsilon\, z\, \lambda(z)$, which satisfies the same regularity conditions stated in Assumption **??**-*(1)*. □

F Asymptotic Behavior of the Solution to the ODE (51)

Proposition F.1. *The solution v of the differential equation* (51) *satisfies*

$$v(z) \sim -\frac{\ln z}{2\alpha} \qquad \text{as } z \to +\infty. \tag{55}$$

More precisely

$$v(z) = -\frac{\ln z}{2\alpha} + \frac{\ln\ln z}{2\alpha} - \frac{\ln\Big(4\alpha^3\lambda(\alpha)\epsilon\Big)}{2\alpha} + o(1) \qquad \textit{as } z \to +\infty. \tag{56}$$

Proof. We solve the Cauchy problem associated with Eq. (51). From $\frac{\mathrm{d}v}{\mathrm{d}z} = \mathcal{I}(v)$, we can write

$$\frac{\mathrm{d}v}{\mathcal{I}(v)} = \mathrm{d}z.$$

Integrating both sides yields

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

where F is any primitive of $\frac{1}{\mathcal{I}}$, and $K = F(w^0_{1,\mathbf{r}} - w^0_{1,\mathbf{s}})$.

Asymptotic behavior of $1/\mathcal{I}$. As $x \to -\infty$, Corollary E.2 ensures

$$\frac{1}{\mathcal{I}(x)} \sim C\, x\, e^{-2\alpha x}, \qquad \text{with } C = \frac{1}{\varepsilon\, \alpha\, \lambda(\alpha)}.$$

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

$$F(x) = \int^x \frac{1}{\mathcal{I}(z)}\,\mathrm{d}z \sim C\int^x z\, e^{-2\alpha z}\mathrm{d}z = -\frac{C\, e^{-2\alpha x}}{2\alpha}\left(x + \frac{1}{2\alpha}\right) \qquad \text{as } x \to -\infty.$$

Asymptotic inversion. Since $v(z) \to -\infty$ (see Remark D.6*(i)*), we have

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

Hence,

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

Using the principal real branch of Lambert W function we get

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

Therefore,

$$v(z) = -\frac{\ln z}{2\alpha} + \frac{\ln\ln z}{2\alpha} - \frac{\ln\Big(4\alpha^3\lambda(\alpha)\epsilon\Big)}{2\alpha} + o(1) \qquad \text{as } z \to +\infty.$$

□