IMPLICIT BIAS OF GRADIENT DESCENT UNDER FEATURE-MEDIATED SPURIOUS CORRELATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We study how *feature-mediated* spurious correlations affect the learning dynamics of gradient descent on linearly separable data. Unlike the widely studied *label-mediated* setting, where the spurious feature depends only on the class label and is conditionally independent of the causal features, feature-mediated correlations entangle spurious and causal signals at the instance level through group-specific operators, making them both harder to detect in practice and more complex to analyse theoretically. We first generalise the implicit-bias result of Soudry et al. (2018) to continuous distributions. Building on this, we derive closed-form group-wise classification error rates. In the general setting, each group's error decays at a rate governed by its hard-margin γ_g. In the *isotropic regime*, we identify a single exponent α that captures the competition between the minority's geometric margin advantage and the coupling induced by the spurious alignment operators. A phase transition occurs at $\alpha = 1$: below it, both groups decay at rate $\kappa_g^t/(\varepsilon_g\, z_t)$ and group proportion directly controls learning speed; above it, the minority escapes the ε-dependence entirely and converges at the faster polynomial rate $\ln(z_t)^{\alpha_\xi - 1}\, z_t^{-\alpha_\xi(1+\delta_{\text{maj}})+\delta_{\text{min}}}$. These results formalise in closed form the widely observed phenomenon that group imbalance implicitly biases optimisation dynamics toward the majority, while revealing that the effect of imbalance can be entirely superseded by geometric margin advantages. Finally we establish a non-vanishing error at test time, under spurious correlation shift. Synthetic experiments confirm the rates, the prefactors and the phase transition, and last-layer retraining on Waterbirds shows the predicted $1/(\varepsilon_g z_t)$ scaling of each group's error on five frozen representations.

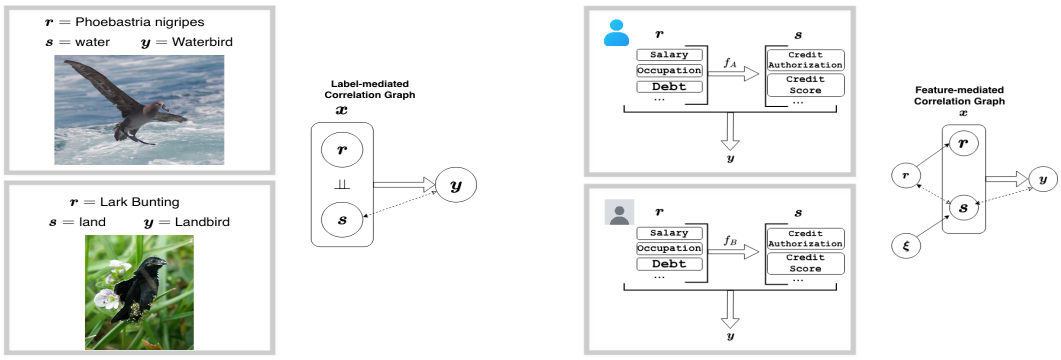

Figure 1: Two structural mechanisms behind spurious correlations. **Left (label-mediated):** the spurious feature **s** (background) depends on the label y but is conditionally independent of the causal features **r** given y and the group. **Right (feature-mediated):** the spurious features **s** (credit score etc.) depend directly on **r** through two group-specific operators f_A, f_B, entangling the two signals.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Modern neural networks trained via Empirical Risk Minimisation (ERM) (Vapnik, 1999) are susceptible to *spurious correlations*: statistical associations between non-causal features and the target label that hold for most but not all of the training distribution. A classifier that relies on such associations performs well on the majority of the data but fails on minority subgroups where the correlation is reversed or absent, a failure mode documented across vision (Sagawa et al., 2020a; Idrissi et al., 2022), language and medical imaging applications.

These correlations can arise through different mechanisms (see Figure 1). In *label-mediated* spurious correlations, the spurious feature depends on the class label y but is conditionally independent of the causal features $\mathbf{r}$ given y and the group, as in the Waterbirds benchmark (Sagawa et al., 2020a) where the background scene correlates with the bird species through the labelling process. Because the conditional independence $\mathbf{s} \perp \mathbf{r} \mid (y,\, g)$ allows the spurious and causal features to be manipulated independently, label-mediated correlations are comparatively straightforward to study, and the vast majority of both empirical benchmarks and theoretical analyses have focused on this case.

In *feature-mediated* spurious correlations, by contrast, the spurious feature $\mathbf{s}$ depends on the *specific value* of $\mathbf{r}$, not merely on y. The causal and spurious signals are entangled at the instance level, making the spurious component harder to isolate experimentally and substantially more complex to analyse theoretically. To our knowledge, no prior work has specifically provided a theoretical analysis of learning dynamics under feature-mediated spurious correlations.

This work. We provide, to the best of our knowledge, the first closed-form characterisation of how features-mediated spurious correlations shape the *per-group* learning speed of gradient descent. Our analysis operates at the population level: we study gradient descent on the expected risk under a neural net learned representation distribution rather than on an empirical average over a finite training set. This yields closed-form dynamics free of finite-sample fluctuations, while remaining predictive of the behaviour observed in practice. Our contributions can be summarized this way:

- We generalise Soudry et al. (2018, *Theorem 9*) to continuous, compactly supported distributions (Proposition 5.4). The iterates still converge in direction to the max-margin classifier $\hat{w}$, but the residual grows as $-\hat{w} \log\log t + O(1)$, in contrast to the $O(1)$ residual of the finite-sample case.
- We investigate how features-mediated spurious correlation affects neural networks learned representation $\Phi(x)$ (Sections 4 and B).
- We derive explicit decay rates for the classification error on each group. In the general case, each group's error decays as $\Theta\big(z_t^{-\gamma_g} (\ln z_t)^{\gamma_g + \delta_g + \bar{\delta}_g - 1}\big)$, governed by its hard-margin γ_g (Theorem 5.1). In the isotropic regime, we obtain sharp asymptotics controlled by a single exponent α that encodes the competition between geometric margins and spurious alignment operators (Theorem 5.3).
- We identify a phase transition at $\alpha = 1$. Below this threshold, the error of every group decays as $\kappa_g / (\varepsilon_g\, z_t)$, so the group proportion ε_g directly scales learning speed. Above it, the group with the larger $\mathbf{r}$-margin escapes ε-dependence entirely and converges at the faster rate $z_t^{-\alpha}$. We confirm the sharpness of this transition on synthetic data (Figure 3), and we show that last-layer retraining on Waterbirds follows the $1/(\varepsilon_g z_t)$ law on five frozen representations (Table 1).
- We prove a generalisation failure result (Corollary G.1): under correlation shift at test time, the classifier may reach a non-vanishing error floor when the test margin $\gamma_{\text{test}} < 0$.

2 PROBLEM SETUP

A classical perspective in statistical learning theory frames nonlinear classification as a two-stage process: (i) a data representation $\Phi : \mathcal{X} \to \mathcal{H}$ lifts inputs into a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, in which (ii) a linear classifier finds a max-margin separator (SVM) (Schölkopf & Smola, 2002; Steinwart & Christmann, 2008; Chen et al., 2021). In this work we consider a data-generating distribution of pairs $(\mathbf{x}, y)$, where $\mathbf{x} \in \mathbb{R}^d$. The data representation $\Phi(\mathbf{x}) = (\mathbf{r}, \mathbf{s}) \in \mathbb{R}^{d_r + d_s}$ satisfies $(\mathbf{r}, \mathbf{s}) \sim \mathcal{D}_{\text{train}}$, where $\mathcal{D}_{\text{train}}$ is a continuous compactly supported distribution and where:

- $\mathbf{r} \in \mathbb{R}^{d_r}$ represents the *causal features*. Following the principle of Invariant Risk Minimization (IRM) (Arjovsky et al., 2019), these features are the invariant data representation that elicits an

unbiased optimal classifier $f(\mathbf{r}) = y$ stable across all groups. Following the same formalism of IRM, $\mathbf{r}$ regroups the parents of y in the Structural Equation Model (SEM) of the problem.

- $\mathbf{s} \in \mathbb{R}^{d_s}$ represents the *spurious features*. These features are predictive of y only through group-specific correlations. Consequently, a predictor based on $\mathbf{s}$ cannot be simultaneously optimal across different groups.

Since our analysis operates at the population level rather than on a finite dataset, we require a notion of max-margin separation that is defined directly over the learned representation distribution:

Definition 2.1 ($\mathbf{r}$-separability)**.** *We say that the relevant features are* $\mathbf{r}$-separable *whenever the following minimum-norm problem admits a solution:*

$$\boldsymbol{v} \;=\; \operatorname*{argmin}_{\theta \in \mathbb{R}^{d_r}} \|\theta\|^2 \quad \textit{s.t.} \quad P_{\mathcal{D}_{train}}\big(y\, \theta^\top \boldsymbol{r} \geq 1\big) = 1. \tag{1}$$

Throughout this work we assume $\|\mathbf{v}\| = 1$ to facilitate the analysis. Similarly, we have:

Definition 2.2 (full-separability)**.** *We say that the full feature representation* $\Phi(\boldsymbol{x}) = (\boldsymbol{r}, \boldsymbol{s})$ *is* $(\mathbf{r}, \mathbf{s})$-separable *whenever the following minimum-norm problem admits a solution:*

$$\hat{w} \;=\; \operatorname*{argmin}_{w \in \mathbb{R}^{d_r + d_s}} \|w\|^2 \quad \textit{s.t.} \quad P_{\mathcal{D}_{train}}\big(y\, w \cdot \Phi(\boldsymbol{x}) \geq 1\big) = 1. \tag{2}$$

We are not aware of prior work that explicitly formulates and uses these population-level minimum-norm separators; we therefore state the definitions here for clarity and assume that $\mathbf{v}$ and $\hat{w}$ exist.

Definition 2.3 (Features-mediated linear spurious correlation)**.** *We say that the feature representation* $\Phi(\boldsymbol{x})$ *exhibits* features-mediated linear spurious correlation *when the conditional law* $\boldsymbol{s} \mid \boldsymbol{r}$ *satisfies:*

$$\boldsymbol{s} \mid \boldsymbol{r} \sim \begin{cases} A\boldsymbol{r} + \xi, & \textit{with probability } 1 - \varepsilon, \\ B\boldsymbol{r} + \xi, & \textit{with probability } \varepsilon, \end{cases} \tag{3}$$

where $\varepsilon \in [0, 1]$. *The mappings* $A, B \in \mathbb{R}^{d_s \times d_r}$ *are* spurious alignment operators *encoding how the spurious signal linearly correlates with the relevant features in each group.* ξ *represents the residual and bounded random noise, dependent only on the* $\boldsymbol{r}$*-margin variable:*

$$\xi \perp\!\!\!\perp (\boldsymbol{r}_\perp, y) \mid \boldsymbol{v} \cdot \boldsymbol{r}, \qquad \textit{with} \quad \boldsymbol{r}_\perp := \boldsymbol{r} - (\boldsymbol{v} \cdot \boldsymbol{r})\boldsymbol{v}. \tag{4}$$

Under this definition, the population naturally decomposes into two disjoint groups:

$$\mathcal{G}_1 : (\mathbf{r}, \mathbf{s}, y) \;\text{ s.t. }\; \mathbf{s} = A\mathbf{r} + \xi, \qquad \mathcal{G}_2 : (\mathbf{r}, \mathbf{s}, y) \;\text{ s.t. }\; \mathbf{s} = B\mathbf{r} + \xi, \qquad \varepsilon = P_{\mathcal{D}_{\text{train}}}(\mathcal{G}_2). \tag{5}$$

The spuriousness of $\mathbf{s}$ is defined by the condition $A \neq B$. This ensures that the optimal linear classifier from $\mathbf{s}$ to y varies across groups, thereby satisfying the definition of Arjovsky et al. (2019). Throughout all this work we use the convention $\mathcal{G}_{\text{maj}} = \mathcal{G}_1$, $\mathcal{G}_{\text{min}} = \mathcal{G}_2$, and $\Phi(\mathbf{x}) \leftarrow \mathbf{x}$ for simplicity.

3 Related Works

Theoretical analyses of spurious correlations. Prior theoretical explanations for why ERM relies on spurious correlations broadly fall into two classes. The first posits that both invariant and spurious features are only *partially* predictive of the label, so that the accuracy-maximising classifier cannot ignore the spurious component (Tsipras et al., 2018; Sagawa et al., 2020b; Arjovsky et al., 2019; Ilyas et al., 2019; Khani & Liang, 2020). The second invokes the *simplicity bias* of gradient-based training (Rahaman et al., 2019; Kalimeris et al., 2019; Shah et al., 2020): even when both features are fully predictive, this model posits that the spurious feature is inherently "simpler" to learn than the invariant one, so that gradient descent preferentially relies on it (Shah et al., 2020; Hermann & Lampinen, 2020). Both frameworks have been developed primarily in the context of *label-mediated* spurious correlations, and therefore structurally differ from our setting.

At the model level, Sagawa et al. (2020b) show that overparameterisation creates a non-vanishing worst-group error floor in a linear model with spurious features. Nagarajan et al. (2020) identify geometric and statistical failure modes of ERM and show that the spurious weight component decays slowly under gradient descent. In the high-dimensional regime, Bombari & Mondelli (2025)

characterises the spurious correlation learned by ridge regression through the covariance spectrum. Closest to our setting, Jain et al. (2024) derive closed-form ODE solutions for the SGD dynamics of a linear classifier on a Gaussian mixture with heterogeneous subpopulations, uncovering a three-phase transient behaviour where variance, representation, and spurious shift dominate at different timescales. Our work differs in loss function (logistic vs. squared), data model (compact support vs. Gaussian), and the nature of the results: we obtain *asymptotic* per-group error exponents and a sharp phase transition at $\alpha = 1$, whereas their analysis characterises the full transient trajectory without producing such exponents. Furthermore, Yang et al. (2024) prove that the spurious contribution grows linearly with correlation strength during early training, and Bachoc et al. (2025) provide lower bounds on the additional training time needed to escape stereotypical predictors that neglect minority features. These works are complementary to our population-level, linear, asymptotic analysis.

Group robustness and mitigation strategies. Group robustness (Sagawa et al., 2020a) measures a model's ability to maintain consistent performance across subgroups. Because the global empirical loss can mask poor minority-group accuracy, it is common to evaluate via the Worst-Group Accuracy (WGA). Distributionally Robust Optimisation (DRO) (Sagawa et al., 2020a) directly optimises WGA but requires group labels at training time. A wide range of methods instead attempt to achieve robustness without group annotations, including data augmentation (Yao et al., 2022), two-stage retraining (Liu et al., 2021; Kirichenko et al., 2023), balanced sampling (Idrissi et al., 2022), committee-based debiasing (Kim et al., 2022), shortcut probing (Zheng et al., 2025), and gradient-based proxy selection (Kenfack et al., 2025). A recurrent empirical finding across these works is that *group balancing*, is a key ingredient. Our theoretical results confirm that group proportion ε_g is indeed a first-order determinant of learning speed in the regime $\alpha < 1$, but reveal that it becomes irrelevant once the margin advantage exceeds the critical threshold ($\alpha \geq 1$).

4 Feature-mediated Spurious Correlations in Practice



Figure 2: Colored-MNIST dataset. **Top row:** majority group (maroon background), digits 0–9. Background intensity increases with digit class. **Bottom row:** minority group (teal background), digits 0–9. Background intensity *decreases* with digit class. The spurious feature (background intensity) depends on the causal feature (foreground image) through the digit class, a nonlinear function of the causal feature. The ground truth label $y = \mathbf{1}_{\text{digit}\geq 5}$, making this a feature-mediated spurious correlation.

While feature-mediated spurious correlation may appear restrictive, it arises naturally in several important settings. The most direct application is the *last-layer retraining* paradigm (Kirichenko et al., 2023), where a feature extractor Φ is frozen and only a linear head is retrained on balanced data. Motivated by the linear representation hypothesis (Park et al., 2023), and by the analysis of feature superposition (Elhage et al., 2022), we experimentally confirmed (Figure 2 and Appendix B) that a correlation (even non linear) mediated by the raw data features becomes approximately linear in the model's learned representation. Beyond learned representations, linear correlation structure appears in sensor fusion (two sensors of the same scene related by a calibration that varies with the conditions), in tabular data (a derived score such as a credit score depends linearly on raw measurements, with coefficients that differ by group or by provider, see Figure 1), and in language model alignment, where response length acts as a linearly correlated proxy for the semantic content (length bias (Park et al., 2024)).

5 MAIN RESULTS

Consider a linear model $f(x, w)$ with $w \in \mathbb{R}^{d_r+d_s}$ trained for a binary classification task: $y \in \{-1, 1\}$, with logistic loss function. We are now ready to state our main results. Our main theoretical results characterizes how features-mediated spurious correlation govern the asymptotic learning pace of each subgroup.

Theorem 5.1 (General Decay Rates). *Assume that $\Phi(\boldsymbol{x}) = (\boldsymbol{r}, \boldsymbol{s})$ satisfies Assumptions I.1–I.4, that the step sizes satisfy* (16) *and $h_t < 2/\beta$ (Proposition I.10), that the infimums in* (20) *are achieved at the boundaries $y\,\boldsymbol{v} \cdot \boldsymbol{r} = \tilde{\gamma}_g$, and that the noise satisfies* (101). *Then:*

$$\mathbb{E}_{\mathcal{G}_{maj}}\big[1 - p_y(\boldsymbol{x}, w^t)\big] = \Theta\left(\frac{(\ln z_t)^{\gamma_{maj}+\delta_{maj}+\tilde{\delta}_{maj}-1}}{z_t^{\gamma_{maj}+o(1)}}\right), \tag{6}$$

$$\mathbb{E}_{\mathcal{G}_{min}}\big[1 - p_y(\boldsymbol{x}, w^t)\big] = \Theta\left(\frac{(\ln z_t)^{\gamma_{min}+\delta_{min}+\tilde{\delta}_{min}-1}}{z_t^{\gamma_{min}+o(1)}}\right), \tag{7}$$

where γ_g are the group-specific hard-margins defined in Eq (20)*, $\tilde{\gamma}_g$ are the group-specific $\boldsymbol{r}$-hard-margins Eq* (21)*, $(\delta_g, \tilde{\delta}_g)$ are defined in Eqs* (113)-(172)*, the $o(1)$ are uniformly noise-dependent, and $z_t := \sum_{k=0}^{t-1} h_k$ is the cumulative learning steps.*

We also focus on this following special spurious correlation regime:

Definition 5.2 (Isotropic Regime). *A spurious correlation setting is isotropic if $\boldsymbol{v}$ is an eigenvector of $A^\top A$, $A^\top B$, $B^\top B$ and $B^\top A$. In plain terms, there exist μ_A, $\mu_B \geq 0$, and μ such that:*

$$A^\top A\boldsymbol{v} = \mu_A \boldsymbol{v}, \qquad A^\top B\boldsymbol{v} = \mu \boldsymbol{v} \tag{8}$$

$$B^\top B\boldsymbol{v} = \mu_B \boldsymbol{v}, \qquad B^\top A\boldsymbol{v} = \mu \boldsymbol{v}. \tag{9}$$

In this regime, we assume

$$-1 \;\leq\; \mu \;<\; \min(\mu_A,\, \mu_B). \tag{10}$$

Concretely, $A^\top A\mathbf{v} = \mu_A \mathbf{v}$ means that $\|A\mathbf{v}\|^2 = \mu_A$ and that $A\mathbf{v}$ is orthogonal to $A\boldsymbol{u}$ for every $\boldsymbol{u} \perp \mathbf{v}$, and the three other conditions read the same way. The isotropic regime therefore only constrains how the *single* direction $\mathbf{v}$ is encoded in the spurious features: $A\mathbf{v}$ and $B\mathbf{v}$ must be orthogonal to the spurious images $A\boldsymbol{u}$, $B\boldsymbol{u}$ of every other direction $\boldsymbol{u} \perp \mathbf{v}$. It does not require A and B to commute. Condition (10) asks that the two groups encode $\mathbf{v}$ differently enough: $A\mathbf{v} \cdot B\mathbf{v}$ is smaller than both $\|A\mathbf{v}\|^2$ and $\|B\mathbf{v}\|^2$ (in particular $A\mathbf{v} \neq B\mathbf{v}$).

We now present our second result.

Theorem 5.3 (Isotropic regime). *Assume the spurious correlation is isotropic (Definition 5.2), that Assumptions D.1 and D.2 hold, that the noise amplitude satisfies* (101)*, and that the step sizes satisfy* (16)*. Without loss of generality, assume $\tilde{\gamma}_{maj} = 1$ and $\tilde{\gamma}_{min} \geq 1$ and define*

$$\alpha := \frac{\tilde{\gamma}_{min}\,(1+\mu)}{1+\mu_A}, \tag{11}$$

where $\tilde{\gamma}_g$ are the group-specific $\boldsymbol{r}$-hard-margins Eq (21)*. Assume finally that on each group the $\boldsymbol{r}$-margin variable $\boldsymbol{v} \cdot \boldsymbol{r}$ has a continuous density, positive at $\tilde{\gamma}_g$ (Eq* (27)*). Then:*
(i) If $\alpha < 1$, there exist two bounded factors $\kappa^t_{maj}(\tilde{\gamma}_{min})$ and $\kappa^t_{min}(\tilde{\gamma}_{min})$ independent of ε and independent of t when the noise $\xi = 0$, such that:

$$\mathbb{E}_{\mathcal{G}_{maj}}\big[1 - p_y^t\big] \underset{t\to\infty}{\sim} \frac{\kappa^t_{maj}(\tilde{\gamma}_{min})}{(1-\varepsilon)\,z_t}, \qquad \mathbb{E}_{\mathcal{G}_{min}}\big[1 - p_y^t\big] \underset{t\to\infty}{\sim} \frac{\kappa^t_{min}(\tilde{\gamma}_{min})}{\varepsilon\, z_t}, \tag{12}$$

(ii) If $\alpha > 1$, the classification error decay at different rates: there exists a positive bounded factor κ^t_{maj} independent on ε and $\tilde{\gamma}_{min}$, and independent on t when the noise $\xi = 0$, such that:

$$\mathbb{E}_{\mathcal{G}_{maj}}\big[1 - p_y^t\big] \underset{t\to\infty}{\sim} \frac{\kappa^t_{maj}}{(1-\varepsilon)\,z_t}, \qquad \mathbb{E}_{\mathcal{G}_{min}}\big[1 - p_y^t\big] = \Theta\left(\frac{(\ln z_t)^{\alpha_\xi - 1}}{z_t^{\alpha_\xi(1+\delta_{maj})-\delta_{min}+o(1)}}\right), \tag{13}$$

with the $o(1)$ uniformly noise-dependent, δ_g defined in Eq (113) noise-dependent, and $\alpha_\xi = \alpha + O(\|\xi\|_\infty) \underset{\|\xi\|\to 0}{\longrightarrow} \alpha$ independently on t.
(iii) Furthermore, these different rate decay are continuous with respect to $\tilde{\gamma}_{min}$, and ξ at the phase transition point $\alpha = 1$.

Note that the part (ii) condition $\alpha{>}1$ is more precisely $\alpha_\xi{>}\frac{1+\delta_{\text{min}}}{1+\delta_{\text{maj}}}{\geq}1$ as we showed in Proposition F.7. Theorem 5.3 naturally extends analogously to the mirror case where $\tilde{\gamma}_{\text{min}}{=}1$ and $\tilde{\gamma}_{\text{maj}} \geq 1$, by substituting $(\mu_A, \tilde{\gamma}_{\text{min}}) \leftarrow (\mu_B, \tilde{\gamma}_{\text{maj}})$. In Appendix E we present a detailed proof of the noise-free version of Theorem 5.3 (Theorem D.4), in Appendix F.2 we add the noise, and in Appendix G.1 we present a proof of Theorem 5.1.

Role of each assumption. Three assumptions carry the phenomenon. *Separability* (Definitions 2.1 and 2.2) is essential: without it gradient descent has no implicit bias towards a max-margin direction. *The linear two-group coupling* with $A \neq B$ (Definition 2.3) is what makes $\mathbf{s}$ spurious and creates the two groups whose rates we compare. *Isotropy* (Definition 5.2) is only needed for the closed form in α: Theorem 5.1 holds without it. The other assumptions are technical. Bounded support (K, R) makes the loss smooth; the continuous, positive margin density at the boundary gives the Laplace (Watson's lemma) expansions behind our rates; Assumptions D.1 and D.2 keep the iterates in $\text{Span}\{\mathbf{v}\}$ and reduce the dynamics to a 2D ODE, and we sketch in Section F.1 how a weaker non-degeneracy condition would only change the prefactors κ_g; the noise bound (101) prevents the noise from closing the margin, and only moves the exponent by $O(\|\xi\|_\infty)$; condition (16) controls the Euler discretisation of the ODE (Lemma H.6).

Our main results rely on the following generalisation of Soudry et al. (2018, *Theorem 9*) for our continuous-distribution setting.

Proposition 5.4 (Implicit bias of GD - continuous-distribution setting)**.** *Under Assumptions I.1-I.4, let w^t denote the iterates of population gradient descent* (207) *with any initial point w^0, and either a constant step size h small enough, or step sizes $(h_t)_t$ satisfying condition* (16) *and $h_t < 2/\beta$ (β defined in Assumption I.2). Then, as $t \to \infty$:*

$$w^t \;=\; \hat{w}\log z_t \;-\; \hat{w}\log\log z_t \;+\; O(1), \tag{14}$$

where $z_t := \sum_{k=0}^{t-1} h_k$ is the cumulative learning steps. In particular, the direction of w^t still converges to the max-margin direction:

$$\lim_{t\to\infty} \frac{w^t}{\|w^t\|} = \frac{\hat{w}}{\|\hat{w}\|}.$$

A proof of this result is provided in Section I with two versions: variable and fixed step-sizes. This result shows that unlike in the finite dataset case, the residual term $\rho^t := w^t - \hat{w}\log t$ is no longer bounded. Section I.4 and Table 3 discuss this difference.

6 Experiments

We test our predictions in two settings. On synthetic data, where α is set by construction, we check the rates, the prefactors and the phase transition of Theorem 5.3. On Waterbirds, we check on the last layer of real networks the prediction that can be tested without knowing the split $(\mathbf{r}, \mathbf{s})$: how each group's error depends on its proportion. All experiments use full-batch gradient descent with logistic loss and a constant step size. Note that our rates are proved for step sizes satisfying (16), which excludes constant steps; we use constant steps because this is what is done in practice, and we still observe the predicted rates. Code is provided in the supplementary material.

Synthetic setup. We sample $(\mathbf{r}, \mathbf{s})$ following Definition 2.3, with operators (A, B) isotropic by construction (Definition 5.2). We fix $\mu_A{=}\mu_B{=}1$, $\mu{=}0$, $d_r{=}d_s{=}8$, $h = 0.01$, and vary $\tilde{\gamma}_{\text{min}}$ through α. All runs last $T = 2 \cdot 10^6$ training steps, and use $N = 10^6$ samples. We used a truncated Gaussian noise according to condition (23).

Asymptotic rates and phase transition. We fix $\varepsilon = 0.1$ and sweep $\alpha \in \{0.5, 0.7, 1.0, 1.3, 1.5\}$. Figure 3 displays the compensated errors for all α values overlaid. The left panel shows the majority error decays at $1/z_t$ rate since all curves plateau at $\kappa_{\text{maj}}(\alpha)/(1-\varepsilon)$. The middle panel shows the minority error decays at $1/z_t^{\max(1,\alpha)}$ rate since all curves plateau, matching Theorem 5.3 predictions. The right panel illustrates the phase transition by plotting the empirical decay exponent β (obtained via log-log regression) against the theoretical α.

Impact of group proportion.

We fix two representative α values (one below and one above the phase transition) and sweep $\varepsilon \in \{0.01, 0.05, 0.1, 0.2, 0.5\}$. Figure 4 shows the rescaled quantity $\varepsilon \cdot \text{err}_{\min} \cdot z_t$ for both α. At $\alpha < 1$ (left panel), all five curves collapse onto the single constant $\kappa_{\min}$, confirming the $\kappa_{\min}/(\varepsilon\, z_t)$ rate decay of the classification error. At $\alpha > 1$ (right panel), all curves tend toward zero together regardless of ε, confirming that the minority rate $\Theta\big(z_t^{-\alpha}(\ln z_t)^{\alpha-1}\big)$ is ε-independent once $\alpha > 1$.

Margin alignment. Aligning Theorems 5.1 and 5.3 predicts $\gamma_{\min} \approx \max(1, \alpha)$ (see the Discussion). We check it directly by solving the hard-margin problem on a smaller synthetic sample ($n = 9{,}000$, $d_r = 5$, $d_s = 8$, $\varepsilon = 0.1$, no noise). With $\alpha = 1.6$ and an **r**-margin ratio $\tilde{\gamma}_{\text{min}}/\tilde{\gamma}_{\text{maj}} = 2.00$, the full-feature margin ratio is $\gamma_{\text{min}}/\gamma_{\text{maj}} = 1.5996$: the spurious coupling shrinks the margin advantage from 2 down to α. With $\alpha = 0.6$ and $\tilde{\gamma}_{\text{min}}/\tilde{\gamma}_{\text{maj}} = 1.05$, the two full-feature margins are equal (1.0000): the coupling erases the advantage entirely.

Last-layer retraining on Waterbirds. We now test on real representations the part of Theorem 5.3 that does not require identifying **r** and **s**: when $\alpha < 1$, each group's error is inversely proportional to its own proportion. We freeze five feature extractors on Waterbirds (Sagawa et al., 2020a): three ResNet-50 trained on Waterbirds with the same hyperparameters as Kirichenko et al. (2023) and three different objectives (ERM, group-reweighted ERM, group DRO), and the self-supervised DINOv2 ViT-B/14 and ViT-L/14 used as they are. On the training features ($n = 4{,}795$, the minority being the 240 images whose background disagrees with the label), we train a linear head with gradient descent ($h = 0.05$) up to $z_t = 10^6$. To vary ε without changing the training set, we reweight the loss: the minority receives a total weight ε and the majority $1 - \varepsilon$, with $\varepsilon \in \{0.01, 0.03, 0.08, 0.2, 0.5\}$. For population gradient descent this is exactly a change of the group proportion, whereas subsampling the minority would also change the maximum margin of the training set. We measure two exponents for each group g: β_g, the slope of $-\log \mathbb{E}_{\mathcal{G}_g}[1 - p_y^t]$ against $\log z_t$, and ν_g, the slope of $-\log \mathbb{E}_{\mathcal{G}_g}[1 - p_y^t]$ against $\log \varepsilon_g$ across the five values of ε at fixed z_t (with $\varepsilon_{\text{min}} = \varepsilon$, $\varepsilon_{\text{maj}} = 1 - \varepsilon$). Theorem 5.3-(i) predicts $\beta_g \to 1$ and $\nu_g \to 1$ for both groups.

Table 1 reports the results. On all five representations, both groups decay with an exponent β_g within 0.025 of 1, and each group's error scales with its inverse proportion: ν_{min} and ν_{maj} lie between 0.95 and 0.98, and the five points of each fit lie on a straight line (residuals at most 0.007). In other words, while ε varies 50-fold, the rescaled error $\varepsilon\, z_t\, \mathbb{E}_{\mathcal{G}_{\text{min}}}[1 - p_y^t]$ varies by at most 17%. All the exponents are still increasing towards 1 at $z_t = 10^6$. These results are consistent with Theorem 5.3-(i), even

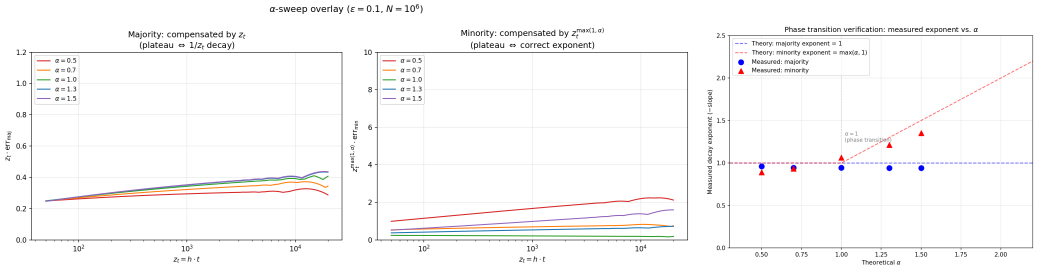

Figure 3: α-sweep overlay ($\varepsilon = 0.1$, $N = 10^6$, $T = 2 \cdot 10^6$). **Left:** majority errors compensated by z_t plateau for all α, confirming the universal $1/z_t$ rate. **Middle:** minority errors compensated by $z_t^{\max(1,\alpha)}$ also plateau, confirming that the theoretical exponent is correct across the full sweep. **Right:** empirical decay exponent β vs. theoretical α. Dashed lines: theoretical predictions (flat at 1 for majority; $\max(\alpha, 1)$ for minority). This provides evidence for the phase transition Theorem 5.3.

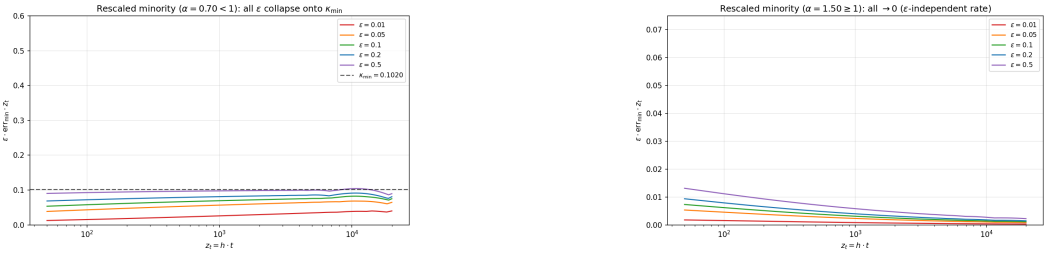

Figure 4: Rescaled minority error $\varepsilon \cdot \mathrm{err}_{\min} \cdot z_t$ for various ε. **Left** ($\alpha = 0.7$): all curves collapse onto the theoretical constant $\kappa_{\min}$ (dashed line), confirming the exact prefactor predicted by Theorem 5.3-(i). **Right** ($\alpha = 1.5$): all curves tend to zero together, confirming ε-independent decay.

Table 1: Last-layer retraining on Waterbirds. ν_g: exponent of the group error in its own proportion ε_g at $z_t = 10^6$ (maximum residual of the 5-point log-log fit in brackets). β_g: exponent of the group error in z_t at $z_t = 3.2 \cdot 10^5$, range over the five values of ε. Theorem 5.3-(i) predicts 1 for all four columns.

Representation	ν_{min}	ν_{maj}	β_{min}	β_{maj}
ResNet-50, ERM	0.978 (0.003)	0.972 (0.004)	0.991–1.002	0.993–0.996
ResNet-50, reweighted	0.961 (0.005)	0.964 (0.005)	0.986–1.006	0.991–0.995
ResNet-50, group DRO	0.973 (0.003)	0.973 (0.004)	0.988–1.002	0.993–0.995
DINOv2 ViT-B/14	0.981 (0.001)	0.977 (0.003)	0.983–0.999	0.989–0.992
DINOv2 ViT-L/14	0.961 (0.002)	0.954 (0.007)	0.977–1.000	0.990–0.993

though the head is trained on a finite training set rather than on the population, and they place the five representations in the regime $\alpha \leq 1$, where balancing is effective.

7 Discussion

Beyond the isotropic regime. Theorem 5.1 does not require isotropy, and it already contains the phase transition in a coarser form. Since $\min(\gamma_{\text{maj}}, \gamma_{\text{min}}) = 1$ (Eq (20)), at least one group always decays with exponent 1, and the other group decays faster if and only if its hard-margin under $\hat{w}$ exceeds 1. What the isotropic regime adds is the closed form of this margin, $\gamma_{\text{min}} \approx \max(1, \alpha)$ (see below), and the exact ε-dependence of the prefactors.

For small departures from isotropy, we show in Section C that any operators (A, B) with $d_s \geq 2d_r$ can be written as an isotropic pair (A', B') plus a residual that only depends on $\mathbf{v} \cdot \mathbf{r}$, which we absorb in the noise ξ. Theorem 5.3 then applies as long as the minimum displacement d^* (Eq (35)) is small enough for (101) to hold, and the exponent only moves by $\alpha_\xi - \alpha = O(\|\xi\|_\infty)$. In the overparameterised setting $d_s \gg d_r$ that is typical of learned representations, this dimensional requirement is easily met.

When is group balancing sufficient? A widely adopted conclusion in the group-robustness literature (Liu et al., 2021; Yao et al., 2022; Idrissi et al., 2022; Kim et al., 2022) is that balancing the group proportions, whether by resampling, reweighting, or data augmentation, is the primary lever for improving worst-group performance. Our isotropic regime results provide a precise theoretical grounding for this claim, and at the same time delineate its limits.

When $\alpha < 1$, the error on each group scales as $\kappa_g^t/(\varepsilon_g\, z_t)$. In this regime, the group proportion ε_g acts as a direct multiplier on the learning speed: a group that is ten times smaller learns ten times slower. Balancing ($\varepsilon = 1/2$) equalises the denominators and is therefore an effective intervention.

However, when $\alpha > 1$, the minority error decays as $\Theta\Big((\ln z_t)^{\alpha_\xi - 1}\, z_t^{-\alpha_\xi(1+\delta_{\text{maj}})+\delta_{\text{min}}+o(1)}\Big)$, a rate that *does not depend on* ε at all: the minority's geometric margin advantage is large enough that it learns faster than the majority regardless of how small it is. In this regime, the majority group is the bottleneck, and its error $\kappa_{\text{maj}}^t/\big((1-\varepsilon)\, z_t\big)$ *increases* as ε moves towards $1/2$: balancing now slows down the group that limits the model, and the useful levers become training time and the majority's

weight. This reveals a range of $\tilde{\gamma}_{\text{min}}$ values, specifically $1 \leq \tilde{\gamma}_{\text{min}} < (1+\mu_A)/(1+\mu)$, where the spurious correlation still influences the learning dynamics through ε, and balancing is a meaningful intervention.

How rare is the regime $\alpha > 1$? Theorem 5.1 gives a simple reading of the escape condition: a group escapes the $1/z_t$ rate only if it puts no mass near the supporting hyperplane of $\hat{w}$ (i.e. $\gamma_g > 1$). The same holds on a finite training set: since $w(t) = \hat{w} \log t + O(1)$ for almost every dataset (Soudry et al., 2018, *Theorem 9*), the loss of a training point x_n decays like $t^{-\hat{w}\cdot x_n}$, so a group decays faster than $1/t$ only if it contains none of the points attaining the maximum margin. When the group variable is unrelated to how hard each example is (a nuisance, such as the background in Waterbirds), this is unlikely: if $|\mathcal{S}|$ of the n training points attain the margin, a random group of m points avoids all of them with probability at most $(1 - |\mathcal{S}|/n)^m$ (very small). The regime $\alpha > 1$ is therefore not impossible, but it requires a group variable that tracks the intrinsic difficulty of the core task (for instance data acquired with a better instrument, or at a different site). Standard group-robustness benchmarks are built the other way around: nuisance group variables and strong spurious correlations.

Interpretation of the main results. *(i) The role of α.* Theorems 5.1 and 5.3 together paint a coherent picture of how spurious correlations govern optimisation dynamics. Theorem 5.1 generalises the classification error analysis of Soudry et al. (2018) to a multi-group population and shows that, in full generality, each group's error decays at a rate controlled by its hard-margin γ_g. Theorem 5.3 then zooms into the isotropic regime and reveals that the *effective* quantity controlling the decay rate is the exponent $\alpha = \tilde{\gamma}_{\text{min}}(1+\mu)/(1+\mu_A)$, not the $\mathbf{r}$-margin $\tilde{\gamma}_{\text{min}}$ alone. The exponent α captures a competition: the minority group's geometric margin advantage ($\tilde{\gamma}_{\text{min}} \geq 1$) is discounted by the coupling ratio $(1+\mu_A)/(1+\mu) \geq 1$, which measures how efficiently the majority's gradient drift propagates to the minority logits. In other words, the spurious features "absorb" part of the minority's margin advantage, and α quantifies exactly how much survives.

Aligning the two theorems in the isotropic regime yields the identity $\gamma_{\text{min}} = \max(1,\, \alpha_\xi(1 + \delta_{\text{maj}}) - \delta_{\text{min}}) \approx \max(1,\, \alpha)$, providing an explicit relationship between the global hard-margin γ_{min} (which determines the coarse rate in Theorem 5.1) and the $\mathbf{r}$-margin $\tilde{\gamma}_{\text{min}}$ modulated by the spurious alignment eigenvalues. The critical threshold $\gamma_{\text{crit}} = (1+\mu_A)/(1+\mu)$ depends on the coupling μ between the spurious alignment operators and the "diagonal" eigenvalue μ_A.

Generalisation failure under correlation shift. Most works on label-mediated spurious correlation observe a generalisation failure of the *worst-group accuracy* when the model is evaluated on minority-group samples underrepresented during training (Kirichenko et al., 2023; Sagawa et al., 2020a; Nagarajan et al., 2020). In our setting, let the test distribution follow $\mathbf{s} \mid \mathbf{r} \sim C\mathbf{r} + \xi$ for an arbitrary spurious alignment operator C. The max-margin classifier then acts on test data through the projected weight $\hat{\chi} := \hat{w}_r + C^\top \hat{w}_s$, and we show in Corollary G.1 (Section G) that if the test margin $\gamma_{\text{test}}(C)$ is positive, the test error still vanishes, at a rate governed by $\gamma_{\text{test}}(C)$, while if $\gamma_{\text{test}}(C) < 0$ it converges to the non-vanishing floor $P_{\mathcal{D}_{\text{test}}}(\hat{\chi} \cdot \mathbf{r} + \hat{w}_s \cdot \xi < 0)$, however long training continues.

8 Conclusion and Limitations

We have presented a mathematical analysis of how features-mediated spurious correlations shape the per-group learning dynamics of gradient descent. At the population level, the group-wise error rates are governed by two quantities: the group proportion ε_g and a composite exponent α that encodes the competition between geometric margins and spurious alignment operators. Below $\alpha = 1$ the group proportion sets the learning speed of each group; above it, the margin geometry supersedes it. Last-layer retraining on Waterbirds with five frozen representations follows the first regime. Beyond training-time dynamics, a correlation shift at test time can create a *non-vanishing error floor*, irrespective of how long training continues (Corollary G.1).

Limitations. First, our analysis is at the population level; a *finite-sample* version, connecting the population rates to empirical risk via uniform concentration bounds, would complete the bridge between theory and practice. Second, our rates are proved for square-summable step sizes, while our experiments use constant steps. Third, we treat *binary* classification with two groups; the *multi-class* generalisation, where the linear relationship takes more than two forms, would significantly broaden the scope of the theory.

POLICY ON LLM USE FOR RESEARCH AND WRITING

We used a large language model (Claude) to implement and run parts of the experimental code, to audit the proofs and point to steps that needed correction or more detail, to give feedback on the experimental methodology, and to edit the text. All theoretical results, their proofs and the final interpretation of the experiments were written and checked by the authors; LLM-generated code was verified and tested by the authors. We take full responsibility for the final content of this work.

REFERENCES

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer, 2005.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

François Bachoc, Jérôme Bolte, Ryan Boustany, and Jean-Michel Loubes. When majority rules, minority loses: bias amplification of gradient descent. *arXiv preprint arXiv:2505.13122*, 2025.

Simone Bombari and Marco Mondelli. Spurious correlations in high dimensional regression: The roles of regularization, simplicity bias and over-parameterization. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 4839–4873. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/bombari25a.html.

Yilan Chen, Wei Huang, Lam Nguyen, and Tsui-Wei Weng. On the equivalence between neural network and support vector machine. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 23478–23490. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/c559da2ba967eb820766939a658022c8-Paper.pdf.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

William Feller et al. *An introduction to probability theory and its applications*. Wiley New York, 1971.

Katherine Hermann and Andrew Lampinen. What shapes feature representations? exploring datasets, architectures, and training. *Advances in Neural Information Processing Systems*, 33:9995–10006, 2020.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

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.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Anchit Jain, Rozhin Nobahari, Aristide Baratin, and Stefano Sarao Mannelli. Bias in motion: Theoretical insights into the dynamics of bias in sgd training. *Advances in Neural Information Processing Systems*, 37:24435–24471, 2024.

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*.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. Sgd on neural networks learns functions of increasing complexity. *Advances in neural information processing systems*, 32, 2019.

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.

Hassan K. Khalil. *Nonlinear Systems*. Prentice Hall, 3rd edition, 2002.

Fereshte Khani and Percy Liang. Feature noise induces loss discrepancy across groups. In *International conference on machine learning*, pp. 5209–5219. PMLR, 2020.

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.

Vaishnavh Nagarajan, Anders Andreassen, and Behnam Neyshabur. Understanding the failure modes of out-of-distribution generalization. *arXiv preprint arXiv:2010.15775*, 2020.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Ryan Park, Rafael Rafailov, Stefano Ermon, and Chelsea Finn. Disentangling length from quality in direct preference optimization. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4998–5017, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.297. URL https://aclanthology.org/2024.findings-acl.297/.

Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pp. 5301–5310. PMLR, 2019.

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

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. An investigation of why overparameterization exacerbates spurious correlations. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 8346–8356. PMLR, 13–18 Jul 2020b. URL https://proceedings.mlr.press/v119/sagawa20a.html.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. *Advances in Neural Information Processing Systems*, 33:9573–9585, 2020.

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.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

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.

Yu Yang, Eric Gan, Gintare Karolina Dziugaite, and Baharan Mirzasoleiman. Identifying spurious biases early in training through the lens of simplicity bias. In Sanjoy Dasgupta, Stephan Mandt, and Yingzhen Li (eds.), *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 2953–2961. PMLR, 02–04 May 2024. URL https://proceedings.mlr.press/v238/yang24c.html.

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.

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

TABLE OF CONTENTS

A Notations **14**

B Features-mediated Spurious Correlation Linearises in Learned Representation **15**

B.1 Experimental Methodology 15

B.2 Results and interpretation 16

C Scope of the Isotropic Regime. **17**

D Relaxed Version of Theorem 5.3 **19**

E Proof of Theorem D.4 **20**

E.1 Auxiliary Results 20

E.2 Discrete and Continuous Error Dynamics 21

E.3 ODE Synchronisation 22

E.4 Conclusion of the Proof when $\alpha < 1$ 26

E.5 Conclusion of the Proof when $\alpha \geq 1$ 26

F Generalizing to the proof of Theorem 5.3 **28**

F.1 Relaxing Assumptions D.1 and D.2 28

F.2 Relaxing Assumption D.3 (feature noise) 28

G Generalisation Failure under Correlation Shift **36**

G.1 Application to the Proof of Theorem 5.1 38

H Toolbox **39**

H.1 Asymptotic Behavior of $\mathcal{J}$ and $\mathcal{I}$ 39

H.2 Asymptotic Behavior of the Solution of 1D ODEs 40

H.3 Euler Scheme Error Bound 43

H.4 Synchronisation Method Lemmas 44

H.5 Discrete Synchronisation 46

I Generalizing Theorem 9 of Soudry et al. **48**

I.1 Setup and assumptions 48

I.2 Proof of Proposition 5.4 49

I.3 Generalisation to variable step-size 53

I.4 Differences between the finite and the continuous settings 53

A NOTATIONS

The model parameters after t training steps are denoted $w^t = [w_r^t \quad w_s^t] \in \mathbb{R}^{d_r+d_s}$.

Denote $p_c(x, w) = \sigma(c\, w \cdot x)$ the predicted probability of input x being in class $c \in \{-1, 1\}$, where σ is the sigmoid function $1/(1 + e^{-u})$.

We use the logistic loss function $\ell\big(f(x, w), y\big) = -\log p_y(x, w)$, and the objective function is $\mathcal{L}(\mathcal{D}_{\text{train}}, w) = \mathbb{E}_{\mathcal{D}_{\text{train}}}\big[\ell(f(\mathbf{x}, w), y)\big]$. The parameters are updated via population gradient descent with step size h_t:

$$w^{t+1} = w^t - h_t \nabla\, \mathcal{L}(\mathcal{D}_{\text{train}}, w^t). \tag{15}$$

We assume that the sequence $(h_t)_{t\geq 0}$ satisfies

$$\sum_{t=0}^{\infty} h_t = +\infty \qquad \text{and} \qquad \sum_{t=0}^{\infty} h_t^2 < \infty, \qquad h_t > 0, \tag{16}$$

and we denote $z_t := \sum_{k=0}^{t-1} h_k$ is the cumulative learning steps.

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

$$\Pi_u : x \in \mathbb{R}^{d_r} \mapsto \frac{u \cdot x}{\|u\|}\, \frac{u}{\|u\|} \in \text{Span}\{u\}, \qquad \bar{\Pi}_u : x \in \mathbb{R}^{d_r} \mapsto x - \frac{u \cdot x}{\|u\|}\, \frac{u}{\|u\|} \in \text{Span}\{u\}^{\perp}, \tag{17}$$

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

$\mathbf{v}$ is the solution of the minimum-norm hard-margin problem:

$$\mathbf{v} \;=\; \underset{\theta \in \mathbb{R}^{d_r}}{\operatorname{argmin}} \|\theta\|^2 \quad \text{s.t.} \quad P_{\mathcal{D}_{\text{train}}}\big(y\, \theta^\top \mathbf{r} \geq 1\big) = 1, \tag{18}$$

and throughout this work we assume $\|\mathbf{v}\| = 1$ to facilitate the analysis.

$\hat{w}$ is the solution of the minimum-norm hard-margin problem:

$$\hat{w} \;=\; \underset{w \in \mathbb{R}^{d_r+d_s}}{\operatorname{argmin}} \|w\|^2 \quad \text{s.t.} \quad P_{\mathcal{D}_{\text{train}}}\big(y\, w \cdot \mathbf{x} \geq 1\big) = 1 \tag{19}$$

Both $\mathbf{v}$ and $\hat{w}$ are assumed to exist in our setting.

Define the group-specific hard-margins for the two groups as:

$$\gamma_{\text{maj}} = \operatorname*{ess\,inf}_{(\mathbf{x},y)|\mathcal{G}_{\text{maj}}} y\, \hat{w} \cdot \mathbf{x}, \qquad \gamma_{\text{min}} = \operatorname*{ess\,inf}_{(\mathbf{x},y)|\mathcal{G}_{\text{min}}} y\, \hat{w} \cdot \mathbf{x}, \qquad \min(\gamma_{\text{maj}},\, \gamma_{\text{min}}) = 1, \tag{20}$$

and the group-specific $\mathbf{r}$-hard-margins for the two groups as

$$\tilde{\gamma}_{\text{maj}} = \operatorname*{ess\,inf}_{(\mathbf{r},y)|\mathcal{G}_{\text{maj}}} y\, \mathbf{v} \cdot \mathbf{r}, \qquad \tilde{\gamma}_{\text{min}} = \operatorname*{ess\,inf}_{(\mathbf{r},y)|\mathcal{G}_{\text{min}}} y\, \mathbf{v} \cdot \mathbf{r}, \qquad \min(\tilde{\gamma}_{\text{maj}},\, \tilde{\gamma}_{\text{min}}) = 1. \tag{21}$$

We work with compactly supported distributions, therefore we define the constants K, and R:

$$\|\mathbf{r}\| \leq K \qquad \text{almost surely,} \tag{22}$$

$$\|\xi\| \leq R \qquad \text{almost surely.} \tag{23}$$

In the isotropic regime Definition 5.2, $\mathbf{v}$ is a common eigenvector for $A^\top A$, $A^\top B$, $B^\top B$ and $B^\top A$. We therefore denote μ_A, μ_B, and μ their corresponding positive eigenvalues:

$$A^\top A\mathbf{v} = \mu_A \mathbf{v}, \qquad A^\top B\mathbf{v} = \mu\mathbf{v} \tag{24}$$

$$B^\top B\mathbf{v} = \mu_B \mathbf{v}, \qquad B^\top A\mathbf{v} = \mu\mathbf{v}. \tag{25}$$

Define the quantity

$$\Sigma := (1+\mu_A)(1+\mu_B) - (1+\mu)^2 \;>\; 0, \tag{26}$$

which is strictly positive under condition 10.

Write $Z := \mathbf{v} \cdot \mathbf{r}$, the $\mathbf{r}$-margin variable. We denote λ_{maj} and λ_{min} the continuous probability density functions, such that $\lambda_{\text{maj}}(\tilde{\gamma}_{\text{maj}}) > 0$, $\lambda_{\text{min}}(\tilde{\gamma}_{\text{min}}) > 0$, and

$$p_Z(z \mid g) = \begin{cases} \lambda_{\text{maj}}(z)\, \mathbb{1}_{[\tilde{\gamma}_{\text{maj}}, K]}(z) & \text{if } g = \mathcal{G}_{\text{maj}}, \\ \lambda_{\text{min}}(z)\, \mathbb{1}_{[\tilde{\gamma}_{\text{min}}, K]}(z) & \text{if } g = \mathcal{G}_{\text{min}}, \end{cases} \tag{27}$$

We absorb the label into the feature: redefine $\mathbf{x} \leftarrow y\mathbf{x}$ so that we may assume $y = +1$ almost surely. In this convention also note $q^t := 1 - p_y^t = 1 - p_1^t$.

B FEATURES-MEDIATED SPURIOUS CORRELATION LINEARISES IN LEARNED REPRESENTATION

A natural question to ask is: what does the feature representation $\Phi(\mathbf{x})$ looks like when the raw-data $\mathbf{x}$ present correlation between the causal and spurious features ? Does the learned representation preserves this correlation? If yes, does that correlation linearises or does it take another form ?

We construct a controlled experiment on a variant of colored MNIST that introduces a nonlinear, feature-mediated spurious correlation by design, train a nonlinear classifier (MLP), and use sparse autoencoders (SAEs) (Huben et al., 2024) to decompose the learned representation into causal and spurious components. The main finding is that even though the input-space correlation is mediated by a highly nonlinear function (digit class, which is a nonlinear function of raw pixels), the representation-space relationship between the causal and spurious SAE features is approximately linear within each group, therefore aligning with the linear feature-mediated model assumed in our theory.

All the codes are available in the supplementary files.

B.1 EXPERIMENTAL METHODOLOGY

Dataset construction. We build on the MNIST dataset ($N = 70{,}000$ images) and introduce a feature-mediated spurious correlation through the background color. Each image is assigned a binary label $y = \mathbf{1}_{\text{digit}\geq 5}$ and, independently of y, is placed into one of two groups:

$$g = \begin{cases} 0 \text{ (majority)} & \text{with probability } 1-\varepsilon, \\ 1 \text{ (minority)} & \text{with probability } \varepsilon, \end{cases} \tag{28}$$

with $\varepsilon = 0.1$. The digit foreground is rendered in grayscale. The background color and intensity are then set as follows:

- *Majority (maroon background):* intensity $= \delta + (1-\delta) \cdot (\text{digit_class}+1)/10$, increasing with digit class.
- *Minority (teal background):* intensity $= \delta + (1-\delta) \cdot (1 - (\text{digit_class}+1)/10)$, decreasing with digit class.

Here $\delta = 0.15$ is a floor that prevents pure-black backgrounds. The intensity is monotonically determined by the digit class, but the digit class is a highly nonlinear function of the raw pixel values $\mathbf{r}$. Figure 2 illustrates the resulting images.

This construction produces a feature-mediated spurious correlation: the spurious feature $\mathbf{s}$ (background intensity) correlates with the label y through the causal feature $\mathbf{r}$ (through the digit class). That correlations takes two forms depending on the group: in the majority, higher digits produce brighter backgrounds; in the minority, higher digits produce darker backgrounds. This mirrors the $A \neq B$ structure of Definition 2.3, but with a nonlinear input-space map because the map from raw pixels to digit class is highly nonlinear.

Model and representation extraction. We train a 4-layer MLP, binary cross-entropy loss, Adam optimiser, learning rate 10^{-3}) and extract the penultimate-layer activations $\phi(x) \in \mathbb{R}^{128}$ at training checkpoints (epochs 1–5). Early stopping stopped the training after 5 epochs, preventing overfitting.

Sparse autoencoder decomposition. To decompose $\phi(x)$ into causal and spurious components, we train a sparse autoencoder (SAE) on the extracted representations at each epoch. Following Huben et al. (2024), we learned a dictionary matrix $M \in \mathbb{R}^{d_h \times 128}$ with $d_h = 512$, four times the feature space dimension of $\phi(x)$ and with an L_1 regularisation coefficient $\alpha = 0.03$. All SAE achieves $> 99.9\%$ variance explained with an average of $\sim$20 active features per sample.

Furthermore, writing

$$c = \mathrm{ReLU}(M\,\phi(x) + b), \qquad \hat{\phi} = M^{\top} c, \tag{29}$$

then c_k represents the feature f_k (whose direction is the $k-$th row of M) activation on instance $\phi(x)$.

Feature classification via group-conditional correlations. Each SAE feature f_k is classified as causal, spurious, or weak/dead using group-conditional correlations with the digit class $\boldsymbol{d}$:

$$\rho_0(k) = \mathrm{corr}(c_k,\, \boldsymbol{d} \mid g{=}0), \qquad \rho_1(k) = \mathrm{corr}(c_k,\, \boldsymbol{d} \mid g{=}1). \tag{30}$$

Within the majority (g=0), background intensity increases with $\boldsymbol{d}$, so both **r**-features and **s**-features have $\rho_0 > 0$. Within the minority (g=1), background intensity *decreases* with $\boldsymbol{d}$, so an **r**-feature (tracking digit shape, invariant to color) retains $\rho_1 > 0$, while an **s**-feature (tracking background color) has $\rho_1 < 0$. This yields the classification rule:

$$f_k \text{ is } \begin{cases} \textbf{r}\text{-type} & \text{if } \rho_0(k)\,\rho_1(k) > 0 \text{ and } \min(|\rho_0(k)|,\, |\rho_1(k)|) \geq \tau, \\ \textbf{s}\text{-type} & \text{if } \rho_0(k)\,\rho_1(k) < 0 \text{ and } \min(|\rho_0(k)|,\, |\rho_1(k)|) \geq \tau, \\ \text{weak} & \text{otherwise}, \end{cases} \tag{31}$$

with threshold $\tau = 0.2$. Figure 6 shows the (ρ_0, ρ_1) scatter for a representative epoch. The **r**-features (blue) cluster along the diagonal $\rho_0 = \rho_1$, while the **s**-features (red) appear in the off-diagonal quadrants ($\rho_0 = -\rho_1$), confirming that the classification cleanly separates the two types.

Linearity of the $\phi_r \to \phi_s$ map. Let $\phi_r = (c_k)_{k\in\mathcal{R}}$ and $\phi_s = (c_k)_{k\in\mathcal{S}}$ denote the subvectors of SAE activations c corresponding to **r**-type and **s**-type features respectively. We fit a Ridge regression $\phi_r \to \phi_s$ within each group and measure R^2 (variance-weighted, multi-output). Table 2 reports the results across all five checkpoints.

Null-distribution control. To verify that the R^2 values reflect genuine structure rather than dimensional artefacts, we perform a permutation test: the rows of ϕ_s are shuffled (as a block, preserving the internal correlation structure among **s**-features) within each group, and the Ridge regression is refit. This is repeated 1000 times to obtain a null baseline. The null R^2 values are three to four orders of magnitude below the observed values (Table 2), with all empirical p-values below 0.001.

Table 2: SAE decomposition and linearity across training epochs ($\alpha_{\mathrm{SAE}} = 0.03$, $\tau = 0.2$). n_r, n_s: number of classified **r**- and **s**-features. R^2_{maj}, R^2_{min}: coefficient of determination of the Ridge regression $\phi_r \to \phi_s$ within each group. R^2_{null}: mean R^2 under 1000 random permutations of ϕ_s (within-group), establishing the null baseline.

Epoch	n_r	n_s	R^2_{maj}	$R^2_{\text{null, maj}}$	R^2_{min}	$R^2_{\text{null, min}}$	Avg. active
1	8	1	0.669	0.0001	0.417	0.001	16.4
2	19	13	0.792	0.0002	0.977	0.002	26.7
3	8	5	0.415	0.0002	0.998	0.001	20.2
4	16	6	0.382	0.0005	0.998	0.001	20.7
5	14	14	0.777	0.0003	0.997	0.001	21.8

B.2 Results and interpretation

Minority group: near-perfect linearity. From epoch 2 onward, $R^2_{\text{min}} \geq 0.977$, indicating that the linear map $\phi_r \to \phi_s$ explains essentially all the variance of the spurious features within the minority group. Furthermore we observe that even though the input-space relationship between digit identity and background intensity is highly nonlinear, the MLP's learned representation linearises this relationship almost perfectly.

Majority group: weaker but significant coupling. The majority R^2_{maj} is lower and more variable across epochs (0.38–0.79), but remains orders of magnitude above the null baseline. This still supports the assumption of correlation linearisation in the learned representation feature space.

Figure 5 displays the first principal component of ϕ_r against the first principal component of ϕ_s for all five epochs, with points colored by group. Two consistent patterns are visible across all epochs: (i) The *minority* (teal) points form a tight, approximately linear cloud with a clear slope. (ii) The *majority* (maroon) points are concentrated near $\phi_s \approx 0$ (flat), indicating that the **s**-features carry little additional information beyond what is already captured by the **r**-features.

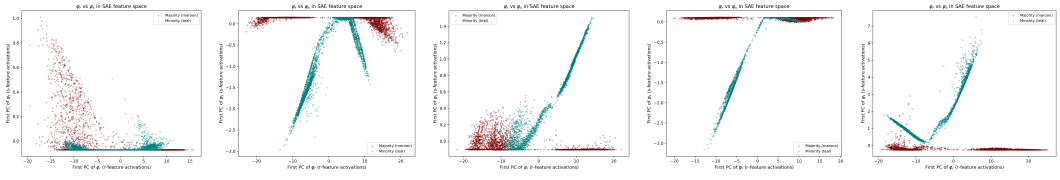

Figure 5: First principal component of ϕ_r (horizontal) vs. first principal component of ϕ_s (vertical) at epochs 1–5 (left to right). **Teal:** minority group; **maroon:** majority group. The minority points form a tight linear cloud whose slope becomes sharper as training progresses. The majority points are concentrated near $\phi_s \approx 0$.

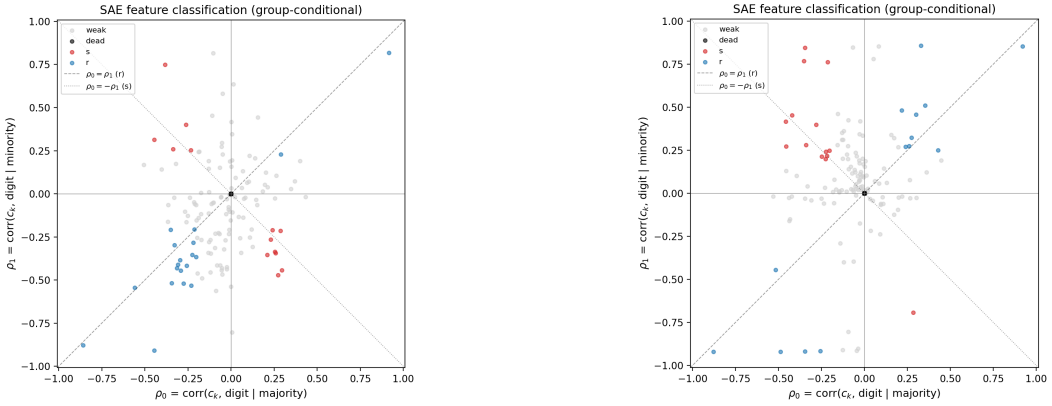

Figure 6: SAE feature classification at epoch 2 (left) and epoch 5 (right). Each point is one SAE feature; axes are group-conditional correlations ρ_0 (majority) and ρ_1 (minority) with digit class. Blue (**r**-features) cluster along $\rho_0 = \rho_1$; red (**s**-features) in the off-diagonal quadrants.

Conclusion. The colored-MNIST experiment demonstrates that the linear feature-mediated model $\mathbf{s} = A_g\,\mathbf{r} + \xi$ can emerge in a learned representation even when the input-space relationship is nonlinear. The key mechanism is that the MLP, through gradient descent, learns a representation in which the nonlinear digit-class information is encoded as a set of approximately linear features. This finding goes further than our initial assumption: rather than merely *preserving* a pre-existing linear relationship, the representation actively *linearises* a nonlinear one.

C Scope of the Isotropic Regime.

Our sharpest rates (Theorem 5.3) require the *isotropic regime* (Definition 5.2), in which $\mathbf{v}$ is a common eigenvector of the four alignment products $A^\top A$, $A^\top B$, $B^\top B$, $B^\top A$. A natural question is whether this condition is genuinely restrictive. We show that the class of operator pairs (A, B) to which Theorem 5.3 applies is strictly larger than those that are already isotropic.

Because the noise ξ depends on the margin variable $Z = \mathbf{v} \cdot \mathbf{r}$ (see Eq (4)), one may decompose any pair (A, B) as

$$A = A' + \boldsymbol{u}_A\,\mathbf{v}^\top, \qquad B = B' + \boldsymbol{u}_B\,\mathbf{v}^\top, \tag{32}$$

for vectors $\boldsymbol{u}_A, \boldsymbol{u}_B \in \mathbb{R}^{d_s}$, and treat the residuals $\xi'_g := Z\,\boldsymbol{u}_g$ as additional noise. By construction ξ'_g depends only on Z, so the conditional independence $\xi' \perp\!\!\!\perp (\mathbf{r}_\perp, y) \mid Z$ holds automatically, and the total noise $\xi + \xi'$ remains compatible with our setup.

For (A', B') to be isotropic, one needs $A'^\top C'\mathbf{v} \propto \mathbf{v}$ for each $C' \in \{A', B'\}$ and similarly with $B'^\top$:

$$A'\mathbf{v},\ B'\mathbf{v}\ \in\ \mathcal{K} := \ker\big(\bar{\Pi}_{\mathbf{v}} \circ A^\top\big)\ \cap\ \ker\big(\bar{\Pi}_{\mathbf{v}} \circ B^\top\big), \tag{33}$$

where $\bar{\Pi}_{\mathbf{v}}$ is a notation introduced in Eq (17). Each map $\bar{\Pi}_{\mathbf{v}} \circ A^\top,\ \bar{\Pi}_{\mathbf{v}} \circ B^\top : \mathbb{R}^{d_s} \to \mathrm{Span}\{\mathbf{v}\}^\perp$ has rank at most $d_r - 1$, so $\dim(\mathcal{K}) \geq d_s - 2(d_r - 1)$. The subspace $\mathcal{K}$ is therefore nontrivial whenever $d_s \geq 2d_r - 1$. Finding the closest isotropic pair amounts to minimising

$$
\begin{aligned}
d_{\mathcal{S}}\big((A', B'),\,(A, B)\big) &= \sup_{r \in \mathrm{supp}(\mathbf{r})} \|(A'-A)r\| + \sup_{r \in \mathrm{supp}(\mathbf{r})} \|(B'-B)r\| \\
&\underset{(1)}{=} K\big(\|\boldsymbol{u}_A\| + \|\boldsymbol{u}_B\|\big) \qquad (34) \\
&= K\big(\|A\mathbf{v} - A'\mathbf{v}\| + \|B\mathbf{v} - B'\mathbf{v}\|\big) \qquad \text{such that,} \\
\inf_{A', B'}\; d_{\mathcal{S}}\big((A', B'),\,(A, B)\big) &\underset{(2)}{\geq} K\big(\|\Pi_{\mathcal{K}^\perp}(A\mathbf{v})\| + \|\Pi_{\mathcal{K}^\perp}(B\mathbf{v})\|\big) \\
&=: d^* \quad (= 0 \quad \text{iff } (A, B) \text{ is already isotropic}), \qquad (35)
\end{aligned}
$$

where (1) uses Eq (32) and (2) uses the condition (33). The resulting eigenvalues are $\mu_{A'} = \|A'\mathbf{v}\|^2$, $\mu_{B'} = \|B'\mathbf{v}\|^2$, and $\mu' = A'\mathbf{v} \cdot B'\mathbf{v}$, and the attractive regime condition (10) is satisfiable whenever $\dim(\mathcal{K}) \geq 2$ (i.e. $d_s \geq 2d_r$), since one can break positive collinearity of $A'\mathbf{v}$ and $B'\mathbf{v}$ by perturbing within $\mathcal{K}$.

When all these conditions are met, the absorbed noise inflates the amplitude bound to $R' = R + K\,\max(\|\boldsymbol{u}_A\|, \|\boldsymbol{u}_B\|)$ and simultaneously reduces the signal strength $\hat{\psi}'^g \cdot \mathbf{v} = \hat{\psi}^g \cdot \mathbf{v} - \boldsymbol{u}_g \cdot \hat{\mathbf{w}}_s$ (see Eq (98)), so the noise condition (101) may fail for large $\|\boldsymbol{u}_g\|$.

In summary, the isotropic regime is less restrictive than it may first appear: any (A, B) with $d_s \geq 2d_r$ admits an isotropic reparametrisation provided the minimum displacement d^* is small enough for the noise condition to hold. In the overparameterised setting $d_s \gg d_r$ that is typical of learned representations (such as the last-layer retraining paradigm), this dimensional requirement is easily met.

D RELAXED VERSION OF THEOREM 5.3

We make the following very natural extra assumptions to make the proof simpler and we explain in Section F how to lift them.

Assumption D.1 (Zero-mean Component). *Write $\boldsymbol{r}_\perp := \boldsymbol{r} - Z\boldsymbol{v} \in \mathrm{Span}\{\boldsymbol{v}\}^\perp$, the component of $\boldsymbol{r}$ perpendicular to $\boldsymbol{v}$. Then,*

$$\mathbb{E}[\boldsymbol{r}_\perp \mid Z,\, g] = 0 \qquad \textit{for all } g \in \{\mathcal{G}_{maj},\, \mathcal{G}_{min}\}. \tag{36}$$

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

$$\psi_t := w_r^t + A^\top w_s^t \in \mathbb{R}^{d_r}, \qquad \phi_t := w_r^t + B^\top w_s^t \in \mathbb{R}^{d_r}, \tag{37}$$

the model initialization satisfies

$$\boldsymbol{v}\cdot\psi_0 \geq 1, \qquad \boldsymbol{v}\cdot\phi_0 \geq 1, \qquad \bar{\Pi}_{\boldsymbol{v}}\psi_0 = 0, \qquad \bar{\Pi}_{\boldsymbol{v}}\phi_0 = 0. \tag{38}$$

Special extra-condition $\star$ *In Section E.4, we further assume $(\boldsymbol{v}\cdot\psi_0, \boldsymbol{v}\cdot\phi_0) \in \mathcal{T}$, where $\mathcal{T}$ is introduced in Eq* (69)*.*

Assumption D.3 (Noise-free case). *For clarity, we analyze the dynamics without perturbation, setting $\xi = 0$.*

Assumption D.1 is simply made to make the perpendicular gradient vanish (Lemma E.1). In Section F.1 we discuss how to lift this assumption.

Assumption D.2 is very reasonable in light of Remark E.4 and guarantees numerical stability as required in the proof of Propositions H.3 and H.4. The orthogonality condition $\bar{\Pi}_{\mathbf{v}}\psi_0 = \bar{\Pi}_{\mathbf{v}}\phi_0 = 0$ goes hand in hand with condition (36). A discussion on how to lift this assumption is presented in Section F.1. The special condition $\star$ is very mild as well as explained in Remark E.6.

Assumption D.3 is introduced solely to simplify the computations in the proof. This assumption might seem to overly simplify the problem, but it is still aligned with formalism of spurious correlation Section 2. In fact, although it makes **s** a deterministic function of **r**, the function itself is not linear ($A \neq B$). Therefore, a linear model that learns to rely on **s** would fall in a spurious pattern trap. A discussion on how to remove this assumption is presented in Section F.2.

We now restate the theorem in the isotropic regime.

Theorem D.4 (Isotropic regime)**.** *Assume the spurious correlation is isotropic 5.2, and that assumptions D.1-D.3 are verified. Without loss of generality, assume $\tilde{\gamma}_{maj} = 1$ and $\tilde{\gamma}_{min} \geq 1$ and define*

$$\alpha := \frac{\tilde{\gamma}_{min}\,(1+\mu)}{1+\mu_A}. \tag{39}$$

Then
(i) If $\alpha < 1$, there exist two positive constants κ_{maj} and κ_{min} independent on ε such that:

$$\mathbb{E}_{\mathcal{G}_{maj}}\big[1-p_y^t\big] \underset{t\to\infty}{\sim} \frac{\kappa_{maj}}{(1-\varepsilon)\,z_t}, \qquad \kappa_{maj} = \frac{\tilde{\gamma}_{min}(1+\mu_B) - (1+\mu)}{\tilde{\gamma}_{min}\,\Sigma}, \tag{40}$$

$$\mathbb{E}_{\mathcal{G}_{min}}\big[1-p_y^t\big] \underset{t\to\infty}{\sim} \frac{\kappa_{min}}{\varepsilon\, z_t}, \qquad \kappa_{min} = \frac{(1+\mu_A) - \tilde{\gamma}_{min}(1+\mu)}{\tilde{\gamma}_{min}^2\,\Sigma}, \tag{41}$$

with Σ defined in Eq (26)*.*
(ii) If $\alpha \geq 1$, the classification error decay at different rates:

$$\mathbb{E}_{\mathcal{G}_{maj}}\big[1-p_y^t\big] \underset{t\to\infty}{\sim} \frac{1}{(1-\varepsilon)(1+\mu_A)\,z_t}, \tag{42}$$

$$\mathbb{E}_{\mathcal{G}_{min}}\big[1-p_y^t\big] = \Theta\left(\frac{(\ln z_t)^{\alpha-1}}{z_t^\alpha}\right). \tag{43}$$

(iii) Furthermore, these different rate decay are continuous with respect to $\tilde{\gamma}_{min}$ at the phase transition point $\alpha = 1$.

Continuity at the phase transition At the threshold $\alpha = 1$ (i.e. $\tilde{\gamma}_{\min} = (1+\mu_A)/(1+\mu)$):

- $\kappa_{\min}(\tilde{\gamma}_{\min}) \to 0$: the factor of the minority's $1/z_t$ rate vanishes continuously.
- $\kappa_{\text{maj}}(\tilde{\gamma}_{\min}) \to 1/(1+\mu_A)$: the majority factor continuously matches the Regime (ii) rate (42).
- The minority rate $(\ln z_t)^{\alpha-1}/z_t^{\alpha}$ with $\alpha \to 1^+$ approaches $1/z_t$, matching Regime (i) rate (41).

E Proof of Theorem D.4

The proof proceeds in four stages. First, we establish geometric properties of the gradient updates (Lemma E.1 and Proposition E.2): the weight symmetry inherited from the loss and the conditional zero-mean structure of $\mathbf{r}_\perp$ together confine the parameter trajectory to the one-dimensional subspace $\mathrm{Span}\{\mathbf{v}\}$, reducing the high-dimensional dynamics to the scalar projections $a_t := \mathbf{v}\cdot\psi_t$ and $b_t := \mathbf{v}\cdot\phi_t$. Second, we derive the coupled recurrences satisfied by (a_t, b_t) (Proposition E.3), identify them as an Euler discretisation of a 2D autonomous ODE, and control the approximation error (Lemma H.6). Third, an asymptotic synchronisation argument shows that the gap $a(z) - b(z)$ converges to an explicit constant d determined by the eigenvalues μ_A, μ_B, μ and the group proportions, effectively reducing the 2D system to a single perturbed 1D equation. Finally, Laplace-method asymptotics and the solution of the resulting ODE (Proposition H.4) combine to yield the group-wise rates κ_{maj} and $\kappa_{\min}$.

E.1 Auxiliary Results

Lemma E.1 (Perpendicular Gradient Vanishes)**.** *For all $t \geq 0$ and $g \in \{\mathcal{G}_{maj}, \mathcal{G}_{min}\}$,*

$$\mathbb{E}_g\big[q^t\, \boldsymbol{r}_\perp\big] = 0, \qquad \textit{where } \boldsymbol{r}_\perp := \boldsymbol{r} - \Pi_{\boldsymbol{v}}\boldsymbol{r} = \bar{\Pi}_{\boldsymbol{v}}\boldsymbol{r}.$$

Proof. We proceed step by step.

Joint induction. We prove the following two facts simultaneously by induction on $t \geq 0$:

(I) $\psi_t \cdot \mathbf{r}_\perp = 0$ and $\phi_t \cdot \mathbf{r}_\perp = 0$ almost surely.

(II) $\mathbb{E}_g\big[q^t \mathbf{r}_\perp\big] = 0$ for $g \in \{\mathcal{G}_{\text{maj}}, \mathcal{G}_{\min}\}$.

Base case ($t = 0$): (I) holds by Assumption D.2. (II) then follows from the cancellation argument below, applied with $\bar{\Pi}_{\mathbf{v}}\psi_0 = \bar{\Pi}_{\mathbf{v}}\phi_0 = 0$.
Inductive step: Assuming (I) at time t, the argument below establishes (II) at time t. Proposition E.2 then uses (II) at time t to establish (I) at time $t+1$.

The model's predicted logit is

$$w^t \cdot \mathbf{x} = (w_{\mathbf{r}}^t) \cdot \mathbf{r} + (w_{\mathbf{s}}^t) \cdot \mathbf{s} = \begin{cases} (w_r^t + A^\top w_s^t) \cdot \mathbf{r} = \psi_t \cdot \mathbf{r} & \text{on } \mathcal{G}_{\text{maj}}, \\ (w_r^t + B^\top w_s^t) \cdot \mathbf{r} = \phi_t \cdot \mathbf{r} & \text{on } \mathcal{G}_{\min}, \end{cases} \tag{44}$$

so that $q^t = \sigma(-\psi_t \cdot \mathbf{r})$ on $\mathcal{G}_{\text{maj}}$, and $q^t = \sigma(-\phi_t \cdot \mathbf{r})$ on $\mathcal{G}_{\min}$. Expanding $\mathbf{r}$ with $\mathbf{r} = Z\mathbf{v} + \mathbf{r}_\perp$ (see Assumption D.1):

$$\text{On } \mathcal{G}_{\text{maj}}: \; q^t = \sigma(-\psi_t \cdot \mathbf{r}) = \sigma(-\mathbf{v}\cdot\psi_t\, Z - \mathbf{r}_\perp \cdot \psi_t) = \sigma(-\mathbf{v}\cdot\psi_t\, Z), \tag{45}$$

$$\text{On } \mathcal{G}_{\min}: \; q^t = \sigma(-\phi_t \cdot \mathbf{r}) = \sigma(-\mathbf{v}\cdot\phi_t\, Z - \mathbf{r}_\perp \cdot \phi_t) = \sigma(-\mathbf{v}\cdot\phi_t\, Z), \tag{46}$$

where we used the inductive hypothesis (I), $\mathbf{r}_\perp \cdot \psi_t = 0$ and $\mathbf{r}_\perp \cdot \phi_t = 0$ almost surely to remove the $\mathbf{r}_\perp$-terms. Therefore, q^t becomes a function of Z and by the total expectation property we have:

$$\mathbb{E}_{\mathcal{G}_{\min}}\big[q^t \mathbf{r}_\perp\big] = \mathbb{E}_{\mathcal{G}_{\min}}\Big[\sigma(-\mathbf{v}\cdot\phi_t\, Z) \cdot \underbrace{\mathbb{E}\big[\mathbf{r}_\perp \mid Z, \mathcal{G}_{\min}\big]}_{=\,0 \text{ by Eq (36)}}\Big] = 0.$$

The same argument with ψ_t instead of ϕ_t yields $\mathbb{E}_{\mathcal{G}_{\text{maj}}}[q^t \mathbf{r}_\perp] = 0$. □

Proposition E.2 (Parameter Updates Stay in $\mathrm{Span}\{\mathbf{v}\}$)**.** *For all $t \geq 0$,*

$$\psi_t \cdot \boldsymbol{r}_\perp = 0 \qquad \textit{and} \qquad \phi_t \cdot \boldsymbol{r}_\perp = 0 \qquad \textit{almost surely}. \tag{47}$$

Proof. Let's write $\psi_{t+1} - \psi_t = -h_t(\nabla_{w_r}\mathcal{L}^t + A^\top \nabla_{w_s}\mathcal{L}^t)$.

On one hand, using Lemma E.1 and $\mathbf{r} = Z\mathbf{v} + \mathbf{r}_\perp$:

$$\nabla_{w_r}\mathcal{L}^t = -\mathbb{E}\Big[q^t \mathbf{r}\Big] = -\mathbb{E}\big[Z\, q^t\big]\mathbf{v} - \underbrace{\mathbb{E}\big[q^t \mathbf{r}_\perp\big]}_{=0} = -\mathbb{E}\big[Z q^t\big]\mathbf{v}. \tag{48}$$

On the other hand, by Eq (5):

$$\begin{aligned}
\nabla_{w_s}\mathcal{L}^t &= -\mathbb{E}\big[q^t \mathbf{s}\big] \\
&= -\Big((1-\varepsilon)\, A\, \mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[q^t \mathbf{r}\big] + \varepsilon\, B\, \mathbb{E}_{\mathcal{G}_{\text{min}}}\big[q^t \mathbf{r}\big]\Big) \\
&= -\Big((1-\varepsilon)\, \mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[Z\, q^t\big] A\mathbf{v} + \varepsilon\, \mathbb{E}_{\mathcal{G}_{\text{min}}}\big[Z\, q^t\big] B\mathbf{v}\Big) \quad \text{(by Lemma E.1)}
\end{aligned} \tag{49}$$

Since $\bar{\Pi}_{\mathbf{v}}\psi_0 = 0$ (Assumption D.2), and the update $\psi_{t+1} - \psi_t = -h_t(\nabla_{w_r}\mathcal{L}^t + A^\top \nabla_{w_s}\mathcal{L}^t)$ lies in $\mathrm{Span}\{\mathbf{v}\} + A^\top \mathrm{Span}\{A\mathbf{v}\} + A^\top \mathrm{Span}\{B\mathbf{v}\} = \mathrm{Span}\{\mathbf{v}\}$ (under isotropic regime 5.2), it follows by induction that $\psi_t \cdot \mathbf{r}_\perp = 0$ almost surely for all t. The same argument gives $\phi_t \cdot \mathbf{r}_\perp = 0$ almost surely. □

E.2 Discrete and Continuous Error Dynamics

Now we can write the scalar equations that characterize the group-wise classification errors. We show that these discrete recurrence equation can be seen as approximations of a 2D-ODE.

Define the group-specific drift functions

$$\mathcal{I}_{\text{maj}}(x) := \int_1^K z\, \sigma(-xz)\, \lambda_{\text{maj}}(z)\, \mathrm{d}z, \qquad \mathcal{I}_{\text{min}}(x) := \int_{\tilde{\gamma}_{\text{min}}}^K z\, \sigma(-xz)\, \lambda_{\text{min}}(z)\, \mathrm{d}z, \tag{50}$$

where λ_{maj} and λ_{min} are the group densities from Assumption D.1. Both functions are decreasing.

Define the scalar projections

$$a_t := \mathbf{v} \cdot \psi_t, \qquad b_t := \mathbf{v} \cdot \phi_t. \tag{51}$$

Proposition E.3 (Coupled Scalar Dynamics)**.** *Under Assumptions D.1-D.3, the scalars a_t and b_t evolve according to the coupled recurrences*

$$a_{t+1} - a_t = h_t(1-\varepsilon)(1+\mu_A)\, \mathcal{I}_{maj}(a_t) + h_t\, \varepsilon\, (1+\mu)\, \mathcal{I}_{min}(b_t), \tag{52}$$

$$b_{t+1} - b_t = h_t\, \varepsilon(1+\mu_B)\, \mathcal{I}_{min}(b_t) + h_t(1-\varepsilon)(1+\mu)\, \mathcal{I}_{maj}(a_t). \tag{53}$$

Moreover, under condition 10 the two sequences are strictly increasing and for all $t \geq 0$,

$$a_t \geq 1 \qquad \textit{and} \qquad b_t \geq 1. \tag{54}$$

Proof. By combining Eqs (48)-(49),

$$\begin{aligned}
\psi_{t+1} - \psi_t &= -h_t\big(\nabla_{w_r}\mathcal{L}^t + A^\top \nabla_{w_s}\mathcal{L}^t\big) \\
&= h_t\, \mathbb{E}\big[Z\, q^t\big]\mathbf{v} + (1-\varepsilon) h_t\, \mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[Z\, q^t\big] A^\top A\mathbf{v} + \varepsilon h_t\, \mathbb{E}_{\mathcal{G}_{\text{min}}}\big[Z\, q^t\big] A^\top B\mathbf{v} \\
&= h_t(1-\varepsilon)\, \mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[Z\, q^t\big](\mathbf{v} + A^\top A\mathbf{v}) + \varepsilon h_t\, \mathbb{E}_{\mathcal{G}_{\text{min}}}\big[Z\, q^t\big](\mathbf{v} + A^\top B\mathbf{v}) \\
&= h_t(1-\varepsilon)(1+\mu_A)\, \mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[Z\, q^t\big]\mathbf{v} + h_t \epsilon (1+\mu)\, \mathbb{E}_{\mathcal{G}_{\text{min}}}\big[Z\, q^t\big]\mathbf{v},
\end{aligned} \tag{55}$$

using Eq (24) in the last line. By an identical argument we also get:

$$\phi_{t+1} - \phi_t = h_t \varepsilon (1+\mu_B)\, \mathbb{E}_{\mathcal{G}_{\text{min}}}\big[Z\, q^t\big]\mathbf{v} + h_t(1-\varepsilon)(1+\mu)\, \mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[Z\, q^t\big]\mathbf{v}. \tag{56}$$

Projecting Eq. (55) onto $\mathbf{v}$:

$$a_{t+1} - a_t = h_t(1-\varepsilon)(1+\mu_A)\,\mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[Z\,q^t\big] + h_t\varepsilon(1+\mu)\,\mathbb{E}_{\mathcal{G}_{\text{min}}}\big[Z\,q^t\big]. \tag{57}$$

Using Eq (45)-(46), $q^t = \sigma(-a_t Z)$ on $\mathcal{G}_{\text{maj}}$ and $\sigma(-b_t Z)$ on $\mathcal{G}_{\text{min}}$, and since $Z \mid \mathcal{G}_g$ has density λ_g for $g \in \{\text{min}, \text{maj}\}$(Assumption D.1) we then have:

$$\mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[Z\,q^t\big] = \mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[Z\,\sigma(-a_t Z)\big] = \int_1^K z\,\sigma(-a_t z)\,\lambda_{\text{maj}}(z)\,\mathrm{d}z = \mathcal{I}_{\text{maj}}(a_t)$$

$$\mathbb{E}_{\mathcal{G}_{\text{min}}}\big[Z\,q^t\big] = \mathbb{E}_{\mathcal{G}_{\text{min}}}\big[Z\,\sigma(-b_t Z)\big] = \int_{\tilde{\gamma}_{\text{min}}}^K z\,\sigma(-b_t z)\,\lambda_{\text{min}}(z)\,\mathrm{d}z = \mathcal{I}_{\text{min}}(b_t).$$

Substituting back:

$$a_{t+1} - a_t = h_t(1-\varepsilon)(1+\mu_A)\,\mathcal{I}_{\text{maj}}(a_t) + h_t\,\varepsilon\,(1+\mu)\,\mathcal{I}_{\text{min}}(b_t).$$

Because $\mathcal{I}_{\text{maj}}(x), \mathcal{I}_{\text{min}}(x) > 0$ for all x, and given that $1+\mu \geq 0$ by condition 10, and $(a_0, b_0) = (\mathbf{v}\cdot\psi_0,\, \mathbf{v}\cdot\phi_0) \geq 1$ by Assumption D.2, the sequences (a_t) and (b_t) are strictly increasing and $a_t,\, b_t \geq 1$ for all t by induction. □

Remark E.4. We highlight two useful observations:

(i) The sequences $(a_t)_{t\geq 0}$ and $(b_t)_{t\geq 0}$ are strictly increasing and diverge to $+\infty$. In fact, if for instance $(a_t)_{t\geq 0}$ was upper bounded, and if we denote $a_\infty,\, b_\infty \in \mathbb{R}\cup\{+\infty\}$ their limits, one would get

$$\Big((1-\varepsilon)(1+\mu_A)\,\mathcal{I}_{\text{maj}}(a_\infty) + \varepsilon(1+\mu_B)\,\mathcal{I}_{\text{min}}(b_\infty)\Big)\sum_{t\geq 0} h_t \leq \sum_{t\geq 0}(a_{t+1}-a_t) < \infty.$$

Under condition (16), this is only possible if $a_\infty = b_\infty = +\infty$, which is impossible since we assumed (a_t) was bounded.

(ii) Equations (52)–(53) are explicit Euler schemes for the 2D autonomous ODE

$$\big(\dot{u}(z),\, \dot{v}(z)\big) = F(u,v), \qquad \big(u(0),\, v(0)\big) = (a_0,\, b_0) \tag{58}$$

where the 2D vector field F is defined as

$$F(a,b) \;:=\; \begin{pmatrix}(1-\varepsilon)(1+\mu_A)\,\mathcal{I}_{\text{maj}}(a) + \varepsilon(1+\mu)\,\mathcal{I}_{\text{min}}(b)\\ \varepsilon(1+\mu_B)\,\mathcal{I}_{\text{min}}(b) + (1-\varepsilon)(1+\mu)\,\mathcal{I}_{\text{maj}}(a)\end{pmatrix}, \tag{59}$$

Defining $z_t := \sum_{k=0}^{t-1} h_k$, we will approximate $a_t \approx u(z_t)$ and $b_t \approx v(z_t)$ in Lemma H.6.

From now on we work with the continuous trajectory $(a(\cdot), b(\cdot))$ for the synchronisation analysis; the discrete dynamics is recovered *a posteriori* via Lemma H.6 and Lemma H.9 in the conclusion (Section E.4).

E.3 ODE Synchronisation

Let $\Delta(z) := a(z) - \tilde{\gamma}_{\text{min}}\, b(z)$ denote the gap between the two trajectories. Subtracting the two coordinates of F (59) and using α introduced in Eq (39),

$$\begin{aligned}\dot{\Delta} \;&=\; (1-\varepsilon)\big((1+\mu_A) - \tilde{\gamma}_{\text{min}}\,(1+\mu)\big)\,\mathcal{I}_{\text{maj}}(a) \;-\; \varepsilon\Big(\tilde{\gamma}_{\text{min}}\,(1+\mu_B) - (1+\mu)\Big)\,\mathcal{I}_{\text{min}}(b)\\ &=\; (1-\varepsilon)(1+\mu_A)(1-\alpha)\,\mathcal{I}_{\text{maj}}(a) \;-\; \varepsilon\Big(\tilde{\gamma}_{\text{min}}\,(1+\mu_B) - (1+\mu)\Big)\,\mathcal{I}_{\text{min}}(b).\end{aligned} \tag{60}$$

We seek a constant gap $d \in \mathbb{R}$ for which the leading-order forcing of the gap dynamics cancels *on* the manifold $\mathcal{M}_d := \{(a,b) \in \mathbb{R}^2 : a = \tilde{\gamma}_{\text{min}}\, b + d\}$. $\mathcal{M}_d$ is never invariant at finite time (i.e a solution initialised in $\mathcal{M}_d$ does not stay inside), however we'd like to keep the trajectory "close" enough, which we'll rigorously formulate below.

As $b \to +\infty$, and for any scalar d, Lemma H.2 gives

$$\mathcal{I}_{\text{maj}}(\tilde{\gamma}_{\min} b + d) = \frac{\lambda_{\text{maj}}(1)}{\tilde{\gamma}_{\min} \lambda_{\min}(\tilde{\gamma}_{\min})} e^{-d} \frac{b}{\tilde{\gamma}_{\min} b + d} \mathcal{I}_{\min}(b) \big(1 + O(1/b)\big)$$
$$= \frac{\lambda_{\text{maj}}(1)}{\tilde{\gamma}^2_{\min} \lambda_{\min}(\tilde{\gamma}_{\min})} e^{-d} \mathcal{I}_{\min}(b) \big(1 + O(1/b)\big), \tag{61}$$

so that, when $\big(a(z),\, b(z)\big) \in \mathcal{M}_d$ i.e $a(z) = \tilde{\gamma}_{\min} b(z) + d$,

$$\dot{\Delta}(z) = \mathcal{I}_{\min}(b(z))\Big[(1-\varepsilon)(1+\mu_A)(1-\alpha) \tfrac{\lambda_{\text{maj}}(1)}{\tilde{\gamma}^2_{\min} \lambda_{\min}(\tilde{\gamma}_{\min})} e^{-d}$$
$$- \varepsilon\Big(\tilde{\gamma}_{\min}(1+\mu_B) - (1+\mu)\Big)\Big] + O\big(\mathcal{I}_{\min}(b(z))/|b(z)|\big). \tag{62}$$

We choose d so that the bracket in (62) vanishes, making $\dot{\Delta}(z) = O(\mathcal{I}_{\min}(b(z))/|b(z)|)$. The resulting equation for d admits an unique real solution only when $\alpha < 1$:

$$e^{-d} \;=\; \frac{\varepsilon}{1-\varepsilon} \frac{\tilde{\gamma}^2_{\min} \lambda_{\min}(\tilde{\gamma}_{\min})}{\lambda_{\text{maj}}(1)} \frac{\tilde{\gamma}_{\min}(1+\mu_B) - (1+\mu)}{(1+\mu_A)(1-\alpha)}, \tag{63}$$

$$d \;=\; -\ln\left(\frac{\varepsilon}{1-\varepsilon} \frac{\tilde{\gamma}^2_{\min} \lambda_{\min}(\tilde{\gamma}_{\min})}{\lambda_{\text{maj}}(1)} \frac{\tilde{\gamma}_{\min}(1+\mu_B) - (1+\mu)}{(1+\mu_A)(1-\alpha)} \right).$$

The complementary case $\alpha \geq 1$ will be covered in Section E.5.

Having fixed the synchronisation constant d, we define the scalar gap $\eta(z) := a(z) - \tilde{\gamma}_{\min} b(z) - d$, which satisfies

$$\dot{\eta} \;=\; (1-\varepsilon)(1+\mu_A)(1-\alpha)\, \mathcal{I}_{\text{maj}}(\tilde{\gamma}_{\min} b + d + \eta) - \varepsilon\Big(\tilde{\gamma}_{\min}(1+\mu_B) - (1+\mu)\Big)\, \mathcal{I}_{\min}(b), \tag{64}$$

by substituting $a = \tilde{\gamma}_{\min} b + d + \eta$ into (60). Thus the 2D ODE reduces to the scalar non-autonomous ODE (64) for η driven by the exogenous signal $b(z)$.

Lemma E.5 (Asymptotic synchronisation)**.** *Under condition 10, there exist $\eta_\star > 0$ and $b_\star > 0$ such that every trajectory $(a,\, b)$ of* (58) *satisfying $|a(z_0) - \tilde{\gamma}_{min}\, b(z_0) - d| \leq \eta_\star$ and $b(z_0) \geq b_\star$ for some $z_0 \geq 0$ verifies*

$$\big|\Delta(z) - d\big| \;=\; O\left(\frac{1}{\ln z}\right) \qquad \text{as } z \to +\infty.$$

In particular, $\Delta(z) \to d$.

Proof. Recall $\eta := \Delta - d = a - \tilde{\gamma}_{\min} b - d$. Writing

$$\mathcal{I}_{\text{maj}}(\tilde{\gamma}_{\min} b + d + \eta) \;=\; \tfrac{\lambda_{\text{maj}}(1)}{\tilde{\gamma}^2_{\min} \lambda_{\min}(\tilde{\gamma}_{\min})} e^{-d-\eta} \big(1 + s(z,\eta)\big)\, \mathcal{I}_{\min}(b), \tag{65}$$
$$s(z,\eta) \;:=\; \tfrac{\tilde{\gamma}_{\min} b}{\tilde{\gamma}_{\min} b + d + \eta} \cdot \tfrac{1 + r_{\text{maj}}(\tilde{\gamma}_{\min} b + d + \eta)}{1 + r_{\min}(b)} - 1,$$

with $r_g(x) := \mathcal{I}_g(x) \big/ \big(\tilde{\gamma}_g\, \lambda_g(\tilde{\gamma}_g) e^{-\tilde{\gamma}_g\, x}/x\big) - 1$, and using Eq (63), Eq (64) becomes

$$\dot{\eta} \;=\; \varepsilon\Big(\tilde{\gamma}_{\min}(1+\mu_B) - (1+\mu)\Big)\, \mathcal{I}_{\min}(b) \Big[e^{-\eta}\big(1 + s(z,\eta)\big) - 1\Big]. \tag{66}$$

Proposition H.2 gives $r_g(x) = \nu_g/x + O(1/x^2)$, so that we have the expansions

$$\frac{\tilde{\gamma}_{\min} b}{\tilde{\gamma}_{\min} b + d + \eta} = 1 - \frac{d+\eta}{\tilde{\gamma}_{\min} b} + O\left(\frac{1}{b^2}\right)$$
$$\frac{1 + r_{\text{maj}}(\tilde{\gamma}_{\min} b + d + \eta)}{1 + r_{\min}(b)} = 1 + \frac{\nu_{\text{maj}}/\tilde{\gamma}_{\min} - \nu_{\min}}{b} + O\left(\frac{1}{b^2}\right),$$

uniformly for $|\eta| \leq 1$. Therefore,

$$s(z,\eta) \;=\; \frac{1}{b(z)}\Big(\nu_{\text{maj}}/\tilde{\gamma}_{\min} - \nu_{\min} - (d+\eta)/\tilde{\gamma}_{\min}\Big) + O\Big(\frac{1}{b(z)^2}\Big), \quad |s(z,\eta)| \;\le\; \frac{K_s}{b(z)} \tag{67}$$

for some constant $K_s > 0$ and all z with $b(z)$ large enough and $|\eta| \le 1$. Moreover

$$|s(z,\eta) - s(z,0)| \le K_s'|\eta|/b(z) \qquad \text{for some } K_s' > 0. \tag{68}$$

From this point, the proof proceeds in three steps: (i) forward invariance of a tube $\{|\eta| \le \eta_\star,\ b \ge b_\star\}$, (ii) the two-sided asymptotic for b via Proposition H.3, and (iii) the Comparison-Lemma argument yielding the $1/\ln z$ rate.

(i) Forward invariance of the tube and asymptotic of b. Choose $\eta_\star \in (0, \frac{1}{2})$. Using $|s| \le K_s/|b|$ from (67), pick $b_\star > 0$ large enough that $\frac{K_s}{b_\star} \;\le\; \sigma$ where $\sigma \;:=\; \frac{1-e^{-\eta_\star}}{1+e^{-\eta_\star}} \;\in\; (0,1)$. Define the tube

$$\mathcal{T} := \{(a,b) \in \mathbb{R}^2 : |a - \tilde{\gamma}_{\min}\, b - d| \le \eta_\star,\ b \ge b_\star\}. \tag{69}$$

On the boundary $\eta = \eta_\star, b \ge b_\star$, $\quad e^{-\eta_\star}(1+s) - 1 \;\le\; e^{-\eta_\star}(1+\sigma) - 1 = \frac{2\,e^{-\eta_\star}}{1+e^{-\eta_\star}} - 1 = -\sigma \;<\; 0$. Since $\mathcal{I}_{\min}(b) \;>\; 0$ and $\varepsilon\Big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\Big) \;>\; 0$, (66) yields $\dot{\eta} < 0$ on this boundary. Symmetrically, on $\eta = -\eta_\star, b \ge b_\star$, $\quad e^{\eta_\star}(1+s) - 1 \;\ge\; e^{\eta_\star}(1-\sigma) - 1 = \frac{2}{1+e^{-\eta_\star}} - 1 = \sigma \;>\; 0$, and $\dot{\eta} > 0$.

Both pushes are inward, so the tube $\mathcal{T}$ is forward invariant (i.e when a trajectory starts inside, it remains inside) as long as $b(z)$ stays over $b_\star$; the latter is guaranteed because $b(z) \to +\infty$ monotonically (by the same argument as in Remark E.4).

(ii) Proposition H.3 requirements. The second line of Eq (58) can be rewritten,

$$\dot{b}(z) \;=\; c(z)\,\mathcal{I}_{\min}\big(b(z)\big), \qquad c(z) \;:=\; \varepsilon(1+\mu_B) + (1-\varepsilon)(1+\mu)\,\frac{\mathcal{I}_{\text{maj}}(a(z))}{\mathcal{I}_{\min}(b(z))}. \tag{70}$$

We show that for every sufficiently small $\eta_\star > 0$, there exist constants $0 < c_-(\eta_\star) \le c_+(\eta_\star) < \infty$ (independent of z) with $c(z) \in [c_-(\eta_\star), c_+(\eta_\star)]$ for all z large enough.

Inside $\mathcal{T}$ we have $a = \tilde{\gamma}_{\min}\, b + d + \eta$ with $\eta = \eta(z) \in [-\eta_\star, \eta_\star]$. Lemma H.2 combined with Eq (63) yields

$$\begin{aligned}
\frac{\mathcal{I}_{\text{maj}}(a(z))}{\mathcal{I}_{\min}(b(z))} \;&\sim\; \frac{\lambda_{\text{maj}}(1)}{\tilde{\gamma}_{\min}\,\lambda_{\min}(\tilde{\gamma}_{\min})}\, e^{-d-\eta(z))}\, \frac{b(z)}{\tilde{\gamma}_{\min}\, b(z) + d + \eta(z)} \\
&\sim\; \underbrace{\frac{\varepsilon\Big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\Big)}{(1-\varepsilon)(1+\mu_A)(1-\alpha)}}_{=:R_0}\, e^{-\eta(z))}
\end{aligned} \tag{71}$$

since $|\eta(z)| \le \eta_\star$ and $b(z) \to +\infty$. Hence there exist $z_{\text{a}}(\eta_\star) \ge z_0$ and constants

$$\begin{aligned}
&R_\pm(\eta_\star) := R_0\, e^{\pm\eta_\star}\big(1 \pm o_{\eta_\star}(1)\big), \qquad \text{s.t.} \quad R_-(\eta_\star) \le \frac{\mathcal{I}_{\text{maj}}(a(z))}{\mathcal{I}_{\min}(b(z))} \le R_+(\eta_\star) \\
&c_\pm(\eta_\star) := \varepsilon(1+\mu_B) + (1-\varepsilon)(1+\mu)\,R_\pm(\eta_\star), \quad c_-(\eta_\star) \;\le\; c(z) \;\le\; c_+(\eta_\star) \quad \forall\, z \ge z_{\text{a}}.
\end{aligned} \tag{72}$$

Both $R_\pm(\eta_\star) \to R_0$ as $\eta_\star \to 0$, so $c_\pm(\eta_\star)$ share the common limit

$$\begin{aligned}
c_0 := \varepsilon(1+\mu_B) + (1-\varepsilon)(1+\mu)R_0 \;&=\; \varepsilon\Big((1+\mu_B) + (1+\mu)\,\tfrac{\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)}{(1+\mu_A)(1-\alpha)}\Big) \\
&=\; \frac{\varepsilon}{(1+\mu_A)(1-\alpha)}\Big((1-\alpha)(1+\mu_A)(1+\mu_B) + \alpha\,(1+\mu_A)(1+\mu_B) - (1+\mu)^2\Big) \\
&=\; \frac{\varepsilon\,\Sigma}{(1+\mu_A)(1-\alpha)},
\end{aligned} \tag{73}$$

with Σ defined in Eq (26). The quantity c_0 is strictly positive with condition 10. By continuity in $\eta_\star$ of $c_\pm(\eta_\star)$, fix $\eta_\star > 0$ small enough that $c_-(\eta_\star) \geq c_0/2 > 0$. With this fixed $\eta_\star$, Eq (72) gives $c(z) \in [c_-(\eta_\star), c_+(\eta_\star)] \subset (0, \infty)$ for all $z \geq z_{\mathrm{a}}$, which is the hypothesis required by Proposition H.3.

(iii) Comparison-lemma. Proposition H.3 applied to b then gives

$$b(z) \;=\; \frac{1}{\tilde{\gamma}_{\min}}\Big(\ln z - \ln\ln z + O(1)\Big), \qquad \frac{M_-}{z} \;\leq\; \mathcal{I}_{\min}\big(b(z)\big) \;\leq\; \frac{M_+}{z} \qquad \text{for all } z \geq z_{\mathrm{b}}, \tag{74}$$

for some constants $0 < M_- \leq M_+$ and $z_{\mathrm{b}} \geq z_{\mathrm{a}}$.

Taylor-expanding $e^{-\eta} = 1 - \eta + \rho(\eta)$ with $|\rho(\eta)| \leq \eta^2 e^{\eta_\star}/2$ for $|\eta| \leq \eta_\star$, and separating the s-contribution, Eq (66) rewrites as

$$\dot{\eta} \;=\; -\gamma(z)\,\eta \;+\; \beta(z) \;+\; R(z,\eta), \tag{75}$$

with

$$\gamma(z) := \varepsilon\Big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\Big)\,\mathcal{I}_{\min}(b(z)), \tag{76}$$

$$\beta(z) := \varepsilon\Big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\Big)\,\mathcal{I}_{\min}(b(z))\,s(z,0), \tag{77}$$

and a remainder $R(z,\eta)$ gathering the remaining $O\big(\rho(\eta)\big)$, $O\big(s(z,\eta){-}s(z,0)\big)$, $O\big(\eta\,s(z,0)\big)$ and $O(\eta\,(s(z,\eta){-}s(z,0)))$ each multiplied by $\varepsilon\Big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\Big)\,\mathcal{I}_{\min}(b)$. By Eqs (74), (67), and (68),

$$\tfrac{\gamma_-}{z} \;\leq\; \gamma(z) \;\leq\; \tfrac{\gamma_+}{z}, \qquad |\beta(z)| \;\leq\; \tfrac{B}{z\,\ln z}, \qquad |R(z,\eta)| \;\leq\; \tfrac{K}{z}\Big(\eta^2 + \tfrac{|\eta|}{\ln z}\Big), \tag{78}$$

for $z \geq z_{\mathrm{b}}$ (possibly enlarging z_{b}), with $\gamma_\pm := \varepsilon\Big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\Big)M_\pm > 0$, and explicit constants $B, K > 0$.

Choose now $\eta_\star$ small enough and $z_{\mathrm{c}} \geq z_{\mathrm{b}}$ large enough that $K\,\eta_\star \;\leq\; \tfrac{\gamma_-}{4}$ and $\tfrac{K}{\ln z_{\mathrm{c}}} \;\leq\; \tfrac{\gamma_-}{4}$.

Then for $z \geq z_{\mathrm{c}}$ and $|\eta| \leq \eta_\star$,

$$\eta\big(-\gamma(z)\eta + R(z,\eta)\big) \;\leq\; -\gamma(z)\,\eta^2 + \tfrac{K}{z}\Big(|\eta|^3 + \tfrac{\eta^2}{\ln z}\Big) \;\leq\; -\tfrac{\gamma_-}{2z}\,\eta^2, \tag{79}$$

so multiplying (75) by $\eta/|\eta|$ and using $\eta\,\dot{\eta} = \frac{\mathrm{d}}{\mathrm{d}z}(\eta^2/2) = |\eta|\,D^+|\eta|$ (Dini derivative Khalil (2002, *Appendix C2*)), one obtains the scalar differential inequality

$$D^+|\eta(z)| \;\leq\; -\frac{\gamma_-}{2\,z}\,|\eta(z)| \;+\; \frac{B}{z\,\ln z}, \qquad z \geq z_{\mathrm{c}}. \tag{80}$$

Khalil's Comparison Lemma (Khalil, 2002, Lemma 3.4) applied to (80) yields

$$|\eta(z)| \;\leq\; |\eta(z_{\mathrm{c}})|\,\exp\Big(-\int_{z_{\mathrm{c}}}^{z} \tfrac{\gamma_-}{2s}\,\mathrm{d}s\Big) + \int_{z_{\mathrm{c}}}^{z} \exp\Big(-\int_{s}^{z} \tfrac{\gamma_-}{2r}\,\mathrm{d}r\Big)\,\tfrac{B}{s\,\ln s}\,\mathrm{d}s. \tag{81}$$

The homogeneous factor evaluates to $\exp(-\tfrac{\gamma_-}{2}\ln(z/z_{\mathrm{c}})) = (z_{\mathrm{c}}/z)^{\gamma_-/2}$, hence the first term of (81) is $O(z^{-\gamma_-/2})$, i.e. polynomial decay. Lemma H.7, the second term equals $O(1/\ln z)$.

Combining, $|\eta(z)| = O(z^{-\gamma_-/2}) + O(1/\ln z) = O(1/\ln z)$, which proves the lemma. □

This synchronisation Lemma naturally generalises to the discrete dynamics as shown in Lemma H.9.

Remark E.6 (On entering the tube). The Lemmas E.5 and H.9 both show that $\mathcal{T}$ is forward invariant. So we just need to prove that both continuous and discrete trajectories eventually fall inside. Noticing that for any initialisation with $a_0, b_0 \geq 1$, the sign of the drift outside the tube pushes η_t inward, the trajectory is expected to reach the tube in finite time. This can be made rigorous at the expense of a longer argument (quantifying the discrete drift rate against the decay

of $\mathcal{I}_{\min}(b_t)$), but we opt for the cleaner approach of assuming it directly (*special condition $\star$ in Assumption D.2*, in the same spirit as the magnitude condition $\mathbf{v} \cdot \psi_0 \geq 1$.

E.4 Conclusion of the Proof when $\alpha < 1$

Proof. We now complete the proof of Theorem D.4 when $\alpha < 1$.

By Lemma H.9 with Assumption D.2 (special condition $\star$), we have $\eta_t = a_t - \tilde{\gamma}_{\min} b_t - d \to 0$. Therefore the coefficient

$$\tilde{c}_t := \varepsilon(1+\mu_B) + (1-\varepsilon)(1+\mu)\,\frac{\mathcal{I}_{\text{maj}}(a_t)}{\mathcal{I}_{\min}(b_t)} = \tilde{c}_t = c_0 + \tilde{\delta}_t, \tag{82}$$

with $\tilde{\delta}_t \to 0$, by exactly the same computation as (71) (with $a_t = \tilde{\gamma}_{\min} b_t + d + \eta_t$ replacing $a(z) = \tilde{\gamma}_{\min} b(z) + d + \eta(z)$). The recurrence (53) rewrites as $b_{t+1} = b_t + h_t\,\tilde{c}_t\,\mathcal{I}_{\min}(b_t)$, which is exactly the setup of Lemma H.6 with $c_t = \tilde{c}_t$. Part (iii) of that lemma gives

$$b_t \;=\; b(z_t) + o(1) \qquad \text{as } t \to \infty. \tag{83}$$

By symmetry (exchanging the roles of a and b), the same argument gives $a_t = a(z_t) + o(1)$.

Minority group error. Proposition H.4 with c_0 defined in Eq (73), $\lambda_0 = \lambda_{\min}(\gamma)$, and $\gamma = \tilde{\gamma}_{\min}$ gives:

$$\tilde{\gamma}_{\min}\, b(z) = \ln z - \ln\ln z + \ln\big(\tilde{\gamma}_{\min}^3\, c_0\, \lambda_{\min}(\tilde{\gamma}_{\min})\big) + o(1),$$

so $\tilde{\gamma}_{\min}\, b(z_t) = \ln(z_t) - \ln\ln(z_t) + \ln\big(\tilde{\gamma}_{\min}^3\, c_0\, \lambda_{\min}(\tilde{\gamma}_{\min})\big) + o(1)$ since $z_t \to +\infty$ by Condition (16). Using Eq (83), we have:

$$\begin{aligned}
\frac{e^{-\tilde{\gamma}_{\min} b_t}}{b_t} &= \frac{e^{-\tilde{\gamma}_{\min} b(z_t) + o(1)}}{b(z_t) + o(1)} = \frac{e^{-b(z_t)}}{b(z_t)} \cdot (1 + o(1)) \\
&= \tilde{\gamma}_{\min}\, \frac{e^{-\ln z_t + \ln\ln z_t - \ln(\tilde{\gamma}_{\min}^3\, c_0 \lambda_{\min}(\tilde{\gamma}_{\min})) + o(1)}}{\ln z_t + o(\ln z_t)} \\
&= \frac{1}{\tilde{\gamma}_{\min}^2\, c_0\, \lambda_{\min}(\tilde{\gamma}_{\min})\, z_t} \cdot (1 + o(1)).
\end{aligned}$$

Therefore with $\mathbb{E}_{\mathcal{G}_{\min}}\big[q^t\big] = \int_{\tilde{\gamma}_{\min}}^{K} \sigma(-b_t\, z)\, \lambda_{\min}(z) \mathrm{d}z$, Lemma H.1 gives

$$\mathbb{E}_{\mathcal{G}_{\min}}\big[q^t\big] \sim \frac{\lambda_{\min}(\tilde{\gamma}_{\min})\, e^{-\tilde{\gamma}_{\min} b_t}}{b_t} \sim \frac{1}{\tilde{\gamma}_{\min}^2\, c_0\, z_t}.$$

We retrieve Eq (41) by expanding c_0 with Eq (73) and using Eq (39).

Majority group error. The identical argument applied to (a_t) with the analogue value for c_0, λ_0 gives Eq (40). □

E.5 Conclusion of the Proof when $\alpha \geq 1$

Proof. When $\alpha \geq 1$, the synchronisation observed in Section E.3 does no longer happen. The proof proceeds in three steps: (i) the drift ratio $\mathcal{I}_{\min}(b)/\mathcal{I}_{\text{maj}}(a) \to 0$ is deduced, (ii) the decoupled asymptotics for a and b are derived, and (iii) the classification errors are computed.

(i) The drift ratio vanishes. Recall that by Eq (60),

$$\dot{\Delta} = \underbrace{(1-\varepsilon)(1+\mu_A)(1-\alpha)}_{\leq 0}\, \mathcal{I}_{\text{maj}}(a) \;+\; \underbrace{\varepsilon\Big((1+\mu) - \tilde{\gamma}_{\min}\,(1+\mu_B)\Big)}_{\leq 0}\, \mathcal{I}_{\min}(b). \tag{84}$$

Since both $\mathcal{I}_{\text{maj}}(a)$ and $\mathcal{I}_{\min}(b)$ are strictly positive and since both factors $(1-\varepsilon)(1+\mu_A)(1-\alpha)$ and $\varepsilon\Big((1+\mu) - \tilde{\gamma}_{\min}\,(1+\mu_B)\Big)$ are not zeros simultaneously, $\Delta \to -\infty$ as soon as one of $\int_0^\infty \mathcal{I}_{\text{maj}}(a)\,\mathrm{d}s$ or $\int_0^\infty \mathcal{I}_{\min}(b)\,\mathrm{d}s$ is ∞. From the first line of the 2D-ODE drift function (59),

$$a(z) - a(0) \;=\; (1{-}\varepsilon)(1{+}\mu_A)\int_0^z \mathcal{I}_{\text{maj}}\big(a(s)\big)\,\mathrm{d}s \;+\; \varepsilon(1{+}\mu)\int_0^z \mathcal{I}_{\text{min}}\big(b(s)\big)\,\mathrm{d}s.$$

Since $a(z) \to +\infty$, their sum diverges:

$$(1{-}\varepsilon)(1{+}\mu_A)\int_0^\infty \mathcal{I}_{\text{maj}}(a)\,\mathrm{d}s \;+\; \varepsilon(1{+}\mu)\int_0^\infty \mathcal{I}_{\text{min}}(b)\,\mathrm{d}s \;=\; +\infty,$$

and $\Delta(z) \to -\infty$. By Lemma H.2, for large enough a and b,

$$\frac{\mathcal{I}_{\text{min}}(b)}{\mathcal{I}_{\text{maj}}(a)} \;=\; \frac{\tilde\gamma_{\text{min}}\,\lambda_{\text{min}}(\tilde\gamma_{\text{min}})}{\lambda_{\text{maj}}(1)}\,\frac{a}{b}\,e^{\Delta}\,\big(1+o(1)\big) \;\le\; \frac{\tilde\gamma_{\text{min}}^2\,\lambda_{\text{min}}(\tilde\gamma_{\text{min}})}{\lambda_{\text{maj}}(1)}\,e^{\Delta}\,\big(1+o(1)\big) \overset{z\to\infty}{\longrightarrow} 0. \tag{85}$$

where the last inequality uses $\Delta = a - \tilde\gamma_{\text{min}}\,b \le 0$ since $\Delta \to -\infty$.

(ii): Decoupled asymptotics for a and b. Denote $R(z) := \frac{\mathcal{I}_{\text{min}}(b(z))}{\mathcal{I}_{\text{maj}}(a(z))}$. From the first line of the 2D-ODE drift function (59),

$$\dot a(z) = (1{-}\varepsilon)(1{+}\mu_A)\,\mathcal{I}_{\text{maj}}(a)\,\Big(1 + \frac{\varepsilon(1{+}\mu)}{(1{-}\varepsilon)(1{+}\mu_A)}\,R(z)\Big) = (1{-}\varepsilon)(1{+}\mu_A)\,\mathcal{I}_{\text{maj}}(a)\,\big(1+o(1)\big). \tag{86}$$

This is a single scalar ODE of the form treated in Propositions H.3 and H.4 with $c_0 = (1{-}\varepsilon)(1{+}\mu_A)$, $\gamma = 1$ and $\lambda_0 = \lambda_{\text{maj}}(1)$. Hence

$$a(z) = \ln z - \ln\ln z + \ln\Big((1{-}\varepsilon)(1{+}\mu_A)\,\lambda_{\text{maj}}(1)\Big) + o(1) \tag{87}$$

$$\mathcal{I}_{\text{maj}}(a(z)) = \Theta\big(1/z\big). \tag{88}$$

On another hand, dividing the two lines of the 2D-ODE drift function (59) and using $R(z) \to 0$:

$$\frac{\dot b}{\dot a} = \frac{\varepsilon(1{+}\mu_B)\,R(z) + (1{-}\varepsilon)(1{+}\mu)}{(1{-}\varepsilon)(1{+}\mu_A) + \varepsilon(1{+}\mu)\,R(z)} \;=\; \frac{1{+}\mu}{1{+}\mu_A} + O\big(R(z)\big),$$

$$\dot b = \frac{1{+}\mu}{1{+}\mu_A}\,\dot a + O\big(\dot a\,R(z)\big). \tag{89}$$

Combining Eqs (86) and (88) gives $\dot a = \Theta(1/z)$, while combining combining Eqs (84) and (88) gives

$$\begin{aligned}
\dot\Delta \;&\le\; (1{-}\varepsilon)(1{+}\mu_A)(1{-}\alpha)\,\mathcal{I}_{\text{maj}}(a) \;\le\; -\frac{C}{z} \\
\Delta(z) \;&\le\; -\,C\ln z + O(1), \qquad \text{s.t.} \\
R(z) \;&\le\; \frac{\tilde\gamma_{\text{min}}^2\,\lambda_{\text{min}}(\tilde\gamma_{\text{min}})}{\lambda_{\text{maj}}(1)}\;e^{\Delta(z)}\,\big(1+o(1)\big) \;\le\; M/z^C \qquad \text{(by Eq (85))}
\end{aligned} \tag{90}$$

for some positive constants C and M. Therefore $\dot a\,R(z) = O(1/z^{C+1})$ and integrating (89):

$$\begin{aligned}
b(z) - b(z_0) \;&=\; \frac{1{+}\mu}{1{+}\mu_A}\,\big(a(z) - a(z_0)\big) + O\big(1\big), \qquad \text{i.e.} \\
\tilde\gamma_{\text{min}}\,b(z) \;&=\; \alpha\,a(z) + O(1) \qquad \text{using Eq (39)} \\
&=\; \alpha\,\ln z - \alpha\,\ln\ln z + O(1) \qquad \text{by Eq (87).}
\end{aligned} \tag{91}$$

(iii) Classification errors. We conclude as in Section E.4. From Eq (87) and Lemma H.1 :

$$\mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[1{-}p_y^t\big] \;\sim\; \frac{\lambda_{\text{maj}}(1)\,e^{-a_t}}{a_t} \;\sim\; \frac{\lambda_{\text{maj}}(1)\,e^{-a(z_t)}}{a(z_t)} \;\sim\; \frac{1}{(1{-}\varepsilon)(1{+}\mu_A)\,z_t}.$$

Likewise, Eq (91) and Lemma H.1 give:

$$\mathbb{E}_{\mathcal{G}_{\text{min}}}\big[1{-}p_y^t\big] \;\sim\; \frac{\lambda_{\text{min}}(\tilde\gamma_{\text{min}})\,e^{-\tilde\gamma_{\text{min}}\,b_t}}{b_t} \;\sim\; \frac{\lambda_{\text{min}}(\tilde\gamma_{\text{min}})\,e^{-\tilde\gamma_{\text{min}}\,b(z_t)}}{b(z_t)} \;=\; \Theta\bigg(\frac{(\ln z_t)^{\alpha-1}}{z_t^{\alpha}}\bigg).$$

□

F GENERALIZING TO THE PROOF OF THEOREM 5.3

The proof of Theorem D.4 rests on Assumptions D.1–D.3. This section discusses these assumptions: their necessity, what the proof can still say when they are relaxed, and which generalisations remain open.

F.1 RELAXING ASSUMPTIONS D.1 AND D.2

Assumptions D.1 and D.2 can be weakened by assuming simply that:

Assumption F.1 (Non-degeneracy of $\mathbf{r}_\perp$ at the margin boundary). *For every unit direction $\hat{\boldsymbol{\delta}} \in \mathbb{R}^{d_r}$ with $\boldsymbol{v}^\top \hat{\boldsymbol{\delta}} = 0$, the conditional distribution of $\zeta := \hat{\boldsymbol{\delta}}^\top \boldsymbol{r}_\perp$ given $Z = 1$ and any group g satisfies*

$$\Pr\big(\zeta < 0 \;\big|\; Z{=}1,\, g\big) > 0. \tag{92}$$

This non-degeneracy assumption simply requires that for every direction $\hat{\boldsymbol{\delta}}$ such that $\hat{\boldsymbol{\delta}}^\top \mathbf{v} = 0$, $\mathbf{r}_\perp \mid Z{=}1$ never puts all its mass on one single side of the hyperplane $(\hat{\boldsymbol{\delta}})^\perp$. In geometric terms, the origin lies in the interior of the convex hull of the support of $\mathbf{r}_\perp \mid Z{=}1$ in the subspace $(\mathbf{v})^\perp$.

In this case, the parameters ψ_t and ϕ_t acquire a perpendicular component:

$$\psi_t = a_t\, \mathbf{v} + \boldsymbol{\delta}_t^\psi, \qquad \phi_t = b_t\, \mathbf{v} + \boldsymbol{\delta}_t^\phi, \qquad \boldsymbol{\delta}_t^\psi,\, \boldsymbol{\delta}_t^\phi \in (\mathrm{Span}\{\mathbf{v}\})^\perp. \tag{93}$$

Furthermore $\|\boldsymbol{\delta}_t^\psi\|$ and $\|\boldsymbol{\delta}_t^\phi\|$ remain $O(1)$, and when $\tilde{\gamma}_{\text{maj}} = 1$, $\tilde{\gamma}_{\text{min}} \geq 1$ (with analogous result in the mirror case), the asymptotic rates are slightly changed:

(i) If $\alpha < 1$,

$$\kappa_{\text{maj}} = \frac{\tilde{\gamma}_{\text{min}}(1{+}\mu_B) - (1{+}\mu)}{\tilde{\gamma}_{\text{min}}\,\Sigma} \cdot \frac{1}{\Lambda_{\text{maj}}}, \tag{94}$$

$$\kappa_{\text{min}} = \frac{(1{+}\mu_A) - \tilde{\gamma}_{\text{min}}(1{+}\mu)}{\tilde{\gamma}_{\text{min}}^2\,\Sigma} \cdot \frac{1}{\Lambda_{\text{min}}}, \tag{95}$$

(ii) If $\alpha \geq 1$,

$$\mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[1 - p_y^t\big] \;\underset{t\to\infty}{\sim}\; \frac{1}{(1{-}\varepsilon)(1{+}\mu_A)\,\Lambda_{\text{maj}}\, z_t}, \tag{96}$$

$$\mathbb{E}_{\mathcal{G}_{\text{min}}}\big[1 - p_y^t\big] \;=\; \Theta\!\left(\frac{(\ln z_t)^{\alpha-1}}{z_t^{\alpha}}\right), \tag{97}$$

with Λ_{maj} and Λ_{min} only depending on the group-wise distributions of $\mathbf{r}_\perp \mid Z = \tilde{\gamma}_g$. More precisely, when $\mathrm{Law}(\mathbf{r}_\perp \mid Z = \tilde{\gamma}_{\text{maj}}, \mathcal{G}_{\text{maj}}) = \mathrm{Law}(\mathbf{r}_\perp \mid Z = \tilde{\gamma}_{\text{min}}, \mathcal{G}_{\text{min}})$, we get that $\Lambda_{\text{maj}} = \Lambda_{\text{min}}$ which falls back to Eqs (40)-(41).

However, this would require considerable more complex mathematical development, therefore we leave this generalisation opened.

F.2 RELAXING ASSUMPTION D.3 (FEATURE NOISE)

We simplify the notations in this section this way: a_t^g, A_g, ψ_t^g, μ_g, $\bar{g}$ represent in this order a_t, A, ψ_t, μ_A, $\mathcal{G}_{\text{min}}$ when $g = \text{maj}$ and b_t, B, ϕ_t, μ_B, $\mathcal{G}_{\text{maj}}$ when $g = \text{min}$.

When $\xi \neq 0$, the logit by Eqs (44) becomes $w^t \cdot \mathbf{x} = \begin{cases} \psi_t \cdot \mathbf{r} + w_s^t \cdot \xi & \text{on } \mathcal{G}_{\text{maj}} \\ \phi_t \cdot \mathbf{r} + w_s^t \cdot \xi & \text{on } \mathcal{G}_{\text{min}} \end{cases}$, where we combined Eq (37) and Proposition 5.4 to define

$$\psi_t^g = \hat{\psi}^g\,(\log z_t - \log\log z_t) + O(1), \qquad \hat{\psi}^g := \hat{\mathbf{w}}_r + A_g^\top \hat{\mathbf{w}}_s. \tag{98}$$

Define the effective sigmoids:

$$\sigma^t_{\text{eff}}(u,z) := \mathbb{E}_{\xi|Z=z}\big[\sigma(u - w^t_s \cdot \xi)\big], \tag{99}$$

and the noise symmetrical moment-generating functions:

$$\Lambda_\xi(w_s, z) := \mathbb{E}_{\xi|Z=z}\big[e^{-w_s\cdot\xi}\big]. \tag{100}$$

We have the upper bound $\Lambda_\xi(w^t_s, z) \le e^{\|w^t_s\|R} =: e^{M_t}$, where $M_t := \|w^t_s\|R$.

(i) Relaxed assumption: Noise bound. Assumption D.3 can be greatly weakened by assuming simply that R (see Eq (23)) is sufficiently small in the sense of:

Assumption F.2 (Noise bound). *The noise bound satisfies for all $g \in \{maj, min\}$*

$$(\hat{\psi}^g \cdot \boldsymbol{\nu})\,\tilde{\gamma}_g > \|\hat{\mathbf{w}}_s\|\,R \tag{101}$$

To understand the utility of this relaxed assumption, let's first show the following result.

Lemma F.3. *For any u, z with $u > M_t$, $z \ge 1$:*

$$\sigma^t_{\text{eff}}(-u,z) = e^{-u}\,\Lambda_\xi(w^t_s, z)\,\big(1 + \rho_t(u,z)\big), \tag{102}$$

with

$$|\rho_t(u,z)| \;\le\; e^{-(u-M_t)}. \tag{103}$$

Proof. For $u > M_t$, $u - w^t_s \cdot \xi \ge u - M_t > 0$ almost surely. Using the identity $\sigma(-v) = e^{-v}/(1+e^{-v})$:

$$\sigma(-u - w^t_s\cdot\xi) = e^{-(u+w^t_s\cdot\xi)} \cdot \frac{1}{1+e^{-(u+w^t_s\cdot\xi)}} = e^{-u}\,e^{-w^t_s\cdot\xi}\,h(u,\xi), \tag{104}$$

where $h(u,\xi) := 1/(1+e^{-(u+w^t_s\cdot\xi)})$. Since $u + w^t_s\cdot\xi > 0$, $h \in (1/2, 1)$ a.s. More precisely:

$$1 - e^{-(u+w^t_s\cdot\xi)} \;\le\; h(u,\xi) \;\le\; 1 \qquad \text{a.s.} \tag{105}$$

Taking expectations in (104) under $Z = z$:

$$\begin{aligned}
&\sigma^t_{\text{eff}}(-u,z) = e^{-u}\,\mathbb{E}_{Z=z}\big[e^{-w^t_s\cdot\xi}\,h(u,\xi)\big] \qquad \text{s.t.}\\
&\sigma^t_{\text{eff}}(-u,z) \;\le\; e^{-u}\,\Lambda_\xi(w^t_s,z) \qquad \text{since } h \le 1,\\
&\sigma^t_{\text{eff}}(-u,z) \;\ge\; e^{-u}\Big(\Lambda_\xi(w^t_s,z) - e^{-u}\,\mathbb{E}_{Z=z}\big[e^{-2w^t_s\cdot\xi}\big]\Big) \qquad \text{since } h \ge 1 - e^{-(u+w^t_s\cdot\xi)}\\
&\qquad\qquad = e^{-u}\,\Lambda_\xi(w^t_s,z)\Big(1 \underbrace{-e^{-u}\,\Lambda_\xi(2\,w^t_s,z)/\Lambda_\xi(w^t_s,z)}_{:=\rho_t(u,z)}\Big),
\end{aligned} \tag{106}$$

with

$$|\rho_t(u,z)| \;\le\; e^{-u}\,\frac{\Lambda_\xi(2\,w^t_s,z)}{\Lambda_\xi(w^t_s,z)} \;\le\; e^{-(u-M_t)}. \tag{107}$$

The last inequality is due to the pointwise inequality $e^{-2w^t_s\cdot\xi} \le e^{-w^t_s\cdot\xi}e^{M_t}$ a.s. □

By Proposition I.5 and Eq (98), for all group g

$$\frac{a^g_t}{\|w^t\|} = \frac{\psi^g_t\cdot\mathbf{v}}{\|w^t\|} = \frac{\hat{\psi}^g\cdot\mathbf{v}\,(\log z_t - \log\log z_t) + O(1)}{\|\hat{\mathbf{w}}\|\,(\log z_t - \log\log z_t) + O(1)} \longrightarrow (\hat{\psi}^g\cdot\mathbf{v})/\|\hat{\mathbf{w}}\| \tag{108}$$
$$\|w^t_s\|/\|w^t\| \longrightarrow \|\hat{\mathbf{w}}_s\|/\|\hat{\mathbf{w}}\|.$$

Therefore:

$$\frac{a_t\,\tilde{\gamma}_{\text{maj}}}{\|w^t_s\|\,R} \longrightarrow \frac{\tilde{\gamma}_{\text{maj}}\,(\hat{\psi}\cdot\mathbf{v})}{\|\hat{\mathbf{w}}_s\|\,R} > 1, \qquad \frac{b_t\,\tilde{\gamma}_{\text{min}}}{\|w^t_s\|\,R} \longrightarrow \frac{\tilde{\gamma}_{\text{min}}\,(\hat{\phi}\cdot\mathbf{v})}{\|\hat{\mathbf{w}}_s\|\,R} > 1, \tag{109}$$

the last inequalities come from Eq (101), thereby emphasizing their utility.

Consequently, for all t large enough and all group g:

$$a_t^g\,\tilde{\gamma}_g - M_t = \left(\frac{a_t^g\,\tilde{\gamma}_g}{\|w_s^t\|R} - 1\right) M_t \;\geq\; c_0\,M_t \;\geq\; c_0'\,\log z_t, \tag{110}$$

for constants $c_0, c_0' > 0$. Combining Lemma F.3 and Eq (110) yields:

$$\sigma_{\text{eff}}^t(-a_t^g Z, Z) = e^{-a_t^g Z}\,\Lambda_\xi(w_s^t, Z)\left(1 + O(z_t^{-c_0'})\right) \qquad \mathcal{G}_g - \text{almost surely}, \tag{111}$$

and the $O(...)$ is uniform in $Z \mid \mathcal{G}_g$.

(ii) Λ_ξ *growth.* For $z \geq 1$, define the noise exponent function:

$$\delta_+(z) \;:=\; \sup_{\xi \in \text{supp}(\xi \mid Z = z)} -\hat{\mathbf{w}}_s \cdot \xi, \tag{112}$$

and the boundary noise exponents for each group:

$$\delta_g \;:=\; \delta_+(\tilde{\gamma}_g), \qquad g \in \{\text{maj}, \text{min}\}. \tag{113}$$

The function $\delta_+(z)$ is non-decreasing in z: as Z grows (farther from the margin boundary), the margin condition $\hat{\psi} \cdot \mathbf{r} + \hat{\mathbf{w}}_s \cdot \xi \geq \gamma_{\text{maj}}$ becomes less restrictive, allowing a wider range of ξ realisations, and hence a larger $-\hat{\mathbf{w}}_s \cdot \xi$ supremum. In particular, $\delta_g = \delta_+(\tilde{\gamma}_g)$ is the *smallest* value of δ_+ over the support of $Z \mid \mathcal{G}_g$. We also have the upper bound $\delta_g \leq \|\hat{\mathbf{w}}_s\|R$ since $|-\hat{\mathbf{w}}_s \cdot \xi| \leq \|\hat{\mathbf{w}}_s\|R$ a.s.

This gives us this following result.

Lemma F.4 (Λ_ξ growth rate)**.** *For each $z \geq 1$ with $\text{supp}(\xi \mid Z = z) \neq \emptyset$,*

$$\Lambda_\xi(w_s^t, z) \;=\; z_t^{\delta_+(z)+o(1)} \qquad \textit{as } t \to \infty, \tag{114}$$

with $o(1)$ uniform in z.

Proof. Write $w_s^t = \hat{\mathbf{w}}_s \log z_t - \hat{\mathbf{w}}_s \log\log z_t + O(1)$ (Proposition 5.4), so:

$$w_s^t \cdot \xi = (\hat{\mathbf{w}}_s \cdot \xi)\log z_t + O(\log\log z_t) \qquad \text{a.s. uniformly in } \xi$$

Then on one hand,

$$\Lambda_\xi(w_s^t, z) \leq \exp\Big(\sup_{\xi \in \text{supp}(\xi \mid Z = z)} -w_s^t \cdot \xi\Big) = \exp\big(\delta_+(z)\log z_t + O(\log\log z_t)\big) = z_t^{\delta_+(z)} \cdot O\big((\log z_t)^C\big)$$

giving the upper bound $\Lambda_\xi(w_s^t, z) \leq z_t^{\delta_+(z)+o(1)}$, with $o(1)$ uniform in z.

On the other hand, fix $\eta > 0$. The set $S_\eta(z) := \{\xi \in \text{supp}(\xi \mid Z = z) : -\hat{\mathbf{w}}_s \cdot \xi > \delta_+(z) - \eta\}$ has positive $(\xi \mid Z = z)$-probability by definition of $\delta_+(z)$. Let $p_\eta(z) := P(\xi \in S_\eta(z) \mid Z = z) > 0$. Then:

$$\Lambda_\xi(w_s^t, z) \geq p_\eta(z) \cdot \exp\big((\delta_+(z) - \eta)\log z_t + O(\log\log z_t)\big) \geq z_t^{\delta_+(z)} \cdot p_\eta(z)\, z_t^{-\eta}\, O\big((\log z_t)^{-C}\big)$$

where the $O(...)$ is uniform in η. Sending $\eta \to 0$ gives the lower bound $\Lambda_\xi(w_s^t, z) \geq z_t^{\delta_+(z)+o(1)}$ with $o(1)$ uniform in z. □

(iii) ODE (58) *reformulation.*

Lemma F.5. *For all group $g \in \{min, maj\}$, the group-g drift function becomes*

$$\mathcal{I}_g^{\text{eff}}(a_t^g, w_s^t) := \mathbb{E}_{\mathcal{G}_g}\big[Z\,\sigma_{\text{eff}}^t(-a_t^g Z, Z)\big] = \Lambda_\xi(w_s^t, z_g^\star)\cdot \frac{\tilde\gamma_g\,\lambda_g(\tilde\gamma_g)\,e^{-\tilde\gamma_g a_t^g}}{a_t^g}\,\big(1+O(1/a_t^g)\big), \tag{115}$$

for some t-dependent $z_g^\star \in (\tilde\gamma_g, K)$.
Similarly, the classification error integral becomes:

$$\mathcal{J}_g^{\text{eff}}(a_t^g, w_s^t) := \mathbb{E}_{\mathcal{G}_g}\big[\sigma_{\text{eff}}^t(-a_t^g Z, Z)\big] = \Lambda_\xi(w_s^t, z_g^\star)\cdot \frac{\lambda_g(\tilde\gamma_g)\,e^{-\tilde\gamma_g a_t^g}}{a_t^g}\,\big(1+O(1/a_t^g)\big). \tag{116}$$

Proof. By Eq (111):

$$\begin{aligned}
\mathcal{I}_g^{\text{eff}}(a_t^g) &= \int_{\tilde\gamma_g}^K z\,\sigma_{\text{eff}}^t(-a_t^g z, z)\,\lambda_g(z)\,dz \\
&\underset{(1)}{=} \int_{\tilde\gamma_g}^K z\,\Lambda_\xi(w_s^t, z)\,e^{-a_t^g z}\,\lambda_g(z)\,dz\,\big(1+O(z_t^{-c_0'})\big) \\
&\underset{(2)}{=} \Lambda_\xi(w_s^t, z_g^\star)\int_{\tilde\gamma_g}^K z\,e^{-a_t^g z}\,\lambda_g(z)\,dz\,\big(1+O(z_t^{-c_0'})\big)
\end{aligned} \tag{117}$$

where the factorisation (1) used the fact that $O(z_t^{-c_0'})$ is uniform in $Z \mid g$, and equality (2) used the mean value theorem for integrals with $z_g^\star \in [\tilde\gamma_g, K]$. Using Lemma H.2 and more precisely Eq (175) on the integral gives the stated result, since $z_t^{-c_0'} = o(1/a_t^g)$ for $c_0' > 0$ (because $a_t^g = \Theta(\log z_t)$). □

Two direct consequences of Lemma F.5 are

$$\mathcal{J}_g^{\text{eff}} = \mathcal{I}_g^{\text{eff}}/\tilde\gamma_g \cdot (1+O(1/a_t^g)), \tag{118}$$

and that

$$\frac{\mathcal{I}_{\text{maj}}^{\text{eff}}(a_t)}{\mathcal{I}_{\text{min}}^{\text{eff}}(b_t)} = \frac{\Lambda_\xi(w_s^t, z_{\text{maj}}^\star)}{\Lambda_\xi(w_s^t, z_{\text{min}}^\star)}\cdot\frac{\tilde\gamma_{\text{maj}}\,\lambda_{\text{maj}}(\tilde\gamma_{\text{maj}})}{\tilde\gamma_{\text{min}}\,\lambda_{\text{min}}(\tilde\gamma_{\text{min}})}\cdot\frac{b_t}{a_t}\,e^{-\tilde\gamma_{\text{maj}}a_t+\tilde\gamma_{\text{min}}b_t}\,\big(1+O(1/b_t)\big). \tag{119}$$

By Lemma F.4, the Λ_ξ ratio in Eq (119) satisfies:

$$\frac{\Lambda_\xi(w_s^t, z_{\text{maj}}^\star)}{\Lambda_\xi(w_s^t, z_{\text{min}}^\star)} \;=\; z_t^{\,\delta_+(z_{\text{maj}}^\star)-\delta_+(z_{\text{min}}^\star)+o(1)}. \tag{120}$$

Since the Laplace weight $e^{-a_t^g z}$ concentrates all mass on the boundary $z = \tilde\gamma_g$, $z_g^\star \to \tilde\gamma_g$ as $t \to \infty$. Therefore, with δ_+ being continuous and bounded, Eq (120) can be rewritten as:

$$\frac{\Lambda_\xi(w_s^t, z_{\text{maj}}^\star)}{\Lambda_\xi(w_s^t, z_{\text{min}}^\star)} \;=\; z_t^{\,\delta_{\text{maj}}-\delta_{\text{min}}+o(1)}. \tag{121}$$

Regarding the scalar update for a_t^g, Eqs (48)-(49) acquire a noise dependent term, so that Eqs (52)-(53) become:

$$\begin{aligned}
a_{t+1}^g - a_t^g = h_t\Big(&\varepsilon_g(1+\mu_g)\,\mathcal{I}_g^{\text{eff}}(a_t^g) + \varepsilon_{\bar g}(1+\mu)\,\mathcal{I}_{\bar g}^{\text{eff}}(a_t^{\bar g})\Big) \\
&+\; h_t\,(A_g\mathbf{v})\cdot\Big(\varepsilon_g\mathbb{E}_g[q^t\xi] + \varepsilon_{\bar g}\mathbb{E}_{\bar g}[q^t\xi]\Big),
\end{aligned} \tag{122}$$

so that the gap function Δ (see Eq (60)) becomes $\Delta_t := (a_t - \tilde\gamma_{\text{min}}\,b_t) - (\delta_{\text{maj}} - \delta_{\text{min}})\log z_t - d_t$ with the new synchronisation term d_t defined by the following:

$$
\begin{aligned}
\Delta_{t+1} - \Delta_t &= h_t \Big((1-\varepsilon)(1+\mu_A)(1-\alpha)\, \mathcal{I}^{\text{eff}}_{\text{maj}}(a_t) - \varepsilon \big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\big)\, \mathcal{I}^{\text{eff}}_{\min}(b_t) \Big) \\
&\quad + \; h_t\, (A\mathbf{v} - \tilde{\gamma}_{\min} B\mathbf{v})\, \mathbb{E}[q^t \xi] + (\delta_{\text{maj}} - \delta_{\min}) \log(z_{t+1}/z_t) + (d_{t+1} - d_t) \\
&= \varepsilon \big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\big)\, \mathcal{I}^{\text{eff}}_{\min}(b_t) \Bigg(\\
&\qquad \frac{(1-\varepsilon)(1+\mu_A)(1-\alpha)}{\varepsilon \big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\big)} \frac{\mathcal{I}^{\text{eff}}_{\text{maj}}(a_t)}{\mathcal{I}^{\text{eff}}_{\min}(b_t)} - 1 \Bigg) \\
&\quad + \; O\big(h_t\, \mathbb{E}[q^t \xi]\big) + \Theta(h_t/z_t) + (d_{t+1} - d_t) \\
&= \varepsilon \big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\big)\, \mathcal{I}^{\text{eff}}_{\min}(b_t) \Bigg(\frac{(1-\varepsilon)(1+\mu_A)(1-\alpha)}{\varepsilon \big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\big)} \times \\
&\qquad \frac{\lambda_{\text{maj}}(1)}{\tilde{\gamma}_{\min}\, \lambda_{\min}(\tilde{\gamma}_{\min})} \cdot \frac{b_t}{\tilde{\gamma}_{\min}\, b_t + (\delta_{\text{maj}} - \delta_{\min}) \log z_t + d_t + \Delta_t} \times \\
&\qquad \underbrace{e^{-a_t + \tilde{\gamma}_{\min} b_t + \left(\delta_{\text{maj}} - \delta_{\min} + o(1)\right) \log z_t}}_{= e^{-\Delta_t - d_t + o(\log z_t)}} \big(1 + O(1/b_t)\big) - 1 \Bigg) \\
&\quad + \; O\big(h_t\, \mathbb{E}[q^t \xi]\big) + \Theta(h_t/z_t) + (d_{t+1} - d_t), \qquad (123)
\end{aligned}
$$

so that in order to get $\Delta_t \to 0$, we necessarily need

$$
\begin{aligned}
e^{-d_t + o(\log z_t)} &\sim \frac{\varepsilon \big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\big)}{(1-\varepsilon)(1+\mu_A)(1-\alpha)} \cdot \frac{\tilde{\gamma}_{\min}\, \lambda_{\min}(\tilde{\gamma}_{\min})}{\lambda_{\text{maj}}(1)} \cdot \frac{\tilde{\gamma}_{\min}\, b_t + (\delta_{\text{maj}} - \delta_{\min}) \log z_t + d_t}{b_t} \\
&\sim \frac{\varepsilon \big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\big)}{(1-\varepsilon)(1+\mu_A)(1-\alpha)} \cdot \frac{\tilde{\gamma}_{\min}\, \lambda_{\min}(\tilde{\gamma}_{\min})}{\lambda_{\text{maj}}(1)} \cdot \frac{\tilde{\gamma}_{\min}\, \hat{\phi} \cdot \mathbf{v} + \delta_{\text{maj}} - \delta_{\min}}{\hat{\phi} \cdot \mathbf{v}}, \qquad (124)
\end{aligned}
$$

the analogous of the free-noise case condition (63), and where we used the b_t expression from Eq (108).

We now show that the $\Theta(1/z_t)$ error rate from Section E.4 is preserved under feature noise, for both groups.

(iii) Preservation of the rates decay (case $\alpha < 1$).

When $\alpha < 1$, the synchronisation argument (Lemma E.5 extended to the noisy setting) carries over with the same Lyapunov structure, ensuring $\Delta_t \to 0$, and therefore $\mathcal{I}^{\text{eff}}_{\text{maj}}(a_t)/\mathcal{I}^{\text{eff}}_{\min}(b_t) = R_0\,(1+o(1))$ with the same R_0 as in the noise free case Eq (71). Therefore, the b-recurrence (122) reads

$$
\begin{aligned}
b_{t+1} - b_t \;&=\; h_t\, \tilde{c}_t\, \mathcal{I}^{\text{eff}}_{\min}(b_t) + h_t\, O\big(C_\xi\, \mathbb{E}[q^t]\big) \\
&\underset{(1)}{=}\; h_t\, \tilde{c}_t\, \Lambda_\xi(w_s^t, z^\star_{\min})\, \frac{\tilde{\gamma}_{\min}\, \lambda_{\min}(\tilde{\gamma}_{\min})\, e^{-\tilde{\gamma}_{\min} b_t}}{b_t} \,(1 + O(1/b_t)) + h_t\, O\big(C_\xi\, \mathbb{E}[q^t]\big) \\
&\underset{(2)}{=}\; h_t\, \tilde{c}_t\, z_t^{\delta_{\min} + o(1)}\, \frac{\tilde{\gamma}_{\min}\, \lambda_{\min}(\tilde{\gamma}_{\min})\, e^{-\tilde{\gamma}_{\min} b_t}}{b_t} \,(1 + O(1/b_t)) + h_t\, O\big(C_\xi\, \mathbb{E}[q^t]\big), \qquad (125)
\end{aligned}
$$

where $C_\xi := \|B\mathbf{v}\|\, R$, we used Lemma F.5 for line (1), Lemma F.4 and $z^\star_{\min} \to \tilde{\gamma}_{\min}$ for line (2), and where $\tilde{c}_t = c_0 + o(1)$ is exactly the same as in the noise-free setup Eq (82).

Under synchronisation, $\mathcal{I}^{\text{eff}}_{\text{maj}} = R_0\, \mathcal{I}^{\text{eff}}_{\min}\,(1+o(1))$, and combined with Eq (118), we have $\mathbb{E}[q^t] = (1-\varepsilon)\, \mathcal{J}^{\text{eff}}_{\text{maj}}(a_t, w_s^t) + \varepsilon\, \mathcal{J}^{\text{eff}}_{\min}(b_t, w_s^t) \le C'\, \mathcal{I}^{\text{eff}}_{\min}(b_t, w_s^t)$ for some $C' > 0$ and all t large enough. Therefore (125) can be rewritten as

$$
b_{t+1} - b_t = h_t\, \big(\tilde{c}_t + O(C_\xi)\big)\, z_t^{\delta_{\min} + o(1)}\, \frac{\tilde{\gamma}_{\min}\, \lambda_{\min}(\tilde{\gamma}_{\min})\, e^{-\tilde{\gamma}_{\min} b_t}}{b_t} \,(1 + O(1/b_t)), \qquad (126)
$$

with the term $O(C_\xi)$ dependent on the noise and on t and bounded uniformly in ξ. This is the exact prerequisite of Proposition H.5, which yields

$$\tilde{\gamma}_{\min}\, b(z) \;=\; (1{+}\delta_{\min}{+}o(1))\ln z - \ln\ln z + O(1), \tag{127}$$

$$z^{\delta_{\min}+o(1)}\,\mathcal{I}_{\min}\big(b(z)\big) \;=\; \frac{1}{\big(\tilde{\gamma}_{\min}\,c_0 + O(C_\xi)\big)\big(1{+}\delta_{\min}\big)}\;\cdot\;\frac{e^{O(1)}}{z}, \tag{128}$$

with the bounded terms $O(...)$ independent on c_0.

Combining with Lemma H.1 and Lemma H.2: $\mathcal{I}_{\min}/\mathcal{J}_{\min} = \tilde{\gamma}_{\min}\,(1{+}O(1/b_t))$, we have:

$$\begin{aligned} z^{\delta_{\min}+o(1)}\,\mathcal{J}_{\min}\big(b(z)\big) &= \frac{1}{\tilde{\gamma}_{\min}}\; z^{\delta_{\min}+o(1)}\,\mathcal{I}_{\min}\big(b(z)\big)\;(1{+}O(1/b)) \\ &= \frac{1}{\big(\tilde{\gamma}^2_{\min}\,c_0 + O(C_\xi)\big)\big(1{+}\delta_{\min}\big)}\;\cdot\;\frac{e^{O(1)}}{z} \end{aligned} \tag{129}$$

Now, from Eq (116):

$$\mathcal{J}^{\text{eff}}_{\min}(b_t) = \Lambda_\xi(w_s^t, z^\star_{\min})\cdot\mathcal{J}_{\min}(b_t)\,(1{+}o(1)) = z_t^{\delta_{\min}+o(1)}\;\cdot\mathcal{J}_{\min}(b_t)\,(1{+}o(1)),$$

and combining with Eq (129):

Proposition F.6 (Rate decay preserved, $\alpha < 1$)**.** *When $\alpha < 1$*

$$\mathbb{E}_{\mathcal{G}_{min}}\big[1 - p_y^t\big] \;=\; \mathcal{J}^{\text{eff}}_{min}(b_t) \;=\; \frac{e^{O(1)}}{\big(\tilde{\gamma}^2_{min}\,c_0 + O(C_\xi)\big)\big(1{+}\delta_{min}\big)}\;\cdot\;\frac{1}{z_t} \tag{130}$$

the exact same rate as Eq (41) *in the free-noise case (Theorem D.4), with some $O(1)$ and $e^{O(1)}$ extra-terms who depend on the noise. Especially this confirms the ε-dependency on the rates prefactors.*
The same results are also true for $\mathcal{G}_{maj}$ with the same analysis.

(iv) Preservation of the rates decay (case $\alpha > 1$).

From Eq (122), the raw gap $G_t := a_t - \tilde{\gamma}_{\min}\, b_t$ evolves as

$$\begin{aligned} G_{t+1} - G_t = h_t\Big[(1{-}\varepsilon)(1{+}\mu_A)(1{-}\alpha)\,\mathcal{I}^{\text{eff}}_{\text{maj}}(a_t) \;-\; \varepsilon\big(\tilde{\gamma}_{\min}(1{+}\mu_B) - (1{+}\mu)\big)\,\mathcal{I}^{\text{eff}}_{\min}(b_t)\Big] \\ +\; h_t\,(A\mathbf{v} - \tilde{\gamma}_{\min}\,B\mathbf{v})\cdot\mathbb{E}[q^t\xi]. \end{aligned} \tag{131}$$

Since $\big|\mathbb{E}_g[q^t\xi]\big| \;\le\; R\,\mathcal{J}^{\text{eff}}_g(a_t^g) \;\le\; \frac{R}{\tilde{\gamma}_g}\,\mathcal{I}^{\text{eff}}_g(a_t^g)\,\big(1 + O(1/a_t^g)\big)$ for t large (by Eq (118)), $\mathbb{E}[q^t] = (1{-}\varepsilon)\,\mathcal{J}^{\text{eff}}_{\text{maj}}(a_t){+}\varepsilon\,\mathcal{J}^{\text{eff}}_{\min}(b_t) \le C'(\mathcal{I}^{\text{eff}}_{\text{maj}}(a_t){+}\mathcal{I}^{\text{eff}}_{\min}(b_t))$ for some structural constant $C' > 0$. The noise contribution to (131) is therefore bounded by $C_\xi\,(\mathcal{I}^{\text{eff}}_{\text{maj}} + \mathcal{I}^{\text{eff}}_{\min})$, where $C_\xi := R\,|A\mathbf{v} - \tilde{\gamma}_{\min}B\mathbf{v}|\,C'$.

Both coefficients of $\mathcal{I}^{\text{eff}}_{\text{maj}}$, and $\mathcal{I}^{\text{eff}}_{\min}$ in Eq (131) are non-positive (because $\alpha, \tilde{\gamma}_{\min} \ge 1$), therefore the same argument as in Section E.5 shows $G_t \to -\infty$, since the noise contribution is dominated by the non-positive terms for reasonably small R or C_ξ.

Furthermore, on one hand, from the continuous version of Eq (131), we have

$$\dot{G} \;\le\; \big[(1{-}\varepsilon)(1{+}\mu_A)(1{-}\alpha) + C_\xi\big]\,\mathcal{I}^{\text{eff}}_{\text{maj}}(a), \tag{132}$$

and on another hand, from Eq (122), we also have

$$\dot{a} \;\ge\; \big[(1{-}\varepsilon)(1{+}\mu_A) - C_\xi\big]\,\mathcal{I}^{\text{eff}}_{\text{maj}}(a). \tag{133}$$

Dividing (132) by (133):

$$\frac{\dot{G}}{\dot{a}} \;\le\; \frac{(1{-}\varepsilon)(1{+}\mu_A)(1{-}\alpha) + C_\xi}{(1{-}\varepsilon)(1{+}\mu_A) - C_\xi} \;=:\; 1 - \alpha_\xi \;<\; 0, \tag{134}$$

where

$$\alpha_\xi \;:=\; \frac{(1{-}\varepsilon)(1{+}\mu_A)\,\alpha - 2C_\xi}{(1{-}\varepsilon)(1{+}\mu_A) - C_\xi} \;=\; \alpha + O(C_\xi) \;\underset{R\to 0}{\longrightarrow}\; \alpha \qquad \text{uniformly in } z. \tag{135}$$

Integrating:

$$G(z) \;\le\; (1-\alpha_\xi)\, a(z) + O(1). \tag{136}$$

Finally, since (133) is a scalar inequality of the form $\dot{a} \;\ge\; c_-\, z^{\delta_{\text{maj}}+o(1)}\, \mathcal{I}_{\text{maj}}(a)$ with $c_- := (1-\varepsilon)(1+\mu_A) - C_\xi > 0$ (using $\mathcal{I}_g^{\text{eff}} = z^{\delta_g+o(1)}\, \mathcal{I}_g$ from Lemma F.5 and F.4), Proposition H.5 gives the lower bound:

$$a(z) \;\ge\; \big(1+\delta_{\text{maj}}+o(1)\big)\, \ln z - \ln\ln z + O(1). \tag{137}$$

Defining the effective drift ratio

$$R^{\text{eff}}(z) := \frac{\mathcal{I}_{\text{min}}^{\text{eff}}(b)}{\mathcal{I}_{\text{maj}}^{\text{eff}}(a)}, \tag{138}$$

Eq (121) and Lemma F.5 give

$$R^{\text{eff}}(z) = z^{\delta_{\text{min}}-\delta_{\text{maj}}+o(1)}\; \frac{\tilde{\gamma}_{\text{min}}\, \lambda_{\text{min}}(\tilde{\gamma}_{\text{min}})}{\lambda_{\text{maj}}(1)}\; \frac{a}{b}\; e^{G}\,(1+o(1)). \tag{139}$$

Since $G \le 0$ for z large ($G \to -\infty$), $a/b = O(1)$. Substituting in Eq (139) along with Eq (136) and Eq (137):

$$\begin{aligned} R^{\text{eff}}(z) &\le z^{\delta_{\text{min}}-\delta_{\text{maj}}+o(1)} \;\cdot\; O(1) \;\cdot\; e^{(1-\alpha_\xi)\, a(z)+O(1)} \\ &\le z^{\delta_{\text{min}}-\delta_{\text{maj}}+(1-\alpha_\xi)(1+\delta_{\text{maj}})+o(1)} \;\cdot\; O\Big((\ln z)^{-(1-\alpha_\xi)}\Big) \\ &= z^{1+\delta_{\text{min}}-\alpha_\xi(1+\delta_{\text{maj}})+o(1)} \;\cdot\; O\Big((\ln z)^{-(1-\alpha_\xi)}\Big). \end{aligned} \tag{140}$$

When the noise satisfies $\alpha_\xi > \frac{1+\delta_{\text{min}}}{1+\delta_{\text{maj}}} \ge 1$, the bound forces $R^{\text{eff}} \to 0$ and the dynamics decouple. In this case, the a-equation (Eq (122)) reads

$$a_{t+1} - a_t = h_t\, \big[(1-\varepsilon)(1+\mu_A) + O(1)\big]\, z_t^{\delta_{\text{maj}}+o(1)}\; \frac{\lambda_{\text{maj}}(1)\, e^{-a_t}}{a_t}\,(1 + O(1/a_t)), \tag{141}$$

where the $O(1)$ absorbs $\varepsilon(1+\mu)\, R^{\text{eff}} + O(\mathbb{E}[q^t]/\mathcal{I}_{\text{maj}}^{\text{eff}})$. This is the exact prerequisite of Proposition H.5 which yields

$$a(z) = (1+\delta_{\text{maj}}+o(1))\ln z - \ln\ln z + O(1), \tag{142}$$

$$z^{\delta_{\text{maj}}+o(1)}\, \mathcal{I}_{\text{maj}}\big(a(z)\big) = \frac{1}{\big((1-\varepsilon)(1+\mu_A)+O(1)\big)\big(1+\delta_{\text{maj}}\big)} \;\cdot\; \frac{e^{O(1)}}{z} \;=\; \Theta(1/z). \tag{143}$$

leading to

$$\mathbb{E}_{\mathcal{G}_{\text{maj}}}\big[1-p_y^t\big] = \mathcal{J}_{\text{maj}}^{\text{eff}}(a_t) = \frac{e^{O(1)}}{\big((1-\varepsilon)(1+\mu_A)+O(1)\big)\big(1+\delta_{\text{maj}}\big)} \;\cdot\; \frac{1}{z_t}. \tag{144}$$

Regarding the b-dynamic, the ratio $\dot{b}/\dot{a}$ (Eq (122)), with $R^{\text{eff}} \to 0$ and $\mathbb{E}[q^t] \le C'\, \mathcal{I}_{\text{maj}}^{\text{eff}}(a_t)\,(1+o(1))$:

$$\begin{aligned} \frac{\dot{b}}{\dot{a}} &= \frac{(1-\varepsilon)(1+\mu) + O(R^{\text{eff}}) + O(\mathbb{E}[q^t]/\mathcal{I}_{\text{maj}}^{\text{eff}})}{(1-\varepsilon)(1+\mu_A) + O(R^{\text{eff}}) + O(\mathbb{E}[q^t]/\mathcal{I}_{\text{maj}}^{\text{eff}})} \\ &= \frac{(1-\varepsilon)(1+\mu) + O(C_\xi)}{(1-\varepsilon)(1+\mu_A) + O(C_\xi)} + O(R^{\text{eff}}) \\ &= \frac{\alpha}{\tilde{\gamma}_{\text{min}}} + O(C_\xi) + O(R^{\text{eff}}), \end{aligned} \tag{145}$$

where the $O(C_\xi)$ terms absorb the noise contribution $O(\mathbb{E}[q^t]/\mathcal{I}_{\text{maj}}^{\text{eff}}) = O(R\,|A_g\mathbf{v}|/\tilde{\gamma}_{\text{maj}}) = O(C_\xi)$ and converge to 0 as $R \to 0$ uniformly in t.

From (140) with $\alpha_\xi > \frac{1+\delta_{\text{min}}}{1+\delta_{\text{maj}}}$: $R^{\text{eff}} = O(z^{-p})$ for some $p > 0$, so $\dot{a}\, R^{\text{eff}} = O(z^{-p-1})$. Therefore, integrating Eq (145):

$$\tilde{\gamma}_{\text{min}}\, b(z) \;=\; \big(\alpha + O(C_\xi)\big)\, a(z) + O(1) \;=\; \alpha_\xi\, a(z) + O(1), \tag{146}$$

with α_ξ defined in Eq (135). Substituting (142):

$$\tilde{\gamma}_{\text{min}}\, b(z) = \alpha_\xi \left(1{+}\delta_{\text{maj}}{+}o(1)\right) \ln z - \alpha_\xi \ln\ln z + O(1), \tag{147}$$

$$e^{-\tilde{\gamma}_{\text{min}}\, b_t} = z_t^{-\alpha_\xi(1+\delta_{\text{maj}})+o(1)} \;\cdot\; (\ln z_t)^{\alpha_\xi} \;\cdot\; e^{O(1)} \tag{148}$$

Mixing everything with Eq (116):

$$\mathbb{E}_{\mathcal{G}_{\text{min}}}\big[1{-}p_y^t\big] = \mathcal{J}^{\text{eff}}_{\text{min}}(b_t) = z_t^{\delta_{\text{min}}+o(1)}\, \frac{\lambda_{\text{min}}(\tilde{\gamma}_{\text{min}})\, e^{-\tilde{\gamma}_{\text{min}}\, b_t}}{b_t} \tag{149}$$

$$= \; z_t^{\delta_{\text{min}}-\alpha_\xi(1+\delta_{\text{maj}})+o(1)}\, (\ln z_t)^{\alpha_\xi - 1} \cdot \Theta(1)$$

$$= \Theta\left(\frac{(\ln z_t)^{\alpha_\xi - 1}}{z_t^{\alpha_\xi(1+\delta_{\text{maj}})-\delta_{\text{min}}+o(1)}} \right). \tag{150}$$

Proposition F.7 (Rate decay, $\alpha_\xi > \frac{1+\delta_{\text{min}}}{1+\delta_{\text{maj}}}$)**.** *When $\alpha_\xi > \frac{1+\delta_{min}+o(1)}{1+\delta_{maj}} \geq 1$ with α_ξ defined in Eq* (135)*, we obtain the rates Eq* (144) *and Eq* (150)*. They present the same rates as in the noise-free case (Theorem D.4)-part(ii), with the $O(1)$ and $o(1)$ extra-terms depending on the noise ξ. Especially when $R = 0$, we recover the exact noise-free rates.*

G GENERALISATION FAILURE UNDER CORRELATION SHIFT

Under ERM, the classifier places excessive weight on the spurious component, making it vulnerable to any shift in the spurious correlation structure between training and test. This naturally extends to our *Expected* Risk Minimization setup by defining a continuous test distribution $\mathcal{D}_{\text{test}}$ with a shifted spurious correlation:

$$\tilde{\gamma}_{\text{test}} := \operatorname*{ess\,inf}_{\mathbf{r}|\mathcal{D}_{\text{test}}} y\,\mathbf{v}\cdot\mathbf{r}, \qquad \mathbf{s}\mid\mathbf{r} \;\sim\; C\mathbf{r}+\xi, \qquad \text{under } \mathcal{D}_{\text{test}}, \tag{151}$$

where $C \in \mathbb{R}^{d_s\times d_r}$ is an arbitrary spurious alignment operator. Define the *projected test weight* and test margins: writing $\hat{w} = [\hat{w}_r \quad \hat{w}_s]$ (see Eq (2)),

$$\hat{\chi} := \hat{w}_r + C^\top \hat{w}_s, \quad \gamma_{\text{test}}(C) := \operatorname*{ess\,inf}_{(\mathbf{r},\xi)\sim\mathcal{D}_{\text{test}}} \big(\hat{\chi}\cdot\mathbf{r}+\hat{w}_s\cdot\xi\big), \quad \delta_{\text{test}} := \delta_+(\tilde{\gamma}_{\text{test}}), \quad \tilde{\delta}_{\text{test}} := \tilde{\delta}_+(\tilde{\gamma}_{\text{test}}), \tag{152}$$

with $\delta_+, \tilde{\delta}_+$ defined in Eqs (112) and (165). We formulate these following generalisation results:

Corollary G.1 (Accuracy ceiling under correlation shift)**.** *Assume the infimum in* (152) *is achieved at the boundary* $Z = \tilde{\gamma}_{test}$ *:*

$$\gamma_{test} \;=\; \gamma^{(r)}_{test} - \;\delta_{test}, \qquad \textit{with} \quad \gamma^{(r)}_{test} := \operatorname*{ess\,inf}_{\boldsymbol{r}|Z=\tilde{\gamma}_{test}} \{\hat{\chi}\cdot\boldsymbol{r}\}. \tag{153}$$

We have the following cases:
(i) If $\gamma_{test}(C) > 0$ (test group still separated by $\hat{w}$)*, then*

$$\mathbb{E}_{\mathcal{D}_{test}}\big[1-p_y(\boldsymbol{x},w^t)\big] \;=\; \Theta\left(\frac{(\log z_t)^{\gamma_{test}+\delta_{test}+\tilde{\delta}_{test}-1}}{z_t^{\gamma_{test}+o(1)}}\right), \tag{154}$$

where the $o(1)$ *is uniformly noise dependent. The test group error still vanishes, but at a rate governed by* $\gamma_{test}(C)$*.*
(ii) If $\gamma_{test}(C) < 0$ (test group not separated by $\hat{w}$)*:*

$$\lim_{t\to\infty} \mathbb{E}_{\mathcal{D}_{test}}\big[1-p_y(\boldsymbol{x},w^t)\big] \;=\; P_{\mathcal{D}_{test}}\big(\hat{\chi}\cdot\boldsymbol{r}+\hat{w}_s\cdot\xi<0\big) \;>\; 0. \tag{155}$$

Part (ii) reveals a non-vanishing error floor caused by the correlation shift.

Proof. The test logit at time t is: $\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi$ with $\chi_t := w_r^t + C^\top w_s^t$ and by Proposition 5.4,

$$\chi_t = (\log z_t - \log\log z_t)\,\hat{\chi} + O(1), \qquad w_s^t = (\log z_t - \log\log z_t)\,\hat{\mathbf{w}}_s + O(1). \tag{156}$$

The test error is:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_{\text{test}}}[1-p_y^t] &= \mathbb{E}\big[\sigma\big(-(\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi)\big)\big] = \mathbb{E}_Z\Big[\mathbb{E}_{\mathbf{r}_\perp,\xi|Z}\big[\sigma\big(-\chi_t\cdot\mathbf{r}-w_s^t\cdot\xi\big)\big]\Big]\\ &= \int_{\tilde{\gamma}_{\text{test}}}^{K} \mathbb{E}_{\mathbf{r}_\perp,\xi|Z=z}[\sigma(-\chi_t\cdot\mathbf{r}-w_s^t\cdot\xi)]\,\lambda_{\text{test}}(z)\,dz\end{aligned} \tag{157}$$

Proof of (i). Assume $\gamma_{\text{test}} > 0$.

By Eq (152), $\hat{\chi}\cdot\mathbf{r}+\hat{\mathbf{w}}_s\cdot\xi \geq \gamma_{\text{test}} > 0$ a.s., so with Eq (156):

$$\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi \;=\; (\hat{\chi}\cdot\mathbf{r}+\hat{\mathbf{w}}_s\cdot\xi)(\log z_t-\log\log z_t)+O(1) \;\geq\; \gamma_{\text{test}}\,(\log z_t-\log\log z_t)+O(1) \;>\; 0 \tag{158}$$

a.s. for t large. Using the identity $\sigma(-v) = e^{-v}/(1+e^{-v})$ for $v>0$, we write pointwise:

$$\sigma\big(-(\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi)\big) = e^{-(\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi)}\cdot\frac{1}{1+e^{-(\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi)}} = e^{-\chi_t\cdot\mathbf{r}}\,e^{-w_s^t\cdot\xi}\,\tilde{h}_t(\mathbf{r},\xi), \tag{159}$$

where $\tilde{h}_t(\mathbf{r},\xi) := 1/(1+e^{-(\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi)}) \in (1/2,\,1)$ since $\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi > 0$. Therefore

$$\sigma\big(-(\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi)\big) \;=\; \Theta\big(e^{-\chi_t\cdot\mathbf{r}}\,e^{-w_s^t\cdot\xi}\big) \qquad \text{a.s. for } t \text{ large, uniformly in } \mathbf{r},\xi. \tag{160}$$

By Eq (160), the integrand in Eq (157) satisfies

$$\begin{aligned}
\mathbb{E}_{\mathbf{r}_\perp,\xi|Z=z}[\sigma(-\chi_t\cdot\mathbf{r}-w_s^t\cdot\xi)]\,\lambda_{\text{test}}(z) &= \lambda_{\text{test}}(z)\,\mathbb{E}_{\mathbf{r}_\perp,\xi|Z=z}\big[\Theta\big(e^{-\chi_t\cdot\mathbf{r}}\,e^{-w_s^t\cdot\xi}\big)\big]\\
&= \Lambda_\xi(w_s^t,z)\,\lambda_{\text{test}}(z)\,\mathbb{E}_{\mathbf{r}_\perp|Z=z}\big[\Theta\big(e^{-\chi_t\cdot\mathbf{r}}\big)\big] \qquad \text{(see Eq (100))}\\
&= e^{-z\,\chi_t\cdot\mathbf{v}}\,\Lambda_\xi(w_s^t,z)\,\lambda_{\text{test}}(z)\,\Theta\big(\mathbb{E}_{\mathbf{r}_\perp|Z=z}[e^{-\chi_t\cdot\mathbf{r}_\perp}]\big)\ \text{(all }\Theta(...)\text{ are uniform in }\mathbf{r},z)\\
&= \Theta\big(e^{-z\,\chi_t\cdot\mathbf{v}}\,\Lambda_\xi(w_s^t,z)\,\Lambda_{\mathbf{r}_\perp}(\chi^t,z)\,\lambda_{\text{test}}(z)\big) \qquad \text{(with Eq (156))},
\end{aligned} \tag{161}$$

where

$$\Lambda_{\mathbf{r}_\perp}(w_r,z) := \mathbb{E}_{\mathbf{r}_\perp|Z=z}[e^{-w_r\cdot\mathbf{r}_\perp}]. \tag{162}$$

Therefore, Eq (157) rewrites thanks to the mean-value theorem:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{\text{test}}}[1-p_y^t] &= \Theta(1)\cdot\int_{\tilde\gamma_{\text{test}}}^K e^{-z\,\chi_t\cdot\mathbf{v}}\,\Lambda_\xi(w_s^t,z)\,\Lambda_{\mathbf{r}_\perp}(\chi^t,z)\,\lambda_{\text{test}}(z)\,dz\\
&= \Theta(1)\,\Lambda_\xi(w_s^t,z_{\text{test}}^\star)\,\Lambda_{\mathbf{r}_\perp}(\chi^t,z_{\text{test}}^\star)\cdot\int_{\tilde\gamma_{\text{test}}}^K e^{-z\,\chi_t\cdot\mathbf{v}}\,\lambda_{\text{test}}(z)\,dz,
\end{aligned} \tag{163}$$

for some t- dependent $z_{\text{test}}^\star\in[\tilde\gamma_{\text{test}},K]$. By Lemma F.4:

$$\Lambda_\xi(w_s^t,z_{\text{test}}^\star) = z_t^{\delta_+(z_{\text{test}}^\star)+o(1)},\qquad \Lambda_{\mathbf{r}_\perp}(\chi^t,z_{\text{test}}^\star) = z_t^{\tilde\delta_+(z_{\text{test}}^\star)+o(1)}, \tag{164}$$

where $\tilde\delta_+$ is defined exactly as δ_+ in Eq (112) and has the same properties (due to separability $\gamma_{\text{test}}>0$):

$$\tilde\delta_+(z) \;:=\; \sup_{r_\perp\in\text{supp}(\mathbf{r}_\perp|Z=z)} -\hat\chi\cdot r_\perp \tag{165}$$

Since the Laplace weight $e^{-z\,\chi_t\cdot\mathbf{v}} = e^{-z\,(\log z_t-\log\log z_t)\,\hat\chi\cdot\mathbf{v}+O(1)}$ concentrates at $z=\tilde\gamma_{\text{test}}$, the mean value theorem points satisfy $z_{\text{test}}^\star\to\tilde\gamma_{\text{test}}$, hence $\delta_+(z_{\text{test}}^\star)\to\delta_{\text{test}}$ and $\tilde\delta_+(z_{\text{test}}^\star)\to\tilde\delta_+(\tilde\gamma_{\text{test}})=\tilde\delta_{\text{test}}$.

For the remaining integral, a similar Watson Lemma argument as in Lemma H.1 gives:

$$\begin{aligned}
\int_{\tilde\gamma_{\text{test}}}^K e^{-z\,\chi_t\cdot\mathbf{v}}\,\lambda_{\text{test}}(z)\,dz \;&\sim\; \frac{\lambda_{\text{test}}(\tilde\gamma_{\text{test}})\,e^{-\tilde\gamma_{\text{test}}\,\chi_t\cdot\mathbf{v}}}{\chi_t\cdot\mathbf{v}}\\
&\sim\; \frac{\lambda_{\text{test}}(\tilde\gamma_{\text{test}})\,z_t^{-\tilde\gamma_{\text{test}}\hat\chi\cdot\mathbf{v}}\,(\log z_t)^{\tilde\gamma_{\text{test}}\hat\chi\cdot\mathbf{v}}\,e^{O(1)}}{\hat\chi\cdot\mathbf{v}\,\log z_t}\\
&=\; \Theta\Big(z_t^{-\tilde\gamma_{\text{test}}\hat\chi\cdot\mathbf{v}}\,(\log z_t)^{\tilde\gamma_{\text{test}}\hat\chi\cdot\mathbf{v}-1}\Big)
\end{aligned} \tag{166}$$

But the Assumption (153) reads,

$$\begin{aligned}
\gamma_{\text{test}}^{(r)} &= \operatorname*{ess\,inf}_{\mathbf{r}|Z=\tilde\gamma_{\text{test}}}\{\hat\chi\cdot\mathbf{r}\} = \tilde\gamma_{\text{test}}\,\hat\chi\cdot\mathbf{v} + \operatorname*{ess\,inf}_{\mathbf{r}_\perp|Z=\tilde\gamma_{\text{test}}}\{\hat\chi\cdot\mathbf{r}_\perp\} = \tilde\gamma_{\text{test}}\,\hat\chi\cdot\mathbf{v}-\tilde\delta_{\text{test}},\\
\gamma_{\text{test}} &= \gamma_{\text{test}}^{(r)}-\delta_{\text{test}} = \tilde\gamma_{\text{test}}\,\hat\chi\cdot\mathbf{v}-\tilde\delta_{\text{test}}-\delta_{\text{test}}.
\end{aligned} \tag{167}$$

Therefore combining all factors:

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_{\text{test}}}[1-p_y^t] &= \Theta(1)\cdot z_t^{\delta_{\text{test}}+\tilde\delta_{\text{test}}+o(1)}\cdot z_t^{-\tilde\gamma_{\text{test}}\hat\chi\cdot\mathbf{v}}\,(\log z_t)^{\tilde\gamma_{\text{test}}\hat\chi\cdot\mathbf{v}-1}\\
&= \Theta(1)\cdot z_t^{-\gamma_{\text{test}}+o(1)}\cdot(\log z_t)^{\tilde\gamma_{\text{test}}\hat\chi\cdot\mathbf{v}-1}\\
&= \Theta\left(\frac{(\log z_t)^{\gamma_{\text{test}}+\delta_{\text{test}}+\tilde\delta_{\text{test}}-1}}{z_t^{\gamma_{\text{test}}+o(1)}}\right).
\end{aligned} \tag{168}$$

Proof of (ii). Assume $\gamma_{\text{test}}<0$.

Writing the test-time logit $\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi \;=\; (\hat\chi\cdot\mathbf{r}+\hat{\mathbf{w}}_s\cdot\xi)(\log z_t-\log\log z_t)+O(1)$, we have:

On $\{\hat\chi\cdot\mathbf{r}+\hat{\mathbf{w}}_s\cdot\xi>0\}$: $\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi\to+\infty$, hence $\sigma\big(-(\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi)\big)\to 0$.

On $\{\hat\chi\cdot\mathbf{r}+\hat{\mathbf{w}}_s\cdot\xi<0\}$: $\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi\to-\infty$, hence $\sigma\big(-(\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi)\big)\to 1$.

Since $\gamma_{\text{test}}=\operatorname{ess\,inf}\hat\chi\cdot\mathbf{r}+\hat{\mathbf{w}}_s\cdot\xi<0$, the set $\{\hat\chi\cdot\mathbf{r}+\hat{\mathbf{w}}_s\cdot\xi<0\}$ has positive probability. By dominated convergence:

$$\lim_{t\to\infty}\mathbb{E}[\sigma\big(-(\chi_t\cdot\mathbf{r}+w_s^t\cdot\xi)\big)] = P(\hat\chi\cdot\mathbf{r}+\hat{\mathbf{w}}_s\cdot\xi<0)>0. \tag{169}$$

□

G.1 Application to the Proof of Theorem 5.1

A natural application of the analysis presented above is the proof of Theorem 5.1.

The isotropic regime in Theorem 5.3 was mathematically load-bearing because it pushed the parameter updates to remain confined to $\mathrm{Span}\{\mathbf{v}\}$ (Proposition E.2), facilitating the reduction of the problem into a 2D-ODE (58). When this condition fails, the gradient continuously injects a perpendicular component into ψ_t, ϕ_t, physically rotating the parameters out of $\mathrm{Span}\{\mathbf{v}\}$. This removes the possibility to reduce the problem to its projection on $\mathrm{Span}\{\mathbf{v}\}$ utilized in Proposition E.3.

Recall the theorem describing learning dynamics in this general regime:

Theorem G.2 (General Decay Rates)**.** *Assume the infimums in* (20) *are achieved at the boundaries $Z = \tilde{\gamma}_{maj}$, and $Z = \tilde{\gamma}_{min}$, then*

$$\mathbb{E}_{\mathcal{G}_{maj}}\big[1 - p_y(\boldsymbol{x}, w^t)\big] = \Theta\left(\frac{(\log z_t)^{\gamma_{maj}+\delta_{maj}+\tilde{\delta}_{maj}-1}}{z_t^{\gamma_{maj}+o(1)}}\right), \tag{170}$$

$$\mathbb{E}_{\mathcal{G}_{min}}\big[1 - p_y(\boldsymbol{x}, w^t)\big] = \Theta\left(\frac{(\log z_t)^{\gamma_{min}+\delta_{min}+\tilde{\delta}_{min}-1}}{z_t^{\gamma_{min}+o(1)}}\right), \tag{171}$$

where γ_g are the group-specific hard-margin defined in Eq (20) *and the $o(1)$ are uniformly noise-dependent.*

Proof. The proof follows directly from part (i) of Corollary G.1 with $\mathcal{D}_{\text{test}} = \mathcal{G}_{\text{maj}}$ and $\mathcal{D}_{\text{test}} = \mathcal{G}_{\text{min}}$ respectively, and with the $\mathbf{r}_\perp$-boundary exponents for each group:

$$\tilde{\delta}_g \;:=\; \tilde{\delta}_+(\tilde{\gamma}_g), \qquad g \in \{\text{maj}, \text{min}\}. \tag{172}$$

□

H TOOLBOX

H.1 ASYMPTOTIC BEHAVIOR OF $\mathcal{J}$ AND $\mathcal{I}$

Lemma H.1. *Let λ be a continuous probability density function on $[\tilde{\gamma}, K]$ with $\lambda(\tilde{\gamma}) > 0$, and let $\mathcal{J}(x) := \int_{\tilde{\gamma}}^{K} \frac{\lambda(z)}{1+e^{xz}}\,\mathrm{d}z$. Then,*

$$\mathcal{J}(x) \sim \frac{\lambda(\tilde{\gamma})\, e^{-\tilde{\gamma}\, x}}{x} \qquad \textit{and} \qquad \mathcal{J}'(x) \sim -\frac{\tilde{\gamma}\, \lambda(\tilde{\gamma})\, e^{-\tilde{\gamma}\, x}}{x} \qquad \textit{as } x \to +\infty. \tag{173}$$

Proof. We proceed in two steps. For $x > 0$,

$$\begin{aligned}\left| \mathcal{J}(x) - \int_{\tilde{\gamma}}^{K} e^{-xz} \lambda(z)\,\mathrm{d}z \right| &= \left| -\int_{\tilde{\gamma}}^{K} \lambda(z)\, \frac{e^{-xz}}{1+e^{xz}}\,\mathrm{d}z \right| \\ &\le e^{-\tilde{\gamma}\, x} \int_{\tilde{\gamma}}^{K} \frac{\lambda(z)}{1+e^{xz}}\,\mathrm{d}z \le e^{-\tilde{\gamma}\, x} \int_{\tilde{\gamma}}^{K} e^{-xz}\lambda(z)\,\mathrm{d}z \\ &= o\left(\int_{\tilde{\gamma}}^{K} e^{-xz}\lambda(z)\,\mathrm{d}z \right) \quad \text{as } x \to +\infty,\end{aligned}$$

Hence $\mathcal{J}(x) \sim \int_{\tilde{\gamma}}^{K} e^{-xz}\lambda(z)\,\mathrm{d}z$. By Watson's Lemma Wong (2001, *Chapter 1*):

$$\int_{\tilde{\gamma}}^{K} e^{-xz}\lambda(z)\,\mathrm{d}z \sim \frac{\lambda(\tilde{\gamma})\, e^{-\tilde{\gamma}\, x}}{x} \qquad \text{as } x \to +\infty.$$

Combining everything yields $\mathcal{J}(x) \sim \lambda(\tilde{\gamma})\, e^{-\tilde{\gamma}\, x}/x$.

We use the same arguments for $\mathcal{J}'(x) = -\int_{\tilde{\gamma}}^{K} \frac{z\,\lambda(z)e^{xz}}{(1+e^{xz})^2}\mathrm{d}z \sim -\int_{\tilde{\gamma}}^{K} z\,\lambda(z)e^{-xz}\mathrm{d}z$, which yields $\mathcal{J}'(x) \sim -\tilde{\gamma}\,\lambda(\tilde{\gamma})\, e^{-\tilde{\gamma}\, x}/x$. □

Lemma H.2. *Assume $\lambda \in C^1([\gamma, K])$ with $\lambda(\gamma) > 0$. Define $\mathcal{I}(x) := \int_{\gamma}^{K} \frac{z\lambda(z)}{1+e^{xz}}\,\mathrm{d}z$, then*

$$\mathcal{I}(x) \;=\; \frac{\gamma\,\lambda(\gamma)\, e^{-\gamma x}}{x}\left[1 + \frac{\nu}{x} + O\left(\frac{1}{x^2}\right)\right] \qquad \textit{as } x \to +\infty, \tag{174}$$

where $\nu := 1/\gamma + \lambda'(\gamma)/\lambda(\gamma)$.
In particular, writing $r(x) := \mathcal{I}(x)\big/(\gamma\,\lambda(\gamma)\, e^{-\gamma x}/x) - 1$, one has $|r(x)| = O(1/x)$.

Proof. We reuse the same approach as in Lemma H.1:

$$\begin{aligned}\left| \mathcal{I}(x) - \int_{\gamma}^{K} z\,\lambda(z)e^{-xz}\,\mathrm{d}z \right| &= \left| \int_{\gamma}^{K} z\,\lambda(z)\, \frac{e^{-xz}}{1+e^{xz}}\,\mathrm{d}z \right| \\ &\le e^{-2\gamma x} \int_{1}^{K} z\,\lambda(z)\,\mathrm{d}z = O(e^{-2\gamma x}) \quad \text{as } x \to +\infty.\end{aligned}$$

Hence as $x \to +\infty$,

$$\mathcal{I}(x) = \int_{\gamma}^{K} z\,\lambda(z)e^{-xz}\,\mathrm{d}z + O(e^{-2\gamma x}) = \int_{\gamma}^{K} z\,\lambda(z)e^{-xz}\,\mathrm{d}z + O\Big(\frac{e^{-\gamma x}}{|x|^3}\Big) \tag{175}$$

.

Now, substitute $z = \gamma + t/x$ in $\int_{\gamma}^{K} \phi(z)e^{-xz}\,\mathrm{d}z$ with $\phi(z) := z\,\lambda(z)$, giving

$$\int_{\gamma}^{K} \phi(z) e^{-xz}\,\mathrm{d}z = \frac{e^{-\gamma x}}{x} \int_{0}^{x(K-\gamma)} e^{-t}\,\phi\big(\gamma + \tfrac{t}{x}\big)\,\mathrm{d}t.$$

Taylor-expanding $\phi(\gamma + s) = \phi(\gamma) + \phi'(\gamma)\,s + \epsilon(s)\,s^2$ with ϵ bounded on $[0, K-\gamma]$ by $\|\epsilon\|_\infty$. Integrating term by term then yields

$$\int_{0}^{x(K-\gamma)} e^{-t}\,\phi(\gamma)\,\mathrm{d}t = \phi(\gamma) + O\big(e^{-x(K-\gamma)}\big) = \phi(\gamma) + O\Big(\frac{1}{x^2}\Big)$$

$$\int_{0}^{x(K-\gamma)} e^{-t}\,\frac{t}{x}\,\phi'(\gamma)\mathrm{d}t = \frac{\phi'(\gamma)}{x} + O\big(e^{-x(K-\gamma)}\big) = \frac{\phi'(\gamma)}{x} + O\Big(\frac{1}{x^2}\Big)$$

$$\left| \int_{0}^{x(K-\gamma)} e^{-t}\,\epsilon\Big(\frac{t}{x}\Big)\frac{t^2}{x^2}\mathrm{d}t \right| \le \frac{\|\epsilon\|_\infty}{x^2} + O\big(e^{-x(K-\gamma)}\big) = O\Big(\frac{1}{x^2}\Big).$$

Finally,

$$\int_{\gamma}^{K} \phi(z) e^{-xz}\,\mathrm{d}z = \frac{e^{-\gamma x}}{x}\left(\phi(\gamma) + \frac{\phi'(\gamma)}{x} + O\Big(\frac{1}{x^2}\Big)\right),$$

with $\phi(\gamma) = \gamma\lambda(\gamma)$, $\phi'(\gamma) = \lambda(\gamma) + \gamma\,\lambda'(\gamma)$. We then get

$$\mathcal{I}(x) = \frac{\gamma\,\lambda(\gamma)\,e^{-\gamma x}}{x} + \frac{\big(\lambda(\gamma) + \gamma\,\lambda'(\gamma)\big)\,e^{-\gamma x}}{x^2} + O\Big(\frac{e^{-\gamma x}}{x^3}\Big),$$

which rearranges into (174). □

H.2 Asymptotic Behavior of the Solution of 1D ODEs

Proposition H.3 (Non-autonomous ODE). *Let $\mathcal{I} \in C(\mathbb{R}^+, \mathbb{R}_+^\star)$ satisfy $\mathcal{I}(x) \sim \gamma\lambda_0\,e^{-\gamma x}/x$ as $x \to +\infty$ for some $\lambda_0 > 0$ and $\gamma > 0$. Let $c : [z_0, \infty) \to \mathbb{R}$ be continuous and assume there exist $0 < c_- \le c_+ < \infty$ such that $c_- \le c(z) \le c_+$ for all $z \ge z_0$. Then any solution of the Cauchy problem*

$$\dot v(z) \;=\; c(z)\,\mathcal{I}\big(v(z)\big), \qquad v(z_0) \in \mathbb{R}^+,$$

with $v(z_0)$ sufficiently large, satisfies $v(z) \to +\infty$ monotonically and

$$v(z) \;=\; \frac{1}{\gamma}\Big(\ln z - \ln\ln z + O(1)\Big) \qquad \text{as } z \to +\infty. \tag{176}$$

Consequently $\mathcal{I}(v(z)) = \Theta(1/z)$: there exist constants $0 < M_- \le M_+ < \infty$ and $z_1 \ge z_0$ such that

$$\frac{M_-}{z} \;\le\; \mathcal{I}\big(v(z)\big) \;\le\; \frac{M_+}{z} \qquad \text{for all } z \ge z_1.$$

Proof. Since $\mathcal{I}(v) > 0$ and $c(z) > 0$, $\dot v(z) > 0$, hence $v(\cdot)$ is strictly increasing and $v(z) \to +\infty$ (monotonicity rules out any finite limit because $\mathcal{I}(l) > 0$ at any finite $l > 0$).

Let $F(x)$ be a primitive of $1/\mathcal{I}$ on $\mathbb{R}^+$. Separating variables,

$$F\big(v(z)\big) - F\big(v(z_0)\big) \;=\; \int_{z_0}^{z} c(s)\,\mathrm{d}s. \tag{177}$$

Given that, $1/\mathcal{I}(x) \sim C\,x\,e^{\gamma x}$ as $x \to +\infty$ with $C = 1/(\gamma\lambda_0)$, and since $x\,e^{\gamma x}$ is non-integrable on $(0, \infty)$ the theorem of integration of asymptotic equivalents (Jourdain, 2018, *Theorem 4.11*) yields

$$F(x) \;\sim\; C\,x\,e^{\gamma x}/\gamma \qquad \text{as } x \to +\infty. \tag{178}$$

Meanwhile, the hypothesis $c_- \le c \le c_+$ gives that the right-hand side of (177) is a $\Theta(z)$. Combining with (178),

$$C\,v(z)\,e^{\gamma\,v(z)}\,(1/\gamma + o(1)) = \Theta(z), \qquad \text{i.e.} \qquad \gamma\,v(z)\,e^{\gamma\,v(z)} = \Theta(z).$$

This means $\gamma\, v e^{\gamma v} \in [\alpha z,\, \beta z]$ (for some $0 < \alpha \le \beta$ and z large). We apply the principal real branch of the Lambert W function (Visser, 2018, *Eq. A2*): $\gamma\, v(z) \in [W(\alpha z),\, W(\beta z)]$ with $W(y) = \ln y - \ln\ln y + O(1)$ as $y \to +\infty$. Hence

$$v(z) \;=\; \frac{1}{\gamma}\Big(\ln z - \ln\ln z + O(1)\Big),$$

which proves (176).

Plugging into $\mathcal{I}(x) = \gamma\, \lambda_0\, e^{-\gamma x} x^{-1}\,(1 + o(1))$ gives

$$\begin{aligned}\mathcal{I}(v(z)) &= \frac{\gamma\, \lambda_0\, e^{-\gamma\, v(z)}}{v(z)}\,\big(1 + o(1)\big) \\ &= \frac{\gamma^2\, \lambda_0\, e^{-\ln z + \ln\ln z + O(1)}}{\ln z\big(1 + o(1)\big)}\,\big(1 + o(1)\big) \\ &= \frac{\gamma^2\, \lambda_0}{z} e^{O(1)}\big(1 + o(1)\big) = \Theta(1/z),\end{aligned}$$

which gives the stated $M_\pm$ bounds. □

Proposition H.4 (Perturbed autonomous ODE)**.** *Let $c_0 > 0$, let $\mathcal{I}$ be a continuous function satisfying $\mathcal{I}(x) \sim \gamma\, \lambda_0\, e^{-\gamma x}/x$ as $x \to +\infty$ for some $\lambda_0 > 0$, $\gamma > 0$, and let $\delta\colon [0,\infty) \to \mathbb{R}$ be a continuous function with $\delta(z) \to 0$ as $z \to \infty$.*
The solution v of the perturbed *Cauchy problem*

$$v'(z) = c_0\, \mathcal{I}\big(v(z)\big)\,\big(1 + \delta(z)\big), \qquad v(0) = v_0, \tag{179}$$

satisfies the asymptotic expansion:

$$\gamma\, v(z) = \ln z - \ln\ln z + \ln(\gamma^3\, c_0\, \lambda_0) + o(1) \qquad \text{as } z \to +\infty. \tag{180}$$

Proof. We proceed step by step.

Dividing both sides of (179) by $c_0\, \mathcal{I}(v)$ and integrating from 0 to z:

$$F\big(v(z)\big) - F(v_0) = z + \underbrace{\int_0^z \delta(s)\,\mathrm{d}s}_{=:\, D(z)}, \tag{181}$$

where F is any primitive of $1/(c_0\, \mathcal{I})$. Since $\delta(z) \to 0$ as $z \to \infty$, the Cesàro mean converges to the same limit:

$$\frac{D(z)}{z} = \frac{1}{z}\int_0^z \delta(s)\,\mathrm{d}s \;\xrightarrow{z\to\infty}\; 0,$$

i.e. $D(z) = o(z)$, and Eq (181) becomes

$$F\big(v(z)\big) = z\,\big(1 + o(1)\big) + F(v_0). \tag{182}$$

By the hypothesis $\mathcal{I}(x) \sim \gamma\lambda_0\, e^{-\gamma x}/x$ and the integration-of-equivalents theorem (Jourdain, 2018, *Theorem 4.11*) (the integrand $x\, e^{\gamma x}$ is not integrable on $[0, +\infty)$, so the theorem applies):

$$F(x) = \int^x \frac{\mathrm{d}u}{c_0\, \mathcal{I}(u)} \sim C\int^x u\, e^{\gamma u}\,\mathrm{d}u = \frac{C}{\gamma^2}\, e^{\gamma x}(\gamma x - 1)\,, \qquad C := \frac{1}{c_0\, \gamma\, \lambda_0}, \tag{183}$$

Since $v(z) \to +\infty$ (by the same argument as in Proposition H.3), combining Eq (182) with Eq (183):

$$\begin{aligned}\frac{C}{\gamma^2}\,\gamma\, v\, e^{\gamma v}\,(1 + o(1)) &= z\,(1 + o(1)) \\ \gamma\, v\, e^{\gamma v} &= \frac{\gamma^2\, z}{C}\,(1 + o(1)) = \gamma^3\, c_0\, \lambda_0\, z\,(1 + o(1)).\end{aligned} \tag{184}$$

Using the principal real branch of the Lambert W function (Visser, 2018, *Eq. A2*):

$$\begin{aligned}\gamma\, v(z) &= W\big(\gamma^3\, c_0\, \lambda_0\, z\,(1+o(1))\big)\\ &= \ln\big(\gamma^3\, c_0\, \lambda_0\, z\big) - \ln\ln\big(\gamma^3\, c_0\, \lambda_0\, z\big) + o(1)\\ &= \ln z - \ln\ln z + \ln(\gamma^3\, c_0\, \lambda_0) + o(1).\end{aligned}$$

□

Proposition H.5 (Non-autonomous ODE with polynomial factor)**.** *Let $\mathcal{I} \in C(\mathbb{R}^+, \mathbb{R}_+^\star)$ satisfy $\mathcal{I}(x) \sim \gamma\lambda_0\, e^{-\gamma x}/x$ as $x \to +\infty$ for some $\lambda_0 > 0$ and $\gamma > 0$. Let $\delta \geq 0$, let $\epsilon(z) = o(1)$, and let $c : [z_0, \infty) \to \mathbb{R}$ be continuous with $c(z) = c_0 + O(1)$, where $0 < c_- \leq c(z) \leq c_+$ for all $z \geq z_0$. Then the solution of*

$$\dot v(z) \;=\; c(z)\, z^{\delta+\epsilon(z)}\, \mathcal{I}\big(v(z)\big), \qquad v(z_0) \in \mathbb{R}^+,$$

with $v(z_0)$ large enough, satisfies

$$v(z) \;=\; \frac{1}{\gamma}\Big(\big(1{+}\delta{+}\epsilon(z)\big)\ln z - \ln\ln z + \ln\big(\gamma^3\lambda_0\, c_0 + O(1)\big) + O(1)\Big) \qquad \text{as } z \to +\infty, \tag{185}$$

where both $O(1)$ are independent on c_0, $O(1)$ is proportional to the input $O(1) = c(z) - c_0$, and

$$z^{\delta+\epsilon(z)}\, \mathcal{I}\big(v(z)\big) \;=\; \frac{1}{\big(\gamma\, c_0 + O(1)\big)\big(1{+}\delta\big)} \cdot \frac{e^{O(1)}}{z} \;=\; \Theta(1/z), \tag{186}$$

with both $O(1)$ independent on c_0, and $O(1)$ is proportional to the input $O(1) = c(z) - c_0$.

Proof. Let $G(z) := \int_{z_0}^{z} c(s)\, s^{\delta+\epsilon(s)}\, \mathrm{d}s$. Separating variables: $F\big(v(z)\big) - F\big(v(z_0)\big) = G(z)$ with F as in Proposition H.3.

Since $c_- \leq c \leq c_+$:

$$c_- \int_{z_0}^{z} s^{\delta+\epsilon(s)}\, \mathrm{d}s \;\leq\; G(z) \;\leq\; c_+ \int_{z_0}^{z} s^{\delta+\epsilon(s)}\, \mathrm{d}s.$$

By Karamata's theorem (Feller et al., 1971, *Chapter VIII, Section 9, Theorem 1*) (applicable since $\delta \geq 0 > -1$ and with ϵ decreasing fast enough):

$$\int_{z_0}^{z} s^{\delta+\epsilon(s)}\, \mathrm{d}s \;\sim\; \frac{z^{1+\delta+\epsilon(z)}}{1+\delta} \qquad \text{as } z \to +\infty. \tag{187}$$

Therefore $G(z) = z^{1+\delta+\epsilon(z)}\big(c_0 + O(1)\big)$. From $F(v) \sim v\, e^{\gamma v}/(\gamma^2\lambda_0)$ (see Proposition H.3) and $F(v(z)) \sim G(z)$:

$$v(z)\, e^{\gamma v(z)} \;=\; \gamma^2\lambda_0\, G(z)\,(1{+}o(1)) \;=\; z^{1+\delta+\epsilon(z)}\,\big(\gamma^2\lambda_0\, c_0 + O(1)\big).$$

Applying the Lambert W function:

$$\begin{aligned}\gamma\, v(z) &= W\Big(z^{1+\delta+\epsilon(z)}\,\big(\gamma^3\lambda_0\, c_0 + O(1)\big)\Big)\\ &= \ln\Big(z^{1+\delta+\epsilon(z)}\,\big(\gamma^3\lambda_0\, c_0 + O(1)\big)\Big) - \ln\ln\Big(z^{1+\delta+\epsilon(z)}\,\big(\gamma^3\lambda_0\, c_0 + O(1)\big)\Big) + O(1)\\ &= \big(1{+}\delta{+}\epsilon(z)\big)\ln z + \ln\big(\gamma^3\lambda_0\, c_0 + O(1)\big) - \ln\Big(\big(1{+}\delta{+}\epsilon(z)\big)\ln z + O(1)\Big) + O(1)\\ &= \big(1{+}\delta{+}\epsilon(z)\big)\ln z - \ln\ln z + \ln\big(\gamma^3\lambda_0\, c_0 + O(1)\big) + O(1),\end{aligned}$$

giving Eq (185).

Plug (185) into $\mathcal{I}(x) = \gamma\lambda_0\, e^{-\gamma x}\, x^{-1}\,(1{+}o(1))$:

$$z^{\delta+\epsilon(z)}\,\mathcal{I}\big(v(z)\big) = z^{\delta+\epsilon(z)}\ \cdot\ \frac{\gamma\lambda_0\, e^{-(1+\delta+\epsilon(z))\ln z + \ln\ln z - \ln\left(\gamma^3\lambda_0\, c_0 + O(1)\right) + O(1)}}{v(z)}\,(1+o(1))$$
$$= z^{\delta+\epsilon(z)}\ \cdot\ \frac{\gamma\lambda_0\, z^{-(1+\delta+\epsilon(z))}\ \ln z\ \cdot\ e^{O(1)}}{\big(\gamma^3\lambda_0\, c_0 + O(1)\big)^{\frac{1+\delta+\epsilon(z)}{\gamma}}\ln z\,(1+o(1))}\,(1+o(1))$$
$$= \frac{1}{\big(\gamma\, c_0 + O(1)\big)\big(1+\delta\big)}\ \cdot\ \frac{e^{O(1)}}{z}\,(1+o(1))\ =\ \Theta(1/z).$$

□

H.3 Euler Scheme Error Bound

Lemma H.6. *Let $\mathcal{I} \in C^1(\mathbb{R}^+, \mathbb{R}_+^\star)$ one of the function defined in Eq (50). Let $(c_t)_{t\geq 0}$ be a sequence satisfying $c_t \in [c_-, c_+]$ for some $0 < c_- \leq c_+ < \infty$ and all t large enough. Consider the scalar recurrence*

$$b_{t+1} = b_t + h_t\, c_t\, \mathcal{I}(b_t), \qquad b_0 \geq 1, \tag{188}$$

where $(h_t)_{t\geq 0}$ satisfies Condition (16), and set $z_t := \sum_{k=0}^{t-1} h_k$.
Let $b(\cdot)$ be the solution of the perturbed ODE $\dot{b}(z) = c(z)\,\mathcal{I}(b(z))$ with coefficient $c(z) \in [c_-, c_+]$ and the same initial condition $b(0) = b_0$. Then:

(i) *$b_t \to +\infty$ monotonically.*
(ii) *$b_t = \ln z_t - \ln\ln z_t + O(1)$ as $t \to \infty$.*
(iii) *If additionally $c_t = c_0 + \delta_t$ with $\delta_t \to 0$, then*

$$b_t\ =\ b(z_t) + o(1) \qquad \text{as } t \to \infty. \tag{189}$$

Proof. (i) Since $\mathcal{I}(b_t) > 0$, $c_t > 0$, and $h_t > 0$, the increment $h_t c_t \mathcal{I}(b_t) > 0$, so (b_t) is strictly increasing. If it were upper bounded by some $l < \infty$, then since $\mathcal{I}$ is decreasing, $\infty > \sum_t b_{t+1} - b_t \geq c_- \mathcal{I}(l) \sum h_t = +\infty$, contradicting boundedness. Hence $b_t \to +\infty$.

(ii) Define the primitive $G(x) := \int_{b_0}^{x} \frac{\mathrm{d}u}{\mathcal{I}(u)}$.

By Taylor's theorem with the increment $\eta_t := b_{t+1} - b_t = h_t\, c_t\, \mathcal{I}(b_t)$:

$$G(b_{t+1}) - G(b_t) = G'(b_t)\,\eta_t + \tfrac{1}{2}\,G''(\xi_t)\,\eta_t^2$$
$$= \frac{h_t\, c_t\, \mathcal{I}(b_t)}{\mathcal{I}(b_t)} + \tfrac{1}{2}\,G''(\xi_t)\,\eta_t^2$$
$$= h_t\, c_t + \tfrac{1}{2}\,G''(\xi_t)\,\eta_t^2, \tag{190}$$

for some $\xi_t \in [b_t,\, b_{t+1}]$. Since $G''(x) = -\mathcal{I}'(x)/\mathcal{I}(x)^2$, and by Lemma H.1 (with $z\lambda(z)$ instead of $\lambda(z)$):

$$\mathcal{I}(x) \sim \frac{\gamma\,\lambda(\gamma)\, e^{-\gamma x}}{x}, \qquad \mathcal{I}'(x) \sim -\frac{\gamma^2\lambda(\gamma)\, e^{-\gamma x}}{x}, \qquad G''(x) = -\frac{\mathcal{I}'(x)}{\mathcal{I}(x)^2} \sim \frac{x e^{\gamma x}}{\lambda(\gamma)} \tag{191}$$

Therefore

$$G''(\xi_t)\,\eta_t^2\ \sim\ \frac{\xi_t\, e^{\gamma\,\xi_t}}{\lambda(\gamma)}\,h_t^2\, c_t^2\, \mathcal{I}(b_t)^2\ \sim\ \frac{\xi_t\, e^{\gamma(\xi_t - 2\,b_t)}}{b_t^2}\,\gamma^2\,\lambda(\gamma)\,h_t^2\,c_t^2\ =\ o\big(h_t^2\big),$$

where we used that $\xi_t - 2\,b_t \leq \eta_t - b_t \to -\infty$ because $\eta_t = h_t\, c_t\, \mathcal{I}(b_t)$ is bounded.

Summing (190) from 0 to $t-1$:

$$G(b_t) = \underbrace{G(b_0)}_{=0} + \sum_{k=0}^{t-1} h_k\, c_k\ +\ \sum_{k=0}^{t-1} o\big(h_t^2\big)$$
$$= \Theta(z_t)\ +\ O(1) = \Theta(z_t). \tag{192}$$

With the asymptotics $G(x) \sim \frac{x\, e^{\gamma x}}{\gamma^2\, \lambda(\gamma)}$ (derived in the proof of Proposition H.3 or H.4), Eq (192) becomes $\gamma b_t\, e^{\gamma b_t} = \Theta(z_t)$. We then derive $\gamma\, b_t = \ln z_t - \ln\ln z_t + O(1)$ by the same Lambert W asymptotics (see Proposition H.3 or H.4).

(iii) Assume $c_t = c_0 + \delta_t$ with $\delta_t \to 0$. Resumming (190) gives

$$G(b_t) = c_0\, z_t + \sum_{k=0}^{t-1} h_k\, \delta_k + O(1). \tag{193}$$

By the Cesàro, since $\delta_k \to 0$ and $\sum h_k = +\infty$:

$$\frac{\sum_{k=0}^{t-1} h_k\, \delta_k}{\sum_{k=0}^{t-1} h_k} \xrightarrow{t\to\infty} 0, \qquad \text{i.e.} \quad \sum_{k=0}^{t-1} h_k\, \delta_k = o(z_t).$$

Therefore $G(b_t) = c_0\, z_t\, (1 + o(1))$. The ODE trajectory satisfies the same relation: $G\big(b(z_t)\big) = c_0\, z_t\, (1 + o(1))$, by (182) in the proof of Proposition H.4. Hence $G(b_t) - G\big(b(z_t)\big) = o(z_t)$. By the mean-value theorem,

$$b_t - b(z_t) \;=\; \frac{G(b_t) - G(b(z_t))}{G'(\xi_t)}$$

for some ξ_t between b_t and $b(z_t)$. Both satisfy $\gamma\, \xi_t = \ln z_t - \ln\ln z_t + O(1)$, so

$$G'(\xi_t) = \frac{1}{\mathcal{I}(\xi_t)} \;\sim\; \frac{\xi_t\, e^{\gamma\, \xi_t}}{\gamma\lambda(\gamma)} \;=\; \Theta(z_t).$$

Therefore

$$|b_t - b(z_t)| = \frac{o(z_t)}{\Theta(z_t)} = o(1). \qquad \square$$

H.4 Synchronisation Method Lemmas

Lemma H.7. *Let α, and β strictly positive. As $z \to +\infty$, we have the asymptotic*

$$\int_{z_c}^{z} \exp\Big(-\int_s^z \tfrac{\alpha}{r}\, \mathrm{d}r\Big)\, \frac{B}{s \ln s}\, \mathrm{d}s = O\Big(\frac{1}{\ln z}\Big). \tag{194}$$

Proof.

$$\int_{z_c}^{z} \exp\Big(-\int_s^z \tfrac{\alpha}{r}\, \mathrm{d}r\Big)\, \frac{B}{s \ln s}\, \mathrm{d}s \;=\; B\, z^{-\alpha} \int_{z_c}^{z} \frac{s^{\alpha-1}}{\ln s}\, \mathrm{d}s.$$

Integration by parts with $u = 1/\ln s$, $\mathrm{d}v = s^{\alpha-1}\, \mathrm{d}s$, so that $\mathrm{d}u = -\mathrm{d}s/(s(\ln s)^2)$ and $v = s^\alpha/\alpha$, gives

$$\int_{z_c}^{z} \frac{s^{\alpha-1}}{\ln s}\, \mathrm{d}s \;=\; \frac{z^\alpha}{\alpha\, \ln z} - \frac{z_c^\alpha}{\alpha\, \ln z_c} + \frac{1}{\alpha} \int_{z_c}^{z} \frac{s^{\alpha-1}}{(\ln s)^2}\, \mathrm{d}s.$$

Using the theorem of integration of asymptotic equivalents (Jourdain, 2018, *Theorem 4.10*) gives $\int_{z_c}^{z} s^{\alpha-1}/(\ln s)^2\, \mathrm{d}s = O\big(z^\alpha/(\ln z)^2\big)$. Therefore

$$\int_{z_c}^{z} \frac{s^{\alpha-1}}{\ln s}\, \mathrm{d}s \;=\; \frac{z^\alpha}{\alpha\, \ln z}\big(1 + O(1/\ln z)\big),$$

Combining everything:

$$\int_{z_c}^{z} \exp\Big(-\int_s^z \tfrac{\alpha}{r}\, \mathrm{d}r\Big)\, \frac{B}{s \ln s}\, \mathrm{d}s \;=\; B\, z^{-\alpha}\, \frac{z^\alpha}{\alpha\, \ln z}\big(1 + O(1/\ln z)\big) = O(1/\ln z).$$

$\square$

Lemma H.8. *Consider a sequence $(u_t)_t$ a positive sequence satisfying:*

$$u_{t+1} \;\le\; \big(1 - \tfrac{h_t\,\gamma_-}{2\,z_t}\big)\,u_t + \frac{h_t\,B}{z_t\,\ln z_t}, \qquad t \ge t_0,$$

where γ_- and B are strictly positive, $t_0 > 0$, $(h_t)_t$ a sequence satisfying condition (16)*, and $z_t := \sum_{k=0}^{t-1} h_k$. Then, $u_t = O(1/\ln z_t)$.*

Proof. A quick recurrence gives:

$$u_t \;\le\; u_{t_0} \prod_{j=t_0}^{t-1} \big(1 - \tfrac{h_j\gamma_-}{2z_j}\big) + \sum_{k=t_0}^{t-1} \frac{h_k B}{z_k \ln z_k} \prod_{j=k+1}^{t-1} \big(1 - \tfrac{h_j\gamma_-}{2z_j}\big).$$

Noticing that for j large, $0 < h_j\gamma_-/(2z_j) < 1$, and $\ln(1 - h_j\gamma_-/(2z_j)) \le -h_j\gamma_-/(2z_j)$, we have

$$\prod_{j=k+1}^{t-1} \big(1 - \tfrac{h_j\gamma_-}{2z_j}\big) \le \exp\Big(-\sum_{j=k+1}^{t-1} \frac{h_j\gamma_-}{2z_j}\Big).$$

Now $\sum_{j=k+1}^{t-1} h_j/z_j \;\ge\; \int_{z_{k+1}}^{z_t} \mathrm{d}s/s$ (by the integral comparison, using $h_j{=}z_{j+1}{-}z_j$), giving $\sum_{j=k+1}^{t-1} \frac{h_j}{z_j} \ge \ln \frac{z_t}{z_{k+1}}$. Therefore

$$\prod_{j=k+1}^{t-1} \big(1 - \tfrac{h_j\gamma_-}{2z_j}\big) \le \Big(\frac{z_{k+1}}{z_t}\Big)^{\gamma_-/2}. \tag{195}$$

In particular, $\sum h_j/z_j = +\infty$ (since $z_t \to \infty$), so the homogeneous term $u_{t_1}\prod(\cdots)$ decays polynomially in z_t, hence tends to 0. For the remaining term, setting $\alpha := \gamma_-/2 > 0$:

$$\sum_{k=t_0}^{t-1} \frac{h_k B}{z_k \ln z_k} \prod_{j=k+1}^{t-1} \big(1 - \tfrac{h_j\gamma_-}{2z_j}\big) \;\le\; \frac{B}{z_t^\alpha} \sum_{k=t_0}^{t-1} \frac{h_k\, z_{k+1}^\alpha}{z_k \ln z_k} \;\le\; \frac{2^\alpha B}{z_t^\alpha} \sum_{k=t_0}^{t-1} \frac{h_k\, z_k^{\alpha-1}}{\ln z_k}, \tag{196}$$

using $z_{k+1}^\alpha \le 2^\alpha z_k^\alpha$ for k large (since $z_{k+1}/z_k \to 1$). To bound the remaining sum, we use Abel's summation formula. Define $A_k := \sum_{j=t_0}^{k} h_j$ so that $A_{t-1}{=}z_t{-}z_{t_0}$ and $A_k{-}A_{k-1}{=}h_k$. Writing $f_k := z_k^{\alpha-1}/\ln z_k = \varphi(z_k)$:

$$\sum_{k=t_1}^{t-1} h_k\, f_k = A_{t-1}\, f_{t-1} - \sum_{k=t_1}^{t-2} A_k\,(f_{k+1} - f_k).$$

$A_{t-1} = z_t - z_{t_1} \le z_t$ and $f_{t-1} \sim z_t^{\alpha-1}/\ln z_t$, so $A_{t-1}\, f_{t-1} = O(z_t^\alpha/\ln z_t)$.

By the mean-value theorem, $f_{k+1}{-}f_k = h_k\,\varphi'(\xi_k)$ for some $\xi_k \in [z_k, z_{k+1}]$, with $\varphi'(s) = s^{\alpha-2}\big((\alpha{-}1)\ln s - 1\big)/(\ln s)^2$. For all $\alpha > 0$ and s large, $|\varphi'(s)| \le C_\varphi\, s^{\alpha-2}$ for some $C_\varphi > 0$. Therefore $|f_{k+1} - f_k| \le C_\varphi\, h_k\, z_k^{\alpha-2}$ and

$$\sum_{k=t_1}^{t-2} A_k\,|f_{k+1} - f_k| \le C_\varphi \sum_{k=t_1}^{t-2} z_k\, h_k\, z_k^{\alpha-2}\, h_k = C_\varphi \sum_{k=t_1}^{t-2} h_k^2\, z_k^{\alpha-1}.$$

Since $z_k^{\alpha-1} \le z_k^{\max(\alpha-1,0)} \le \max(1, z_t^{\alpha-1})$ and $\sum h_k^2 < \infty$ by (16), this sum is $O(z_t^{\max(\alpha-1,0)}) = o(z_t^\alpha/\ln z_t)$. Combining: $\sum_{k=t_1}^{t-1} h_k\,\varphi(z_k) = O(z_t^\alpha/\ln z_t)$. Substituting back into (196):

$$\sum_{k=t_0}^{t-1} \frac{h_k B}{z_k \ln z_k} \prod_{j=k+1}^{t-1} \big(1 - \tfrac{h_j\gamma_-}{2z_j}\big) \;\le\; \frac{2^\alpha B}{z_t^\alpha} \cdot \frac{C\, z_t^\alpha}{\alpha\, \ln z_t} = O(1/\ln z_t).$$

Therefore the forcing contribution is $O(1/\ln z_t)$.

Combining: $u_t = O(z_t^{-\alpha}) + O(1/\ln z_t) = O(1/\ln z_t)$ □

H.5 Discrete Synchronisation

The synchronisation Lemma E.5 establishes $a(z) - b(z) - d \to 0$ for the ODE trajectory. To close the proof, we need the analogous statement for the *discrete* iterates (a_t, b_t).

Lemma H.9 (Discrete synchronisation)**.** *Under Condition 10 and Condition* (16)*, if there exists $t_0 > 0$ such that the discrete iterates of Proposition E.3 satisfy*

$$|a_{t_0} - \tilde{\gamma}_{min}\, b_{t_0} - d| \;\le\; \eta_\star, \qquad b_{t_0} \ge b_\star, \qquad \textit{and} \quad h_t \;\textit{sufficiently small for all } t \ge t_0,$$

with $\eta_\star$ and $b_\star$ as described in Lemma E.5. Then

$$|a_t - \tilde{\gamma}_{min}\, b_t - d| \;\le\; \eta_\star \qquad \textit{for all } t \ge t_0.$$

Moreover, the discrete gap converges: $a_t - \tilde{\gamma}_{min}\, b_t \to d$.

Proof. The proof mirrors the previous proof on the continuous case.

(a) Removing z from Lemma E.5 setting. Define the discrete gap $\eta_t := a_t - \tilde{\gamma}_{\min}\, b_t - d$. From Eqs (52)–(53):

$$\eta_{t+1} = \eta_t + h_t\, \Phi(\eta_t, b_t), \tag{197}$$

where

$$\Phi(\eta, b) := (1-\varepsilon)(1+\mu_A)(1-\alpha)\, \mathcal{I}_{\text{maj}}(\tilde{\gamma}_{\min}\, b + d + \eta) - \varepsilon\Big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\Big)\, \mathcal{I}_{\min}(b). \tag{198}$$

Noticing that s in Eq (66) depends on z only through $b(z)$, we can rewrite Φ like:

$$\Phi(\eta, b) = \varepsilon\Big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\Big)\, \mathcal{I}_{\min}(b)\, \big[e^{-\eta}\big(1 + s(b,\eta)\big) - 1\big], \tag{199}$$

with s rewritten as:

$$s(b,\eta) := \frac{\tilde{\gamma}_{\min}\, b}{\tilde{\gamma}_{\min}\, b + d + \eta} \cdot \frac{1 + r_{\text{maj}}(\tilde{\gamma}_{\min}\, b + d + \eta)}{1 + r_{\min}(b)} - 1. \tag{200}$$

The estimates (67)-(68) remain valid for any $b \le b_\star$ and $|\eta| \le 1$:

$$|s(b,\eta)| \le \frac{K_s}{|b|}, \qquad |s(b,\eta) - s(b,0)| \le \frac{K_s'\,|\eta|}{|b|}. \tag{201}$$

(b) Discrete tube invariance. The tube argument is identical to step (i) of Lemma E.5: for $b \ge b_\star$, the same computation as in (62) gives $\Phi(\eta_\star, b) < 0$ and $\Phi(-\eta_\star, b) > 0$. Therefore on the boundary $\eta = \eta_\star$ for instance, $\eta_{t+1} = \eta_\star + h_t\, \Phi(\eta_\star, b_t) \in [-\eta_\star, \eta_\star]$ provided $h_t\, |\Phi(\eta_\star, b_t)| < 2\eta_\star$, which holds for all $t > t_0$ with h_t sufficiently small and given $|\Phi|$ is bounded on $\mathbb{R}^2_{\ge 1}$.

Since $b_t \to +\infty$ monotonically (Remark E.4), b_t eventually remains above $b_\star$. Choosing t_0 as the first time both h_t is small enough and $b_t \ge b_\star$, the tube $\{|\eta_t| \le \eta_\star\}$ is forward invariant for $t \ge t_0$.

(c) Convergence. Like in step (a), noticing that γ, β and R depends on z only through b and η, the linearisation Eq (75) generalizes as well from Eq (199) into:

$$\Phi(\eta, b) = -\hat{\gamma}(b)\, \eta + \hat{\beta}(b) + \hat{R}(b, \eta)$$

with

$$\hat{\gamma}(b) := \varepsilon\Big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\Big)\, \mathcal{I}_{\min}(b), \quad \hat{\beta}(b) := \varepsilon\Big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\Big)\, \mathcal{I}_{\min}(b)\, s(b,0), \tag{202}$$

$$|\hat{R}(b,\eta)| \le \hat{K}\, |\mathcal{I}_{\min}(b)|\, \left(\eta^2 + \frac{|\eta|}{b}\right) \tag{203}$$

with $\hat{K}$ an explicit constant. By the tube invariance (step (b)), $|\eta_t| \le \eta_\star$ for $t \ge t_0$, so the quantity

$$\tilde{c}_t := \varepsilon(1+\mu_B) + (1-\varepsilon)(1+\mu)\,\frac{\mathcal{I}_{\text{maj}}(a_t)}{\mathcal{I}_{\min}(b_t)}; \qquad \text{s.t.} \quad b_{t+1} - b_t = \tilde{c}_t\,\mathcal{I}_{\min}(b_t) \quad \text{(see Eq (53))}$$

satisfies $\tilde{c}_t \in [\tilde{c}_-, \tilde{c}_+] \subset (0,\infty)$ for $t \geq t_1$, by the same bounded-ratio argument as in (72). Lemma H.6, part (ii), then gives

$$\gamma\, b_t = \ln z_t - \ln\ln z_t + O(1), \qquad \text{s.t.} \quad \mathcal{I}_{\min}(b_t) = \Theta(1/z_t) \quad \text{(as in Eq (74))}. \tag{204}$$

Defining $\gamma_t := \hat{\gamma}(b_t)$, $\beta_t := \hat{\beta}(b_t)$, $\mathcal{R}_t := \hat{R}(b_t, \eta_t)$, we also derive the analogous constants and bounds:

(i) $\frac{M_-}{z_t} \leq \mathcal{I}_{\min}(b_t) \leq \frac{M_+}{z_t}$, with $0 < M_- \leq M_+$ and $\frac{\gamma_-}{z_t} \leq \gamma_t \leq \frac{\gamma_+}{z_t}$,

$$\gamma_\pm := \varepsilon\Big(\tilde{\gamma}_{\min}\,(1+\mu_B) - (1+\mu)\Big)\,M_\pm,$$

(ii) $|\beta_t| \leq B/(z_t \ln z_t)$ with $B > 0$ by Eq (201),

(iii) from (203) with $(b,\eta) = (b_t, \eta_t)$ and using $b_t \geq c_b \ln z_t$ for some $c_b > 0$:

$$|\mathcal{R}_t| \leq \hat{K}\,|\mathcal{I}_{\min}(b_t)|\,\Big(\eta_t^2 + \frac{|\eta_t|}{b_t}\Big) \leq \frac{\hat{K}\,M_+}{\min(1,c_b) z_t}\,\Big(\eta_t^2 + \frac{|\eta_t|}{\ln z_t}\Big) = \frac{\check{K}}{z_t}\,\Big(\eta_t^2 + \frac{|\eta_t|}{\ln z_t}\Big),$$

where we wrote $\check{K}$ for $\hat{K}\,M_+/\min(1,c_b)$ to lighten notation). The recurrence (197) with (202) reads

$$\eta_{t+1} = (1 - h_t\gamma_t)\,\eta_t + h_t\,\beta_t + h_t\,\mathcal{R}_t.$$

Setting $u_t := |\eta_t|$ and using $|\eta_t| \leq \eta_\star$ with $\eta_\star$ small enough and t_0 large enough that $\check{K}\eta_\star \leq \gamma_-/4$ and $\check{K}/\ln z_{t_0} \leq \gamma_-/4$ (exactly as we did to get Eq (79), we absorb $\mathcal{R}_t$ into γ_t:

$$u_{t+1} \;\leq\; \big(1 - \tfrac{h_t\,\gamma_-}{2\,z_t}\big)\,u_t + \frac{h_t\,B}{z_t\,\ln z_t}, \qquad t \geq t_0, \tag{205}$$

which is the exact hypothesis of Lemma H.8. We can therefore conclude that $u_t = O(1/\ln z_t)$ and $\eta_t = a_t - \tilde{\gamma}_{\min}\,b_t - d \to 0$. □

I GENERALIZING SOUDRY ET AL. (2018, *Theorem 9*)

I.1 SETUP AND ASSUMPTIONS

We consider a data distribution $\mathcal{D}$ over $(\mathbf{x}, y)$ with $\mathbf{x} \in \mathbb{R}^d$ and binary labels $y \in \{-1, +1\}$. Following the relabelling convention of Soudry et al. (2018), we absorb the label into the feature: redefine $\mathbf{x} \leftarrow y\mathbf{x}$ so that we may assume $y = +1$ for all samples without loss of generality. The population expected risk (in contrast to empirical risk) loss is

$$\mathcal{L}(\mathbf{w}) \;=\; \mathbb{E}\big[\ell(\mathbf{w}^\top \mathbf{x})\big]\,, \tag{206}$$

where ℓ is any loss function satisfying Soudry et al. (2018, *Assumptions 2 and 3*). Without loss of generality we will consider $\ell(u) = e^{-u}$ for simplicity.

We perform *population (full-batch) gradient descent* with fixed learning rate $\eta > 0$:

$$\mathbf{w}(t+1) = \mathbf{w}(t) - \eta\, \nabla \mathcal{L}(\mathbf{w}(t)) = \mathbf{w}(t) + \eta\, \mathbb{E}\Big[e^{-\mathbf{w}(t)^\top \mathbf{x}}\, \mathbf{x}\Big]\,. \tag{207}$$

We impose the following standard assumptions on $\mathcal{D}$ inspired from 2.2 and D.1.

Assumption I.1 (Linear separability). *There exists $\mathbf{w}_\star \in \mathbb{R}^d$ such that $\mathbf{w}_\star^\top \mathbf{x} \geq 1$ almost surely.*

Under Assumption I.1, the population L_2 max-margin (analogous to hard-margin SVM) solution is well-defined:

$$\hat{\mathbf{w}} \;=\; \underset{\mathbf{w}\in\mathbb{R}^d}{\operatorname{argmin}} \|\mathbf{w}\|^2 \qquad \text{s.t.} \qquad \mathbf{w}^\top \mathbf{x} \geq 1 \;\text{ a.s.} \tag{208}$$

Assumption I.2 (Bounded features). *There exists $K > 0$ such that $\|\mathbf{x}\| \leq K$ almost surely. Combined with Assumptions I.1 this assumption gives $\mathcal{L}$ is β-smooth with $\beta = K^2 \mathcal{L}(\mathbf{w}(0))$. For technical reasons in the proof below we set $\beta := \max(\mathcal{L}(\mathbf{w}(0)), 8)\, K^2$ instead.*

Assumption I.3 (Margin density). *$\mathcal{D}$ has continuous density and if we define the* margin variable

$$\mathbf{m} \;:=\; \hat{\mathbf{w}}^\top \mathbf{x}, \tag{209}$$

then $\mathbf{m}$ has a continuous strictly positive density p_m on $[1, M]$ where $M = K\|\hat{\mathbf{w}}\|$.

Assumption I.4. *Let $\mathbf{x}_\perp := \bar{\Pi}_{\hat{\mathbf{w}}} \mathbf{x}$ (see (17) for the notation), then*

$$\mathbb{E}[\mathbf{x}_\perp \mid \mathbf{m}] \;=\; 0. \tag{210}$$

This assumption mirrors Assumption D.1 and similarly could be weakened by requiring that for every direction $\hat{\boldsymbol{\delta}}$ such that $\hat{\boldsymbol{\delta}}^\top \hat{\mathbf{w}} = 0$, $\mathbf{x}_\perp \mid \mathbf{m}{=}1$ never puts all its mass on one single side of the hyperplane $(\hat{\boldsymbol{\delta}})^\perp$ (non-degeneracy F.1):

$$\Pr\big(\hat{\boldsymbol{\delta}}^\top \mathbf{x}_\perp < 0 \;\big|\; m{=}1\big) > 0. \tag{211}$$

In geometric terms, the origin lies in the interior of the convex hull of the support of $\mathbf{x}_\perp \mid \mathbf{m}{=}1$ in the subspace $(\hat{\mathbf{w}})^\perp$. However, this would require considerable more complex mathematical development, therefore we opt for this simpler and relatively general setting.

In Soudry et al. (2018) finite-sample framework, the decomposition $\mathbf{w}(t) = \hat{\mathbf{w}} \log t + \boldsymbol{\rho}(t)$ with *bounded* $\boldsymbol{\rho}$ holds for "almost all datasets" (their Theorem 9); the degenerate case where $\|\boldsymbol{\rho}\| = O(\log\log t)$ is relegated to Theorem 13. Although our continuous-distribution study does not observe such degeneracy, we will show that a $O(\log\log t)$ correction that is absent in the discrete case is nevertheless introduced.

We now state the following proposition generalizing the theorem to our continuous distribution setting.

Proposition I.5 (Implicit bias of GD - continuous-distribution setting)**.** *Under Assumptions I.1-I.4, let $\mathbf{w}(t)$ denote the iterates of population gradient descent* (207) *with any initial point $\mathbf{w}(0)$ and learning rate η small enough (specified in the proof). Then, as $t \to \infty$:*

$$\mathbf{w}(t) \;=\; \hat{\mathbf{w}}\log t \;-\; \hat{\mathbf{w}}\log\log t \;+\; O(1). \tag{212}$$

In particular, the direction of $\mathbf{w}(t)$ still converges to the max-margin direction:

$$\lim_{t\to\infty}\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} = \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}.$$

I.2 Proof of Proposition 5.4

The proof proceeds in four steps inspired from the original discrete proof.

(i) Loss convergence and norm divergence

Lemma I.6. *For η small enough: $\mathcal{L}(\mathbf{w}(t)) \to 0$, and $\|\mathbf{w}(t)\| \to \infty$ as $t \to \infty$.*

Proof. The argument follows Soudry et al. (2018, *Lemma 1*) textually. For any $\mathbf{w}$:

$$\mathbf{w}_\star^\top \nabla\mathcal{L}(\mathbf{w}) = -\mathbb{E}\Big[e^{-\mathbf{w}^\top\mathbf{x}}\,\mathbf{w}_\star^\top\mathbf{x}\Big] < 0 \tag{213}$$

since $\mathbf{w}_\star^\top\mathbf{x} > 0$ a.s. Therefore there's no finite critical point $\mathbf{w}$ such that $\nabla\mathcal{L}(\mathbf{w}) = 0$. Since $\mathcal{L}$ is β-smooth, gradient descent with $\eta < 2/\beta$ converges to a critical point: $\nabla\mathcal{L}(\mathbf{w}(t)) \to 0$ (see Soudry et al. (2018, *Lemma 10*)). The absence of finite critical points forces $\|\mathbf{w}(t)\| \to \infty$. Finally, using $\mathbf{w}_\star^\top\mathbf{x} \geq 1$ a.s, we have the following bound

$$0 \leq \mathcal{L}(\mathbf{w}(t)) = \mathbb{E}\Big[e^{-\mathbf{w}^\top\mathbf{x}}\Big] \leq \mathbb{E}\Big[e^{-\mathbf{w}^\top\mathbf{x}}\,\mathbf{w}_\star^\top\mathbf{x}\Big] = \Big|\mathbf{w}_\star^\top\nabla\mathcal{L}(\mathbf{w}(t))\Big| \leq \|\mathbf{w}_\star\|\|\nabla\mathcal{L}(\mathbf{w}(t))\| \to 0,$$

hence $\mathcal{L}(\mathbf{w}(t)) \to 0$. □

Decomposing $\mathbf{w}(t)$ like this:

$$\mathbf{w}(t) = f(t)\,\hat{\mathbf{w}} + \boldsymbol{\delta}(t), \qquad f(t) := \frac{\mathbf{w}(t)^\top\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|^2}, \qquad \hat{\mathbf{w}}^\top\boldsymbol{\delta}(t) = 0, \tag{214}$$

we now show $f(t) \to \infty$.

(ii) $f(t)$ diverges.

Lemma I.7 (Coarse f growth rate)**.** *$f(t) \to +\infty$, and more precisely, $f(t) = \Theta(\log t)$.*

Proof. Projecting (207) onto $\hat{\mathbf{w}}$ and using Assumption I.1:

$$f(t+1) - f(t) = \frac{\eta}{\|\hat{\mathbf{w}}\|^2}\,\mathbb{E}\big[e^{-\mathbf{w}(t)^\top\mathbf{x}}\,\hat{\mathbf{w}}^\top\mathbf{x}\big] = \frac{\eta}{\|\hat{\mathbf{w}}\|^2}\,\mathbb{E}\big[e^{-\mathbf{w}(t)^\top\mathbf{x}}\,\mathbf{m}\big] \tag{215}$$

$$\geq \frac{\eta}{\|\hat{\mathbf{w}}\|^2}\,\mathbb{E}\big[e^{-\mathbf{w}(t)^\top\mathbf{x}}\big] = \frac{\eta}{\|\hat{\mathbf{w}}\|^2}\,\mathcal{L}(\mathbf{w}(t)) > 0. \tag{216}$$

So f is strictly increasing for all t and summing (216):

$$f(t) - f(0) \;\geq\; \frac{\eta}{\|\hat{\mathbf{w}}\|^2}\sum_{t=0}^{t-1}\mathcal{L}(\mathbf{w}(t)). \tag{217}$$

It remains to show $\sum_t \mathcal{L}(\mathbf{w}(t)) = +\infty$. As in Soudry et al. (2018, *Lemma 10*), using $\mathcal{L}$'s β-smoothness, we have:

$$\mathcal{L}(\mathbf{w}(t)) - \mathcal{L}(\mathbf{w}(t{+}1)) \geq \eta\Big(1 - \frac{\eta\beta}{2}\Big)\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 > 0, \tag{218}$$

$$\mathcal{L}(\mathbf{w}(t)) - \mathcal{L}(\mathbf{w}(t{+}1)) \leq \eta\Big(1 + \frac{\eta\beta}{2}\Big)\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 \leq \eta K^2\Big(1 + \frac{\eta\beta}{2}\Big)\mathcal{L}(\mathbf{w}(t))^2, \tag{219}$$

where the last inequality comes from $\|\nabla\mathcal{L}\| = \|\mathbb{E}[e^{-\mathbf{w}^\top\mathbf{x}}\mathbf{x}]\| \le K\,\mathcal{L}(\mathbf{w})$.

Dividing both sides by $\mathcal{L}(\mathbf{w}(t))\,\mathcal{L}(\mathbf{w}(t{+}1))$ and using $\mathcal{L}(\mathbf{w}(t{+}1)) \le \mathcal{L}(\mathbf{w}(t))$ (by Eq (218)):

$$\begin{aligned}\frac{1}{\mathcal{L}(\mathbf{w}(t{+}1))} - \frac{1}{\mathcal{L}(\mathbf{w}(t))} &= \frac{\mathcal{L}(\mathbf{w}(t)) - \mathcal{L}(\mathbf{w}(t{+}1))}{\mathcal{L}(\mathbf{w}(t))\,\mathcal{L}(\mathbf{w}(t{+}1))} \\ &\le \Big(1+\frac{\eta\beta}{2}\Big)\frac{\eta K^2\,\mathcal{L}(\mathbf{w}(t))^2}{\mathcal{L}(\mathbf{w}(t))\,\mathcal{L}(\mathbf{w}(t{+}1))} \\ &\le \Big(1+\frac{\eta\beta}{2}\Big)\frac{\eta K^2\,\mathcal{L}(\mathbf{w}(t))}{\mathcal{L}(\mathbf{w}(t{+}1))} \;\le\; 4\eta K^2, \end{aligned} \tag{220}$$

where the last step uses $\eta < 2/\beta$ and the fact that for t large enough, $\mathcal{L}(\mathbf{w}(t))/\mathcal{L}(\mathbf{w}(t+1)) \le 2$.

In fact, for t large enough, $\mathcal{L}(\mathbf{w}(t)) < 1$ (by Lemma I.6) which gives by combining to Eq (219):

$$\begin{aligned}\mathcal{L}(\mathbf{w}(t)) - \mathcal{L}(\mathbf{w}(t{+}1)) &\le \eta K^2\Big(1+\frac{\eta\beta}{2}\Big)\mathcal{L}(\mathbf{w}(t)) \\ &= \frac{\eta\beta}{\max(\mathcal{L}(\mathbf{w}(0)),8)}\Big(1+\frac{\eta\beta}{2}\Big)\mathcal{L}(\mathbf{w}(t)) \quad \text{by Assumption I.2} \\ &= \frac{\mathcal{L}(\mathbf{w}(t))}{2} \qquad \text{since } \eta < 2/\beta,\end{aligned}$$

so that $\mathcal{L}(\mathbf{w}(t+1)) \ge \mathcal{L}(\mathbf{w}(t))/2$. Summing Eq (220): $1/\mathcal{L}(\mathbf{w}(t)) \le 1/\mathcal{L}(\mathbf{w}(0)) + 4\eta K^2\,t$, hence

$$\mathcal{L}(\mathbf{w}(t)) \;\ge\; \frac{c'}{t} \qquad \text{for } c' := \frac{1}{4\eta K^2 + 1/\mathcal{L}(\mathbf{w}(0))} > 0. \tag{221}$$

Therefore $\sum_t \mathcal{L}(\mathbf{w}(t)) = +\infty$ and (217) gives $f(t) \to +\infty$.

We can also similarly derive an upper-bound for the loss as follows. By Eq (213), $\hat{\mathbf{w}}^\top\nabla\mathcal{L} = -\mathbb{E}[e^{-\mathbf{w}^\top\mathbf{x}}\,\mathbf{m}] \le -\mathcal{L}(\mathbf{w})$ (since $\mathbf{m} \ge 1$ a.s.), so $\|\nabla\mathcal{L}\| \ge |\hat{\mathbf{w}}^\top\nabla\mathcal{L}|/\|\hat{\mathbf{w}}\| \ge \mathcal{L}/\|\hat{\mathbf{w}}\|$. Therefore with Eq (218):

$$\mathcal{L}(\mathbf{w}(t)) - \mathcal{L}(\mathbf{w}(t{+}1)) \;\ge\; \Big(1-\frac{\eta\beta}{2}\Big)\frac{\eta}{\|\hat{\mathbf{w}}\|^2}\,\mathcal{L}(\mathbf{w}(t))^2. \tag{222}$$

By the same telescoping argument as above: $\frac{1}{\mathcal{L}(\mathbf{w}(t+1))} - \frac{1}{\mathcal{L}(\mathbf{w}(t))} \ge \Big(1-\frac{\eta\beta}{2}\Big)\frac{\eta}{\|\hat{\mathbf{w}}\|^2}$, giving $\mathcal{L}(\mathbf{w}(t)) \le C/t$ for $C := \Big(1-\frac{\eta\beta}{2}\Big)^{-1}\|\hat{\mathbf{w}}\|^2/\eta$. Combined with (221):

$$\mathcal{L}(\mathbf{w}(t)) = \Theta(1/t). \tag{223}$$

A similar analysis as in Eq (216) gives the upper bound $f(t+1) - f(t) \le \eta K\,\mathcal{L}(\mathbf{w}(t))/\|\hat{\mathbf{w}}\|$ (from $\mathbf{m} \le K\|\hat{\mathbf{w}}\|$ a.s.). When combined with $\mathcal{L}(\mathbf{w}(t)) \le C/t$ we get $f(t) \le C'\log t + C''$. Together with $\mathcal{L} \ge c'/t$ and (216):

$$f(t) = \Theta(\log t). \tag{224}$$

□

(iii) $\boldsymbol{\delta}(t)$ is bounded.

Lemma I.8. *$\boldsymbol{\delta}(t) = O(1)$ and therefore $\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} \xrightarrow{t\to\infty} \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}$.*

Proof. Projecting (207) onto $\text{Span}\{\hat{\mathbf{w}}\}^\perp$:

$$\boldsymbol{\delta}(t{+}1) - \boldsymbol{\delta}(t) = \eta\,\mathbb{E}\big[e^{-f(t)\mathbf{m}-\boldsymbol{\delta}(t)^\top\mathbf{x}_\perp}\,\mathbf{x}_\perp\big], \tag{225}$$

so that

$$\|\boldsymbol{\delta}(t{+}1) - \boldsymbol{\delta}(t)\| \;\le\; \eta K\,\mathcal{L}(\mathbf{w}(t)) \;\le\; \eta K\,\frac{C}{t}. \tag{226}$$

Therefore:

$$\sum_{t=1}^{\infty} \|\boldsymbol{\delta}(t{+}1) - \boldsymbol{\delta}(t)\|^2 \;\le\; \eta^2 K^2 C^2 \sum_{t=1}^{\infty} \frac{1}{t^2} \;<\; \infty. \tag{227}$$

Now let's write

$$\begin{aligned}
\|\boldsymbol{\delta}(t{+}1)\|^2 - \|\boldsymbol{\delta}(t)\|^2 &= 2\,\boldsymbol{\delta}(t)^\top \big[\boldsymbol{\delta}(t{+}1) - \boldsymbol{\delta}(t)\big] + \|\boldsymbol{\delta}(t{+}1) - \boldsymbol{\delta}(t)\|^2 \\
&= 2\eta\, \mathbb{E}[e^{-f\mathbf{m} - \boldsymbol{\delta}^\top \mathbf{x}_\perp}\, \boldsymbol{\delta}^\top \mathbf{x}_\perp] \;+\; \|\boldsymbol{\delta}(t{+}1) - \boldsymbol{\delta}(t)\|^2 \quad \text{(by Eq (225)} \\
&= 2\eta \underbrace{\mathbb{E}\big[e^{-f\mathbf{m}}\,(e^{-z} - 1)\,z\big]}_{=:\; R(t) \;\le\; 0} + 2\eta \underbrace{\mathbb{E}\big[e^{-f\mathbf{m}}\,z\big]}_{=:\; L(t)} + \|\boldsymbol{\delta}(t{+}1) - \boldsymbol{\delta}(t)\|^2, \qquad (228)
\end{aligned}$$

where we use $e^{-z} = (e^{-z}{-}1){+}1$ with $z = \boldsymbol{\delta}^\top \mathbf{x}_\perp$ along with the pointwise inequality $z(e^{-z}{-}1) \le 0$ for all $z \in \mathbb{R}$.

Under Assumption I.4 and the total expectation property, we have:

$$L(t) = \boldsymbol{\delta}^\top \mathbb{E}[e^{-f\mathbf{m}}\, \mathbf{x}_\perp] = \boldsymbol{\delta}^\top \mathbb{E}_{\mathbf{m}}\Big[e^{-f\mathbf{m}}\, \mathbb{E}[\mathbf{x}_\perp \mid \mathbf{m}]\Big] = 0. \tag{229}$$

Substituting $L(t) = 0$ into (228):

$$\|\boldsymbol{\delta}(t{+}1)\|^2 - \|\boldsymbol{\delta}(t)\|^2 = 2\eta\, R(t) + \|\boldsymbol{\delta}(t{+}1) - \boldsymbol{\delta}(t)\|^2 \;\le\; \|\boldsymbol{\delta}(t{+}1) - \boldsymbol{\delta}(t)\|^2. \tag{230}$$

Summing from 0 to $t{-}1$ and using Eq (227):

$$\|\boldsymbol{\delta}(t)\|^2 \;\le\; \|\boldsymbol{\delta}(0)\|^2 + \sum_{t=0}^{T-1} \|\boldsymbol{\delta}(t{+}1) - \boldsymbol{\delta}(t)\|^2 \;\le\; \|\boldsymbol{\delta}(0)\|^2 + \eta^2 K^2 C^2 \sum_{t=1}^{\infty} t^{-2}, \tag{231}$$

hence $\|\boldsymbol{\delta}(t)\| = O(1)$. Combining with $f(t) \to +\infty$:

$$\frac{\|\boldsymbol{\delta}(t)\|}{f(t)} \to 0, \qquad \text{and therefore} \qquad \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} \xrightarrow{t\to\infty} \frac{\hat{\mathbf{w}}}{\|\hat{\mathbf{w}}\|}. \tag{232}$$

□

(iv) Sharp growth rate of f.

Lemma I.9. $f(t) = \log t - \log\log t + O(1)$.

Proof. From (215):

$$f(t+1) - f(t) = \frac{\eta}{\|\hat{\mathbf{w}}\|^2}\, \mathbb{E}\big[e^{-f(t)\mathbf{m} - \boldsymbol{\delta}(t)^\top \mathbf{x}_\perp}\, \mathbf{m}\big] = \frac{\eta}{\|\hat{\mathbf{w}}\|^2}\, \mathbb{E}_{\mathbf{m}}\Big[e^{-f(t)\mathbf{m}}\, \mathbf{m}\, \mathbb{E}[e^{-\boldsymbol{\delta}(t)^\top \mathbf{x}_\perp} \mid \mathbf{m}]\Big].$$

By the disintegration theorem Ambrosio et al. (2005, *Theorem 5.3.1*) and Assumption I.3, there exists a continuous function $H(m,t) := m\, \mathbb{E}_{\mathbf{x}|\mathbf{m}=m}[e^{-\boldsymbol{\delta}(t)^\top \mathbf{x}_\perp} \mid \mathbf{m} = m]\, p_m(m)$ such that

$$\mathbb{E}_{\mathbf{m}}\Big[e^{-f(t)\mathbf{m}}\, \mathbf{m}\, \mathbb{E}[e^{-\boldsymbol{\delta}(t)^\top \mathbf{x}_\perp} \mid \mathbf{m}]\Big] = \int_1^{K\|\hat{\mathbf{w}}\|} e^{-f(t)m}\, H(m,t)\, dm. \tag{233}$$

Since $\|\boldsymbol{\delta}\| = O(1)$ and $\|\mathbf{x}_\perp\| \le K$ a.s., there exists a constant $D > 0$ such that for all t:

$$e^{-DK}\, m\, p_m(m) \;\le\; H(m,t) \;\le\; e^{DK}\, m\, p_m(m). \tag{234}$$

Define the t-independent bounding function $h(m) := m\, p_m(m)$ (continuous with $h(1) = p_m(1) > 0$). Eq (234) gives

$$e^{-DK} \int_1^M e^{-fm}\, h(m)\, dm \;\le\; \int_1^M e^{-fm}\, H(m,t)\, dm \;\le\; e^{DK} \int_1^M e^{-fm}\, h(m)\, dm.$$

Watson's lemma Wong (2001, *Chapter 1*) for $f \to +\infty$ applied to the bounding integrals gives:

$$\int_1^M e^{-fm}\, h(m)\, dm \;=\; \frac{p_m(1)}{f}\, e^{-f}\, \big(1 + O(1/f)\big). \tag{235}$$

Combining both results yields:

$$\eta\, c_- \,\frac{e^{-f(t)}}{f(t)}\, \big(1 + O(1/f)\big) \;\le\; f(t+1) - f(t) \;\le\; \eta\, c_+ \,\frac{e^{-f(t)}}{f(t)}\, \big(1 + O(1/f)\big), \tag{236}$$

where $c_\pm := \frac{e^{\pm DK}\, p_m(1)}{\|\hat{\mathbf{w}}\|^2} > 0$. We write this compactly as

$$f(t+1) - f(t) \;=\; \eta\, C(t)\, \frac{e^{-f(t)}}{f(t)}, \qquad 0 < c_- \le C(t) \le c_+, \tag{237}$$

potentially adjusting $c_\pm$ accordingly to absorb the $O(1/f)$.

Let Φ the function $f \longmapsto (f-1)e^f$. By Taylor's theorem applied to Φ at $f(t)$ with increment $a_t := f(t{+}1) - f(t) \ge 0$:

$$\begin{aligned} \Phi(f(t{+}1)) - \Phi(f(t)) &= \Phi'(f(t))\, a_t + \tfrac{1}{2}\, \Phi''(\xi_t)\, a_t^2 \\ &= \eta\, C(t) + \tfrac{1}{2}\, \Phi''(\xi_t)\, a_t^2 \qquad \text{(by Eq (237)} \end{aligned} \tag{238}$$

for some $\xi_t \in (f(t), f(t{+}1))$. Since $f(t)e^{f(t)} \le \Phi''(\xi_t) \le 2f(t)\, e^{f(t)}$ for t large:

$$0 \le \Phi''(\xi_t)\, a_t^2 \le 2f(t)\, e^{f(t)} \cdot \eta^2\, C(t)^2\, \frac{e^{-2f(t)}}{f(t)^2} = 2\eta^2\, C^2(t)\, \frac{e^{-f(t)}}{f(t)} \le 2\eta\, c_+\, a_t.$$

Combining and summing from t_0 to $t-1$ for sufficiently large t:

$$\begin{aligned} \Phi(f(t)) &\le \Phi(f(t_0)) + \eta \sum_{s=t_0}^{t-1} C(s) + 2\eta\, c_+ \sum_{s=t_0}^{t-1} a_s \\ &\le \Phi(f(t_0)) + \eta\, c_+\, (t - t_0) + 2\eta\, c_+ \big(f(t) - f(t_0)\big) \\ &\le \Phi(f(t_0)) + \eta\, c_+\, t + 2\eta\, c_+\, f(t) \end{aligned} \tag{239}$$

$$\begin{aligned} \Phi(f(t)) &\ge \Phi(f(t_0)) + \eta \sum_{s=t_0}^{t-1} C(s) \\ &\ge \Phi(f(t_0)) + \eta\, c_-\, (t - t_0) \\ &\ge \Phi(f(t_0)) + \frac{\eta\, c_-}{2}\, t. \end{aligned} \tag{240}$$

Using the Lambert W function which is increasing, and noticing that $\Phi(f) = (f-1)e^f = b$ implies $f - 1 = W(b/e)$, these inequalities yield:

$$W\Big(\frac{\Phi(f(t_0))}{e} + \frac{\eta\, c_-}{2e}\, t\Big) \le f(t) - 1 \le W\Big(\frac{\Phi(f(t_0))}{e} + \frac{\eta\, c_+}{e}\, t + \frac{2\eta\, c_+}{e} f(t)\Big). \tag{241}$$

By Lambert W asymptotics Visser (2018, *Eq A2*), and Eq (224):

$$\begin{aligned} W\Big(\frac{\Phi(f(t_0))}{e} + \frac{\eta\, c_+}{e}\, t + \frac{2\eta\, c_+}{e} f(t)\Big) &= \log\big(t + \Theta(\log t)\big) - \log\log\big(t + \Theta(\log t)\big) + O(1) \\ &= \log t - \log\log t + O(1), \\ W\Big(\frac{\Phi(f(t_0))}{e} + \frac{\eta\, c_-}{2e}\, t\Big) &= \log t - \log\log t + O(1). \end{aligned}$$

Hence

$$f(t) = \log t - \log\log t + O(1). \tag{242}$$

□

(v) Conclusion From (231) and (242) (Step 2):

$$\mathbf{w}(t) = f(t)\,\hat{\mathbf{w}} + \boldsymbol{\delta}(t) = \hat{\mathbf{w}}\,(\log t - \log\log t + O(1)) + O(1) = \hat{\mathbf{w}}\log t - \hat{\mathbf{w}}\log\log t + O(1).$$

This completes the proof of Proposition I.5.

I.3 Generalisation to variable step-size

The argument extends to step-sizes (h_t) satisfying condition (16). We propose a version of Proposition I.5 and how to adapt the proof.

Proposition I.10 (Variable step-size)**.** *Under Assumptions I.1–I.4, let* $\mathbf{w}(t)$ *denote the iterates of population gradient descent* (207) *with any initial point* $\mathbf{w}(0)$*, and assume our scheduler* $(\eta_t)_t$ *satisfies*

$$2/\beta > \eta_t > 0, \qquad \sum \eta_t = +\infty, \qquad \sum \eta_t^2 < \infty,$$

with β *introduced in Assumption I.2. Then, as* $t \to \infty$*:*

$$\mathbf{w}(t) \;=\; \hat{\mathbf{w}}\log z_t \;-\; \hat{\mathbf{w}}\log\log z_t \;+\; O(1), \tag{243}$$

with $z_t := \sum_{k=0}^{t-1} \eta_k$*.*

Proof. The proof of Lemma I.6 is still mostly valid with a little rearrangement of the deductions. Eq (218) is still valid with η_t instead of η, this means that $\mathcal{L}(\mathbf{w}(t))$ decreases to a limit $\mathcal{L}_\infty$. By Eq (213), $\|\nabla\mathcal{L}\| \geq \mathcal{L}/\|\hat{\mathbf{w}}\|$, so $\|\nabla\mathcal{L}(\mathbf{w}(t))\| \geq \mathcal{L}_\infty/(2\|\hat{\mathbf{w}}\|)$ for $t \geq t_0$. From the proof of Soudry et al. (2018, *Lemma 10*), we have this updated result:

$$\sum_{t=0}^{\infty} \eta_t \Big(1 - \frac{\eta_t\beta}{2}\Big)\|\nabla\mathcal{L}(\mathbf{w}(t))\|^2 < \infty, \tag{244}$$

and given that $\sum_{t=0}^{\infty}\eta_t = \infty$, this implies $\mathcal{L}_\infty = 0$ and $\|\nabla\mathcal{L}(\mathbf{w}(t))\| \to 0$.

The proof of Lemma I.7 undergoes the following adjustments. Eq (220) becomes $\frac{1}{\mathcal{L}(\mathbf{w}(t+1))} - \frac{1}{\mathcal{L}(\mathbf{w}(t))} \leq 4\eta_t K^2$, summing it gives $\mathcal{L}(\mathbf{w}(t)) \geq c'/z_t$ for $c' := \frac{1}{4K^2+1/\mathcal{L}(\mathbf{w}(0))} > 0$, instead of Eq (221). A similar argument gives $\mathcal{L}(\mathbf{w}(t)) \leq C/z_t$ with an explicit constant C. Collecting everything yields $\mathcal{L}(\mathbf{w}(t)) = \Theta(1/z_t)$ and $f(t) = \Theta(\log z_t)$ by the same integral comparison argument.

The proof of Lemma I.8 is valid textually by using the fact that $(\eta_t)_t$ is a bounded sequence.

Regarding Lemma I.9, the proof needs the following adjustment. The sums Eqs (239) and (240) become

$$\Phi(f(t)) \leq \Phi(f(t_0)) + c_+\, z_t + 2\|\eta\|_\infty\, c_+\, f(t), \tag{245}$$

$$\Phi(f(t)) \geq \Phi(f(t_0)) + \frac{c_-}{2}\, z_t, \tag{246}$$

where $\|\eta\|_\infty := \sup_{t\geq 0}\eta_t$. The same Lambert W asymptotics yields the same analogous result $f(t) = \log z_t - \log\log z_t + O(1)$, which concludes the proof. □

I.4 Differences between the finite and the continuous settings

Table 3 summarises the structural differences between *Expected* and *Empirical* Risk Minimization settings Soudry et al. (2018).

(i) Origin of the $\log\log t$ *term in Eq* (14)*.* In the finite-sample setting, the gradient at time t is a finite sum over support vectors, and its leading term matches $\hat{w}/t$ exactly via Soudry et al. (2018, *Eq 21*), since the residual gradient $\dot{r}$ (reusing their notations) is strictly dominated because it is integrable. In the continuous-distribution setting, the gradient is a Laplace integral (via the disintegration theorem, Eq 233). By Watson's Lemma the leading term is $t^{-1}/\log t$. The additional $1/\log t$ factor creates a systematic mismatch between the gradient and $\hat{w}/t$, which accumulates over time into the

Table 3: Summary of differences from Soudry et al. (2018). γ_g is the group-specific margin, ε_g the group proportion, $\tilde{\gamma}_g$ the group-specific **r**-only margin.

	Finite dataset (Soudry)	**Continuous dist. (ours)**
Gradient structure	finite sum	Laplace integral
Support set	finite $\mathcal{S}$ with $\lvert\mathcal{S}\rvert \leq d$	$\{\mathbf{x} \cdot \hat{w} = 1\}$, measure zero
Residual $\rho(t)$	$O(1)$ (Thm. 9)	$-\hat{w} \log\log z_t + O(1)$
Direction convergence	$O(\log\log t / \log t)$	$O(\log\log z_t / \log z_t)$
Classification error	$\Theta(t^{-1})$	$\kappa_g^t(\tilde{\gamma}_g)/(z_t\,\varepsilon_g)$ * $\kappa_g^t(\tilde{\gamma}_g)/(z_t\,\varepsilon_g)$ and $\Theta\Big(z_t^{-\alpha_\xi(1+\delta_{\bar{g}})+\delta_g}(\log z_t)^{\alpha_\xi-1}\Big)$ ** $\Theta\big(z_t^{-\gamma_g}(\log z_t)^{\gamma_g+\delta_g+\tilde{\delta}_g-1}\big)$ ***

* Decay rate in the isotropic regime ($\alpha<1$).
** Decay rate in the isotropic regime ($\alpha>1$), $g = \min$
*** General coarse decay rate.

$-\hat{w} \log\log t$ correction. This mechanism is entirely distinct from the degeneracy-driven $O(\log\log t)$ growth in Soudry et al. (2018, *Theorem 13*).

(ii) Absence of the zero-activation degeneracy. In the finite-sample setting, when certain support vectors have exactly zero activation (i.e. $\hat{w} \cdot x_i = 0$ for some support vector x_i), the residual $\rho(t)$ can grow as $O(\log\log t)$ rather than remaining bounded – the degenerate case treated in Soudry et al. (2018, *Theorem 13*). Our continuous-distribution framework sidesteps this issue entirely: the margin density $\lambda(z)$ is assumed strictly positive and continuous at the boundary $z = 1$, so the gradient contribution is spread over a continuum of margin values rather than concentrated on isolated support vectors. This is at once a theoretical strength – the analysis is cleaner and the resulting rates are sharper – and a modelling limitation, since finite datasets can exhibit such degeneracies in practice. The finite-sample concentration analysis that would bridge the two regimes is left to future work.