How Cross-Entropy Learns Data Modes: Emergence and Implicit Bias in the Unconstrained Features Model

Anonymous authors
Paper under double-blind review

Abstract

A classical result for linear networks trained with mean-squared-error loss is that gradient flow learns the singular modes of the data sequentially, in order of importance. The precise mechanics of such sequential emergence under cross-entropy loss remain largely unknown. We study this question in a minimal nonconvex setting: a two-layer linear network with orthogonal inputs and step-imbalanced classes, equivalent to the unconstrained feature model used in neural collapse analyses. First, we characterize the regularization path in spectral coordinates, showing sequential mode emergence with activation thresholds that, unlike in MSE, need not coincide with the corresponding data singular values. As regularization vanishes, active singular values diverge, while their normalized values converge at a logarithmic rate and can approach their limits from either side. Second, for gradient flow under suitable spectral initialization, we prove majority-first emergence and show that the majority-to-minority mode ratio converges to that of the regularization-path limit. Our analysis relies on a novel imbalance-adapted Hadamard basis in which we show that softmax preserves a diagonal-plus-rank-one structure.

1 Introduction

The training dynamics of deep linear networks under mean-squared-error (MSE) loss have shaped much of deep learning theory (Arora et al., 2018b;a; Gunasekar et al., 2017; Li et al., 2020; Tu et al., 2024; Bach, 2026). A particularly influential example is the spectral-dynamics framework of Saxe et al. (2014; 2019) for *regression*, which showed that, under gradient flow, dominant data modes are learned sequentially and emerge in the network representations. This *sequential-learning* phenomenon has since informed numerous theoretical accounts (e.g., Jacot et al., 2021; Bantzis et al., 2025; Zhang et al., 2025; Musat, 2025; Vasudeva et al., 2025).

By contrast, the analogous picture for *classification* with cross-entropy (CE) loss remains largely incomplete. We still lack a precise characterization of whether representations exhibit comparable emergence, how such emergence unfolds, and which classifier CE selects at convergence. These questions are difficult even in simple models because of the softmax nonlinearity.

Motivated by this gap, we investigate CE dynamics in a two-layer linear neural network with orthogonal inputs under a step-imbalanced classification: K classes partitioned into equal numbers of majority and minority classes, with imbalance ratio $R > 1$. The choice of data and architecture is deliberately minimal. Concretely, *(i)* step-imbalanced data is the minimal label-symmetry distortion that produces a nontrivial mode structure consisting of majority, majority-minority, and minority modes (Zhao & Thrampoulidis, 2025); *(ii)* a non-convex two-layer linear architecture, unlike a single-layer convex model, allows this structure to emerge sequentially; *(iii)* orthogonality simplifies the analysis, matching the original setting of Saxe et al. (2014; 2019), and makes the model coincide with the unconstrained features model (UFM) (Mixon et al., 2020) used in neural-collapse analyses, giving this otherwise minimal model a principled connection to deep networks, whose last-layer geometry the UFM has been shown to predict robustly. Our contributions are as follows:

[1] See App. A for the connection between the RP results shown here and the GF analysis of Saxe et al. (2014).

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.

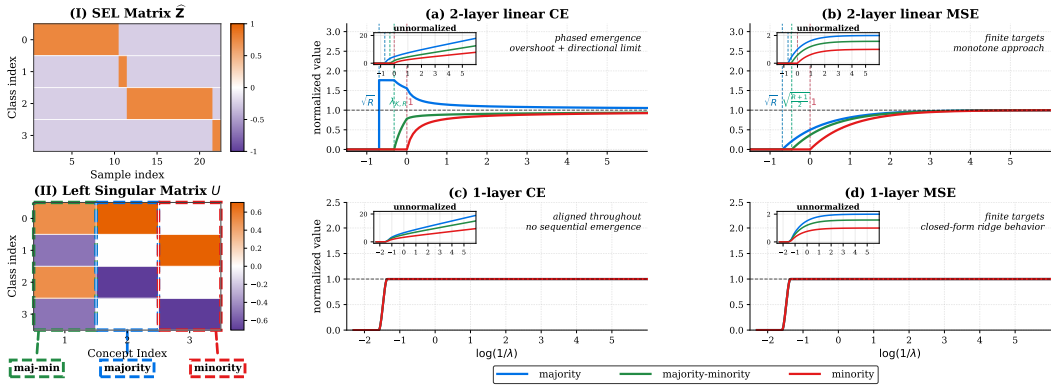


Figure 1: We study sequential learning in step-imbalanced classification with $K/2$ majority and $K/2$ minority classes and imbalance ratio $R > 1$. **(I)** The SEL matrix $\widehat{\boldsymbol{Z}} = (\boldsymbol{I} - \frac{1}{K}\mathbf{1}\mathbf{1}^\top)\boldsymbol{Y}$, a centering of the one-hot encoding matrix, has three distinct singular modes: majority–minority, majority, and minority. **(II)** These are reflected in the left singular matrix $\boldsymbol{U}$ (Zhao & Thrampoulidis, 2025). Panels (a)–(d) show the evolution of logit singular values along the ridge-regularization path ($\lambda > 0$) for two-layer CE, two-layer MSE, one-layer CE, and one-layer MSE networks with orthogonal inputs. All coordinates are normalized so that the limiting profile equals one; unnormalized profiles shown in insets. **(a)** CE exhibits sequential mode activation, but with activation thresholds that differ from those under MSE. Unnormalized singular values diverge, and normalized modes approach their limiting values from different sides. **(b)** MSE exhibits sequential activation and monotone convergence to finite singular-value targets.[1] **(c, d)** One-layer CE and MSE paths show *no* sequential emergence.

• **Approximate softmax diagonalization.** We identify spectral coordinates adapted to the imbalanced label structure in which the softmax operator preserves a *diagonal-plus-rank-one form* (Thm. 1). This approximate diagonalization reduces the matrix-valued CE problem to a lower-dimensional one, pushing the spectral-dynamics framework beyond exact diagonalization of prior works.

• **Regularization path (RP).** Using this reduction, we reduce the ridge-regularized UFM to a four-dimensional convex optimization (Thm. 2), which we then solve explicitly to find sequential spectral emergence with majority activating first, then majority–minority, and finally the minority, as regularization λ decreases (Prop. 1). From this, we obtain an explicit rate for the convergence rate of the modes as $\lambda \to 0$ (Thm. 3), and establish the side from which each normalized mode approaches its limit (Thm. 4). These allow direct comparison of CE with MSE; Fig. 1.

• **Gradient-flow (GF) dynamics.** In Sec. 5, we perform an analogous study of GF under a spectrally aligned initialization compatible with the approximate diagonalization. Using this initialization, we reduce the flow to a vector-valued ODE (Thm. 5). Although the coordinates no longer decouple as in MSE, we prove majority-first emergence (Thm. 6), characterize positive radial points of the flow as RP solutions (Prop. 2), and show that deviations from uniformity within the majority and minority modes decrease monotonically in KL divergence (Thm. 7). Under uniformity, we further prove that the majority and minority modes grow logarithmically and their ratio converges to that of the RP limit (Thm. 8).

Closely related work. Garrod et al. (2025) recently showed that the spectral-dynamics framework for MSE by Saxe et al. (2014; 2019), can be pushed through to CE despite the softmax nonlinearity. We extend this: *(a) Beyond pure diagonalization.* We show that an approximate diagonalization suffices to reduce the dynamics to low dimension. We expect this relaxation to be broadly useful, since prior exact diagonalization is restrictive outside symmetric settings. *(b) Phased mode emergence.* We give the first explicit analysis of sequential-emergence under CE since the balanced regime in (Garrod et al., 2025) has no distinct singular modes. *(c) Regularization path.* We apply the spectral framework to the regularization path, connecting it directly to the GF path. Compared to GF, the RP is better studied in the neural-collapse literature, with the vanishing-regularization limit obtained by Thrampoulidis et al. (2022) and the entire solution path for all λ characterized by Hong & Ling (2023). We revisit these characterizations in spectral coordinates, deriving them directly with the same softmax-diagonalization tool that we use for the GF. This also makes explicit that the solution of Hong & Ling (2023) converges to the geometry of Thrampoulidis et al. (2022), and yields finer

properties (such as, the rate of convergence and the side from which each mode approaches its limit) which allow a direct comparison of CE with MSE. More broadly, by zooming in on a minimal model, studying the RP and GF jointly, and contrasting CE with MSE, our results connect often separate literatures on implicit bias, neural collapse, and spectral dynamics of linear networks.

Notation. Define $[N] \coloneqq \{0, \dots, N-1\}$ for $N \in \mathbb{N}$. Let K classes, set $M \coloneqq K/2$, and use R for the imbalance ratio. Lowercase/uppercase boldface denote vectors/matrices; $\otimes$ and $\odot$ denote Kronecker and Hadamard products. $\boldsymbol{e}_i$ denotes the i-th standard basis vector; to reduce clutter, its dimension is kept implicit when clear from context. We reserve $\boldsymbol{e}_+ \coloneqq (1,0)^\top$ and $\boldsymbol{e}_- \coloneqq (0,1)^\top$ for basis in $\mathbb{R}^2$. Throughout, subscripts $+/-$ flag quantities associated with majorities/minorities. For matrices, $\overline{\boldsymbol{A}}$ denotes L2-normalization. For vectors, $\overline{\boldsymbol{a}}$ denotes L1-normalization. $\|\boldsymbol{A}\|_*$ is the nuclear norm. Inequalities on vectors apply entrywise. $\mathbb{S}(\cdot)$ denotes the softmax map. We denote the Sylvester–Hadamard matrix of order $K = 2^m, m \in \mathbb{N}$ by $\boldsymbol{\Phi}_K \in \{\pm 1\}^{K \times K}$. With $\boldsymbol{\Phi}_2 = \left(\begin{smallmatrix} 1 & 1 \\ 1 & -1 \end{smallmatrix}\right)$, it satisfies the recursion $\boldsymbol{\Phi}_K = \boldsymbol{\Phi}_M \otimes \boldsymbol{\Phi}_2$. We write $\overline{\boldsymbol{\Phi}}_K \in \{\pm 1\}^{K \times (K-1)}$ for the same matrix with its first (all-ones) column removed.

2 Problem Setup and Background

Data. We study imbalanced multiclass classification over $K = 2^m, m \in \mathbb{N}$ classes with n input-label pairs. Let $R \in \mathbb{N}$ denoting the *imbalance ratio*. Data comprises M majority classes with Rn' examples each and M minority classes with n' examples each, where $M \coloneqq K/2$. Without restricting generality of the findings we set $n' = 1$ henceforth; thus $n = (R+1)M$ total examples. For convenience in stating our results, we order classes so that even indices $(0, 2, \dots)$ correspond to majorities and odd indices $(1, 3, \dots)$ to minorities. The one-hot encoding matrix then takes the form

$$\boldsymbol{Y} \coloneqq \boldsymbol{I}_M \otimes \begin{pmatrix} \mathbf{1}_R^\top & 0 \\ \mathbf{0}_R^\top & 1 \end{pmatrix} \in \mathbb{R}^{K \times n}. \tag{1}$$

We also define the *simplex-encoding label (SEL) matrix*, i.e., $\boldsymbol{Y}$ with each column centered:

$$\widehat{\boldsymbol{Z}} \coloneqq \left(\boldsymbol{I}_K - \tfrac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\boldsymbol{Y}.$$

When data are balanced ($R = 1$), $\widehat{\boldsymbol{Z}} = \boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$ and its singular vectors form a simplex equiangular tight frame (ETF) (Papyan et al., 2020). For $R > 1$, $\widehat{\boldsymbol{Z}}$ has three distinct nonzero singular values, i.e., $\sqrt{R}$ and 1, each with multiplicity $M-1$, and $\sqrt{(R+1)/2}$ with multiplicity one. Following Zhao & Thrampoulidis (2025), we call these the *majority*, *minority* and *majority–minority* modes, respectively, since their left singular vectors are supported on the majority classes, are supported on the minority classes, and contrast majorities vs. minorities (see Fig. 1(II), and Sec. 3 for an explicit basis).

Unconstrained Features Model (UFM). The UFM reduces DNN optimization to a bilinear problem over the classifier $\boldsymbol{W} \in \mathbb{R}^{K \times d}$ and a freely optimized embedding matrix $\boldsymbol{H} \in \mathbb{R}^{d \times n}$ (column $\boldsymbol{h}_i$ is the last-layer representation of each example) (Mixon et al., 2020; Fang et al., 2021). Throughout, let $d \geq K$. Then, dropping the DNN's parameterization of the embedding map, the UFM optimizes:

$$\mathcal{L}(\boldsymbol{W}\boldsymbol{H}) \coloneqq \sum\nolimits_{i \in [n]} -\log\left(\boldsymbol{e}_{y_i}^\top \mathbb{S}\left(\boldsymbol{W}\boldsymbol{h}_i\right)\right) = \sum\nolimits_{i \in [n]} \log\left(1 + \sum\nolimits_{c \neq y_i} e^{-(\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i}\right). \tag{UFM}$$

Equivalently, (UFM) can be viewed as CE optimization of a two-layer linear network with orthogonal inputs $\boldsymbol{e}_i \in \mathbb{R}^n$, trainable weight matrices $\boldsymbol{H}$ and $\boldsymbol{W}$, and K outputs.

The (UFM) loss has *no* finite minimizers: any direction pair $(\overline{\boldsymbol{W}}, \overline{\boldsymbol{H}})$ with strictly positive margin drives the loss to zero as parameter norm diverges. We therefore study two complementary ways of making its asymptotic behavior precise: ridge regularization, which defines finite minimizers for each regularization strength, and unregularized gradient flow, which specifies an optimization trajectory.

Regularization path (RP). The ridge-regularized UFM,

$$(\boldsymbol{W}_\lambda, \boldsymbol{H}_\lambda) \in \arg\min_{\boldsymbol{W},\boldsymbol{H}} \mathcal{L}(\boldsymbol{W}\boldsymbol{H}) + \lambda/2\|\boldsymbol{W}\|_F^2 + \lambda/2\|\boldsymbol{H}\|_F^2 \tag{UFM$_\lambda$}$$

models DNN training with weight decay (e.g., Zhu et al., 2021; Han et al., 2021; Súkeník et al., 2024; Garrod & Keating, 2024; Hong & Ling, 2023; Dang et al., 2024). For $\lambda > 0$ (UFM$_\lambda$)

has finite minimizers, which diverge as $\lambda \to 0^+$, but converge in direction (Rosset et al., 2003; Thrampoulidis et al., 2022). A key property that makes the RP accessible for analysis is that its semidefinite-programming (SDP) relaxation in logit space $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{H}$ is tight (Zhu et al., 2021):

$$\boldsymbol{Z}_\lambda = \arg\min_{\boldsymbol{Z}} \mathcal{L}(\boldsymbol{Z}) + \lambda \|\boldsymbol{Z}\|_* \,. \tag{Z-UFM$_\lambda$}$$

Concretely, for all $\lambda > 0$, there exists orthogonal matrix $\boldsymbol{R}\,(\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}_K)$ such that $\boldsymbol{W}_\lambda = \boldsymbol{U}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{R}^\top$ and $\boldsymbol{H}_\lambda = \boldsymbol{R}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{V}^\top$ where $\boldsymbol{Z}_\lambda = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$ is the SVD of the solution to (Z-UFM$_\lambda$). For $R = 1$, it is known for all $\lambda < \lambda_{\max}$ (above which the solution is trivially zero) that $\boldsymbol{Z}_\lambda \propto \widehat{\boldsymbol{Z}}$; thus, both at finite λ and asymptotically as $\lambda \to 0^+$, the weights $(\boldsymbol{W}_\lambda, \boldsymbol{H}_\lambda)$ form a simplex ETF (Zhu et al., 2021). But for $R > 1$, the picture changes: The solution varies nontrivially with λ, and although the normalized logits $\overline{\boldsymbol{Z}}_\lambda$ are known to become proportional to the SEL matrix $\widehat{\boldsymbol{Z}}$ as $\lambda \to 0$, this proportionality fails at finite λ (Thrampoulidis et al., 2022). Indeed, Hong & Ling (2023) provide a characterization $\boldsymbol{Z}_\lambda$ for all λ and show that as λ decreases it goes through four regimes that differ in which blocks of $\boldsymbol{Z}_\lambda$ are nonzero. Here, we ask: *how do these regimes translate into sequential emergence of the majority, majority–minority, and minority modes? Can a direct spectral analysis of the RP provide a framework that also applies to the GF dynamics? How does the resulting picture compare with MSE?*

Gradient flow (GF). The GF dynamics for optimizing the (UFM) are as follows:

$$\dot{\boldsymbol{W}}_t = (\boldsymbol{Y} - \mathbb{S}\,(\boldsymbol{W}_t\boldsymbol{H}_t))\boldsymbol{H}_t^\top \,, \qquad \dot{\boldsymbol{H}}_t = \boldsymbol{W}_t^\top(\boldsymbol{Y} - \mathbb{S}\,(\boldsymbol{W}_t\boldsymbol{H}_t)) \,. \tag{2}$$

Despite their simplicity, (UFM) remains non-convex, so convergence of the GF iterates is itself non-trivial. Specifically, while it is known that $(\overline{\boldsymbol{W}}_t, \overline{\boldsymbol{H}}_t)$ must converge to *some* KKT point of the following *non-convex* max-margin problem (Lyu & Li, 2020; Ji et al., 2022):

$$\min_{\boldsymbol{W},\boldsymbol{H}} \ \|\boldsymbol{W}\|_F^2 + \|\boldsymbol{H}\|_F^2 \qquad \text{sub. to} \quad (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^\top \boldsymbol{W}\boldsymbol{h}_i \geq 1, \ \forall i, c \neq y_i, \tag{3}$$

such a characterization is *not* explicit. There are many KKT points, each with different spectral-mode structure, and, in general, convergence can be either at global or local minima and even at saddles of the max-margin problem in (3) (Vardi et al., 2022). In the binary case ($K = 2$), Ji & Telgarsky (2018) show that the GF dynamics asymptotically satisfy the balancedness condition $\overline{\boldsymbol{W}}_t^\top \overline{\boldsymbol{W}}_t - \overline{\boldsymbol{H}}_t \overline{\boldsymbol{H}}_t^\top = 0$, and Vardi et al. (2022) establish convergence to a global optimum of (3). But $K = 2$ is both restrictive and trivial in our setting, since the only KKT point is the global minimizer. For $K > 2$, Garrod et al. (2025) recently introduced a Hadamard spectral initialization under which the dynamics remain on a fixed subspace, leaving only the singular values to evolve and making analysis tractable. Their construction, however, applies only to $R = 1$, where the SEL matrix has a single distinct nonzero singular value. Here, we ask: *how does gradient flow learn the distinct data modes introduced by class imbalance, and how does its trajectory relate to the regularization path?*

3 Technical Tool: Approximate Softmax Diagonalization

Both our RP and GF analyses rely on a technical observation about how the softmax operator interacts with the singular subspace of the SEL matrix $\widehat{\boldsymbol{Z}}$. We first choose specific left and right singular bases $\boldsymbol{U} \in \mathbb{R}^{K\times(K-1)}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times(K-1)}$, constructed from the Sylvester–Hadamard basis $\overline{\boldsymbol{\Phi}}_M$:

$$\boldsymbol{U} \coloneqq \begin{bmatrix} \frac{1}{\sqrt{2M}}\mathbf{1}_M \otimes \begin{pmatrix} 1 \\ -1 \end{pmatrix} & \frac{1}{\sqrt{M}}\overline{\boldsymbol{\Phi}}_M \otimes \boldsymbol{I}_2 \end{bmatrix}, \quad \boldsymbol{V} \coloneqq \begin{bmatrix} \frac{1}{\sqrt{M(R+1)}}\mathbf{1}_M \otimes \begin{pmatrix} \mathbf{1}_R \\ -1 \end{pmatrix} & \frac{1}{\sqrt{M}}\overline{\boldsymbol{\Phi}}_M \otimes \begin{pmatrix} \frac{\mathbf{1}_R}{\sqrt{R}} & \frac{\mathbf{0}_R}{\sqrt{R}} \\ 0 & 1 \end{pmatrix} \end{bmatrix}.$$

Then, we introduce a rank-1 extension of this basis $\widetilde{\boldsymbol{V}} \coloneqq [\widetilde{\boldsymbol{v}}\,,\, \boldsymbol{V}]$, with $\widetilde{\boldsymbol{v}} \coloneqq \frac{1}{\sqrt{MR(R+1)}}\,\mathbf{1}_M \otimes \begin{pmatrix} \mathbf{1}_R \\ R \end{pmatrix}$. Finally, we define an operator $\mathrm{diag}_{+1}(\cdot) : \mathbb{R}\times\mathbb{R}\times\mathbb{R}^{M-1}\times\mathbb{R}^{M-1} \to \mathbb{R}^{(K-1)\times K}$ that combines a $(K-1)\times(K-1)$ diagonal matrix and a scaled first standard basis vector:

$$\mathrm{diag}_{+1}(\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-) \coloneqq [\alpha\,\boldsymbol{e}_0 \quad \boldsymbol{A}], \qquad \boldsymbol{A} \coloneqq \mathrm{diag}\,(a,\, \boldsymbol{a}_+ \otimes \boldsymbol{e}_+ + \boldsymbol{a}_- \otimes \boldsymbol{e}_-)\,.$$

Pre- and post- multiplying the operator by $\boldsymbol{U}$ and $\widetilde{\boldsymbol{V}}^\top$ gives $\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}^\top + \alpha\boldsymbol{u}_0\widetilde{\boldsymbol{v}}^\top$, where $\boldsymbol{u}_0$ is the first column of $\boldsymbol{U}$. This SVD-like representation augmented by a rank-one term is what we need to state our main result below.

Theorem 1 (Softmax Diagonalization up to Rank-1). *For $\alpha, a \in \mathbb{R}, \boldsymbol{a}_\pm \in \mathbb{R}^{M-1}$ it holds*

$$\boldsymbol{Y} - \mathbb{S}\left(\boldsymbol{U}\,\mathrm{diag}_{+1}\left(\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-\right)\,\widetilde{\boldsymbol{V}}^\top\right) = \boldsymbol{U}\,\mathrm{diag}_{+1}\left(\beta, b, \boldsymbol{b}_+, \boldsymbol{b}_-\right)\,\widetilde{\boldsymbol{V}}^\top. \tag{4}$$

where the map $(\beta, b, \boldsymbol{b}_+, \boldsymbol{b}_-) \coloneqq \mathcal{F}\left((\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-)\right)$ is given in closed-form in (5) below.

To interpret the theorem, recall that $\boldsymbol{Y} = \widehat{\boldsymbol{Z}} + \frac{1}{K}\mathbf{1}_K\mathbf{1}_n^\top$ and that softmax is column-stochastic. Writing $\boldsymbol{a} = (\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-)$ and $\boldsymbol{Z}_{\boldsymbol{a}} = \boldsymbol{U}\,\mathrm{diag}_{+1}\left(\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-\right)\,\widetilde{\boldsymbol{V}}^\top$, (4) gives $\mathbb{S}\left(\boldsymbol{Z}_{\boldsymbol{a}}\right) - \frac{1}{K}\mathbf{1}_K\mathbf{1}_n^\top = \widehat{\boldsymbol{Z}} - \boldsymbol{Z}_{\mathcal{F}(\boldsymbol{a})}$. Since $\widehat{\boldsymbol{Z}}$ is diagonalizable in the bases $\boldsymbol{U}, \boldsymbol{V}$ and thus itself belongs to this family, the theorem states that the *column-centered softmax preserves the diagonal-plus-rank-one structure*.

What makes this statement non-trivial is the softmax nonlinearity, and it holds only thanks to a careful Hadamard parameterization of $\boldsymbol{U}, \widetilde{\boldsymbol{V}}$ and the rank-one extension $\widetilde{\boldsymbol{v}}$. While we build upon the Hadamard-basis idea of Garrod et al. (2025), the diagonal family is no longer preserved under class imbalance. The novelty and main challenge here lie in identifying the parameterization and rank-one extension that restore closure, and proving that no further components are generated by the softmax (App. D; see also the proof sketch below).

Beyond this preservation property, what makes the theorem useful for the subsequent analysis is that we can explicitly identify the mapping $\mathcal{F}(\cdot)$ from the input spectral parameters $(\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-)$ to the residual parameters $(\beta, b, \boldsymbol{b}_+, \boldsymbol{b}_-)$:

$$\beta = \sqrt{\tfrac{R}{2(R+1)}}\left(X_+ - X_-\right), \qquad b = \sqrt{\tfrac{1}{2(R+1)}}\left(RX_+ + X_-\right), \tag{5a}$$

$$\boldsymbol{b}_+ = \sqrt{R}\big(\mathbf{1}_{M-1} - \tfrac{1}{2M}\,X_+\,\overline{\boldsymbol{\Phi}}_M^\top \exp(\boldsymbol{x}_+)\big), \qquad \boldsymbol{b}_- = \mathbf{1}_{M-1} - \tfrac{1}{2M}\,X_-\,\overline{\boldsymbol{\Phi}}_M^\top \exp(\boldsymbol{x}_-), \tag{5b}$$

with $X_\pm \coloneqq \frac{2M}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_\pm) + M}$, $\boldsymbol{x}_+ \coloneqq \frac{1}{M}\big(\gamma\mathbf{1}_M + \frac{1}{\sqrt{R}}\overline{\boldsymbol{\Phi}}_M\boldsymbol{a}_+\big)$, $\boldsymbol{x}_- \coloneqq \frac{1}{M}\big(\delta\mathbf{1}_M + \overline{\boldsymbol{\Phi}}_M\boldsymbol{a}_-\big)$, and,

$$(\gamma\,,\,\delta) \coloneqq \sqrt{\tfrac{2}{R+1}}\Big(a + \tfrac{1}{\sqrt{R}}\,\alpha\,,\,a - \alpha\sqrt{R}\Big)\,. \tag{6}$$

At first, these formulas may seem dauntingly complicated, but they have a nice structure. The intuition to keep in mind as we use them throughout the paper is as follows:

- The spectral parameters $\boldsymbol{a}_\pm$ describe the majority and minority modes, while the pair (α, a) describes the majority–minority mode.
- The majority–minority mode requires *two* parameters because its right singular direction can vary within a fixed two-dimensional subspace. The pair (α, a) tracks both its magnitude and this direction while keeping the basis fixed.
- The transformation (6) is invertible, so (γ, δ) and (α, a) are interchangeable descriptions of the majority–minority mode. We mainly use (γ, δ) below because they lead to cleaner expressions.
- The parameters $\boldsymbol{b}_\pm$ and (β, b) describe the corresponding modes of the softmax residual. The formulas (5) show that the modes remain coupled: (β, b) depends on all the input parameters, while $\boldsymbol{b}_\pm$ depend on the majority–minority pair through (γ, δ) in $\boldsymbol{x}_\pm$.

Proof sketch. The specific property of Hadamard matrices that we leverage is *phase-shift under permutations*: that is, $\boldsymbol{\Pi}_j\boldsymbol{\Phi}_M\boldsymbol{e}_i = (-1)^{\boldsymbol{i}^\top\boldsymbol{j}}\boldsymbol{\Phi}_M\boldsymbol{e}_i$ where $\boldsymbol{\Pi}_j$ is XOR ($\oplus$) permutation and $\boldsymbol{i}, \boldsymbol{j}$ are binary representations of indices i, j. In a bit more detail, the key was to realize that the residual matrix $\boldsymbol{Q}$ on the LHS of (4) can decompose its $n = (R+1)M$ columns into M blocks $\boldsymbol{Q}_i$ satisfying $\boldsymbol{Q}_i = (\boldsymbol{\Pi}_i \otimes \boldsymbol{I}_2)\begin{bmatrix}\boldsymbol{p}_+\mathbf{1}_R^\top & \boldsymbol{p}_-\end{bmatrix} \in \mathbb{R}^{K\times(R+1)}$, for zero-sum $\boldsymbol{p}_+ \coloneqq \boldsymbol{e}_0 - \mathbb{S}\left(\boldsymbol{x}_+ \otimes \boldsymbol{e}_+\right)$ and $\boldsymbol{p}_- \coloneqq \boldsymbol{e}_1 - \mathbb{S}\left(\boldsymbol{x}_- \otimes \boldsymbol{e}_-\right)$. In turn, it can be shown that any matrix with such structure can be decomposed into a block-diagonal-plus-rank-one structure, with the off-diagonal entries vanishing thanks to the specific structure of $\boldsymbol{p}_+, \boldsymbol{p}_-$.

4 Regularization Path

As highlighted in Sec. 2, to study (UFM$_\lambda$), it suffices to study the SDP (Z-UFM$_\lambda$) in the logit space. To begin, we use Thm. 1 to reduce the latter to a strictly convex problem in only four scalars.

Theorem 2 (SDP-Vectorization). *For $\lambda > 0$, the unique minimizer of* (Z-UFM$_\lambda$) *decomposes as* $\boldsymbol{Z}_\lambda = \boldsymbol{U}\,\mathrm{diag}_{+1}\left(\alpha_\lambda, a_\lambda, \boldsymbol{a}_{+,\lambda}, \boldsymbol{a}_{-,\lambda}\right)\widetilde{\boldsymbol{V}}^\top$, *where* $\alpha_\lambda, a_\lambda, \boldsymbol{a}_{+,\lambda}, \boldsymbol{a}_{-,\lambda}$ *are unique minimizers of*

$$\min_{\alpha,a,\boldsymbol{a}_\pm} \mathcal{L}\big(\boldsymbol{U}\,\mathrm{diag}_{+1}\left(\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-\right)\widetilde{\boldsymbol{V}}^\top\big) + \lambda\big(\sqrt{\alpha^2+a^2} + \|\boldsymbol{a}_+\|_1 + \|\boldsymbol{a}_-\|_1\big)\,. \tag{7}$$

Moreover, $\boldsymbol{a}_{\pm,\lambda} = a_{\pm,\lambda}\mathbf{1}_{M-1}$. *Hence* (Z-UFM$_\lambda$) *reduces to a four-dimensional optimization.*

The crux of the proof is using Thm. 1 to show that the global minimizer of (Z-UFM$_\lambda$) lies on the $(\boldsymbol{U}, \widetilde{\boldsymbol{V}})$ subspace for all $\lambda > 0$. The optimality condition for (Z-UFM$_\lambda$) is $\boldsymbol{Y} - \mathbb{S}\left(\boldsymbol{Z}\right) \in \lambda\,\partial\|\boldsymbol{Z}\|_*$. For candidate $\boldsymbol{Z}_{\boldsymbol{a}} = \boldsymbol{U}\,\mathrm{diag}_{+1}\left(\boldsymbol{a}\right)\widetilde{\boldsymbol{V}}^\top$, Thm. 1 rewrites the LHS of this condition as $\boldsymbol{Z}_{\mathcal{F}(\boldsymbol{a})} = \boldsymbol{U}\,\mathrm{diag}_{+1}\left(\mathcal{F}\left(\boldsymbol{a}\right)\right)\widetilde{\boldsymbol{V}}^\top$, so it becomes $\boldsymbol{Z}_{\mathcal{F}(\boldsymbol{a})} \in \lambda\,\partial\|\boldsymbol{Z}_{\boldsymbol{a}}\|_*$, which by orthogonality of $\boldsymbol{U}, \widetilde{\boldsymbol{V}}$ is equivalent to the optimality condition of (7).

We may now solve the optimality conditions of (7) explicitly . To state the result, we use the (γ, δ)–parameterization of Eq. (6). Also, for $Q(x,\theta) \coloneqq \log(x^{M-1}((1+\theta)x+\theta))$, $\theta_{+,\lambda} \coloneqq M[\sqrt{R}\lambda^{-1}-1]_+$, and $\theta_{-,\lambda} \coloneqq M[\lambda^{-1}-1]_+$, define $\gamma_\lambda \coloneqq Q(s_\lambda, \theta_+)$, $\delta_\lambda \coloneqq Q(r_\lambda, \theta_-)$. Finally, let r_λ solve a scalar nonlinear equation and s_λ be a closed-form function of it, both given in App. E.3.

Proposition 1 (Analytic solution). *Let* $\boldsymbol{a}_\lambda = (\alpha_\lambda,\, a_\lambda,\, a_{+,\lambda},\, a_{-,\lambda})$ *the minimizer of* (7) *and* $\lambda_{K,R} \in [1, \sqrt{\frac{R+1}{2}}]$ *the unique solution of* $Q\big(2\lambda/\sqrt{2\lambda^2-1}-1\,,\,\theta_{+,\lambda}\big) = 0$. *Then,* $\boldsymbol{a}_\lambda$ *satisfies:*

(i) *For* $\lambda \geq \sqrt{R}$, $\boldsymbol{a}_\lambda = (0,0,0,0)$.
(ii) *For* $\lambda \in [\lambda_{K,R}, \sqrt{R})$, $\boldsymbol{a}_\lambda = (0,0,a_{+,\lambda},0)$, *where* $\gamma_\lambda = \delta_\lambda = 0$ *and* $a_{+,\lambda} = -M\sqrt{R}\log s_\lambda$.
(iii) *For* $\lambda \in [1, \lambda_{K,R})$, $\boldsymbol{a}_\lambda = (\alpha_\lambda, a_\lambda, a_{+,\lambda}, 0)$, *where* $a_{+,\lambda} = \sqrt{R}(\gamma_\lambda - M\log s_\lambda)$.
(iv) *For* $\lambda < 1$, $\boldsymbol{a}_\lambda = \big(\alpha_\lambda, a_\lambda, a_{+,\lambda}, a_{-,\lambda}\big)$, *where* $a_{+,\lambda} = \sqrt{R}(\gamma_\lambda - M\log s_\lambda)$, $a_{-,\lambda} = \delta_\lambda - M\log r_\lambda$.

Thus, as λ decreases, the majority mode ($a_{+,\lambda}$) appears first, then the majority–minority mode ($\alpha_\lambda, a_\lambda$), and finally the minority mode ($a_{-,\lambda}$).

With the four-dimensional reduction of Thm. 2 in hand, Prop. 1 follows through tedious but straightforward algebraic manipulation of the KKT conditions. In fact, these results can alternatively be obtained from (Hong & Ling, 2023) after reparameterization: the phase structure and $\lambda_{K,R}$ matches their Thm. 3.2(a)–(d), and the formulas for the spectral modes can, in retrospect, be obtained by algebraic manipulation of their formulas for the entries of $\boldsymbol{Z}_\lambda$ in Eqs. (5.32–5.36, 5.40–5.42), appearing in the proof of their Prop. 5.5. Here, we instead derive the results directly in spectral coordinates using Thm. 1 and the reduction it enables in Thm. 2. What follows builds on this spectral form.

Emergence thresholds. Prop. 1 identifies the λs at which modes first become nonzero. Majority and minority activate at $\lambda_+ \coloneqq \sqrt{R}$, $\lambda_- \coloneqq 1$, which coincide with the largest, smallest nonzero singular values of the SEL matrix, and do not vary with K. As expected, increased imbalance R widens the ratio λ_+/λ_-. The intermediate threshold $\lambda_{K,R}$, at which majority–minority mode emerges, is decreasing in K and increasing in R (but slower than λ_+). Although it has no closed form, it solves a scalar nonlinear equation. See App. E.5 for details and Figure 2 for visualization.

Convergence. Beyond the emergence thresholds, Prop. 1 characterizes the evolution within each phase through a single scalar nonlinear equation. In App. E.2, we show that $a_{\pm,\lambda} \geq 0$ for all $\lambda > 0$, with strict positivity after activation. Moreover, as $\lambda \to 0$, all three singular values diverge, but their direction converges to that of the SEL-matrix singular values.

Theorem 3 (Convergence to SEL). *Let* $\boldsymbol{a}_\lambda$ *be as in Proposition 1 and define* $\boldsymbol{a}_\infty = (0, \sqrt{(R+1)/2}, \sqrt{R}, 1)$. *As* $\lambda \to 0^+$, *the normalized spectral parameters satisfy* $\overline{\boldsymbol{a}}_\lambda = \overline{\boldsymbol{a}}_\infty + O\big(1/\log(\lambda^{-1})\big)$. *Consequently, the logit singular-value profile* $\boldsymbol{\sigma}_\lambda \coloneqq (\sqrt{a_\lambda^2+\alpha_\lambda^2}, a_{+,\lambda}, a_{-,\lambda})$ *converges in direction to* $\boldsymbol{\sigma}_\infty = (\sqrt{(R+1)/2}, \sqrt{R}, 1)$ *at the same rate.*

Unlike prior vanishing-regularization convergence results (e.g. (Ji et al., 2020; Thrampoulidis et al., 2022; Tarzanagh et al., 2023; 2024)), which to our knowledge all rely on a non-constructive proof by contradiction following (Rosset et al., 2003), the result above is explicit (by taking the vanishing regularization limit in the formulas of Prop. 1) *and* gives an explicit convergence rate.

Mode emergence without finite targets. Our spectral analysis highlights a key difference between CE and the classical MSE sequential-learning picture. Under MSE gradient flow with small, spectrally

aligned initialization, modes follow sigmoidal paths toward finite target singular values (Saxe et al., 2014; Gidel et al., 2019). Along the MSE regularization path, modes likewise approach their finite targets monotonically from below (App. A.2). Under CE, the active singular values instead diverge as $\lambda \to 0^+$ and converge directionally. Moreover, the normalized CE modes approach their SEL limits from different sides: the majority mode approaches from above, while the majority–minority and minority modes approach from below. Thm. 4 below establishes the corresponding inequalities throughout the nonzero regularization path. See Figs. 1,3 for illustrations. Numerically, we also find that the minority mode can approach its limiting normalized value faster than the majority–minority mode, depending on the values of K and R; see Fig. 8.

Theorem 4 (Direction of convergence of modes)**.** *With $\boldsymbol{\sigma}_\lambda = (\sigma_{0,\lambda}, \sigma_{+,\lambda}, \sigma_{-,\lambda})$ as in Thm. 3, where $0, +, -$ index the majority–minority, majority, and minority modes, respectively, define its normalized entries by $\bar{\sigma}_{i,\lambda} := \sigma_{i,\lambda}/\|\boldsymbol{a}_\lambda\|_{+1}$ for $i \in \{0, +, -\}$, where $\|\boldsymbol{a}_\lambda\|_{+1} := \sigma_{0,\lambda} + (M-1)(\sigma_{+,\lambda} + \sigma_{-,\lambda})$, and define $\bar{\sigma}_{i,\infty}$ analogously. Then, for all $0 < \lambda < \sqrt{R}$,*

$$\bar{\sigma}_{+,\lambda} \geq \bar{\sigma}_{+,\infty}, \qquad \bar{\sigma}_{0,\lambda} \leq \bar{\sigma}_{0,\infty}, \qquad \bar{\sigma}_{-,\lambda} \leq \bar{\sigma}_{-,\infty}. \tag{8}$$

5 Gradient-flow Path

To analyze the GF dynamics (2), we follow the program of (Saxe et al., 2014; Gidel et al., 2019; Garrod et al., 2025), and use a spectral initialization. Concretely, as suggested by Thm. 1, we initialize logits on the $\boldsymbol{U}, \widetilde{\boldsymbol{V}}$ subspace. This initialization is motivated by three observations: *(a)* The conjectured limit of GF, namely the SEL matrix $\widehat{\boldsymbol{Z}}$ (Thrampoulidis et al., 2022), admits this decomposition; *(b)* The entire regularization path lies exactly on this subspace, as shown in Sec. 4; and *(c)* it lets us leverage Thm. 1, which ensures that the logits remain in this decomposition at *all* times, turning the matrix ODE into a vector ODE. The latter is formalized below.

Theorem 5 (Invariance of spectral subspace)**.** *Initialize the GF updates in (2) such that $\boldsymbol{W}_0 = \boldsymbol{U}\,\mathrm{diag}\,(w_0, \boldsymbol{w}_{+,0}, \boldsymbol{w}_{-,0})\,\boldsymbol{R}^\top$ and $\boldsymbol{H}_0 = \boldsymbol{R}\,\mathrm{diag}_{+1}\,(\alpha_{H,0}, h_0, \boldsymbol{h}_{+,0}, \boldsymbol{h}_{-,0})\,\widetilde{\boldsymbol{V}}^\top$ for a partial orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{d\times(K-1)}$. Then, for all times t, it holds that*

$$\boldsymbol{W}_t = \boldsymbol{U}\,\mathrm{diag}\,(w_t, \boldsymbol{w}_{+,t}, \boldsymbol{w}_{-,t})\,\boldsymbol{R}^\top \quad \text{and} \quad \boldsymbol{H}_t = \boldsymbol{R}\,\mathrm{diag}_{+1}\,(\alpha_{H,t}, h_t, \boldsymbol{h}_{+,t}, \boldsymbol{h}_{-,t})\,\widetilde{\boldsymbol{V}}^\top, \tag{9a}$$

where the diagonal entries evolve as (suppressing t-dependence henceforth when clear from context):

$$\dot{w} = \beta\alpha_H + bh, \quad \dot{\alpha}_H = \beta w, \quad \dot{h} = bw, \quad \dot{\boldsymbol{w}}_\pm = \boldsymbol{b}_\pm \odot \boldsymbol{h}_\pm, \quad \dot{\boldsymbol{h}}_\pm = \boldsymbol{b}_\pm \odot \boldsymbol{w}_\pm\,, \tag{10}$$

for the mapping $[\beta, b, \boldsymbol{b}_+, \boldsymbol{b}_-] = \mathcal{F}\,([\alpha_H w\,,\, wh\,,\, \boldsymbol{w}_+ \odot \boldsymbol{h}_+\,,\, \boldsymbol{w}_- \odot \boldsymbol{h}_-])$ as specified in Theorem 1.

5.1 Logit dynamics

As with the RP, we now focus on the logit matrix $\boldsymbol{Z}_t = \boldsymbol{W}_t\boldsymbol{H}_t$, which under the parameterization in Eq. (9) is fully described by the spectral coordinates $\alpha = \alpha_H w$, $a = wh$, and $\boldsymbol{a}_\pm = \boldsymbol{w}_\pm \odot \boldsymbol{h}_\pm$. We focus on their evolution. To simplify the analysis, we adopt the standard balancedness condition at initialization, $\boldsymbol{W}_0^\top\boldsymbol{W}_0 = \boldsymbol{H}_0\boldsymbol{H}_0^\top$ (Cohen & Razin, 2024; Saxe et al., 2014; 2019; Gidel et al., 2019; Garrod et al., 2025), which is preserved by the dynamics. See App. B.2 for a discussion on departure from balanced initialization. As in the RP characterization, we also track the maj-min mode in the (γ, δ) parameterization of Eq. (6).

Corollary 1. *Suppose the spectral initialization of Theorem 5 additionally satisfies the balancedness conditions $w_0 = (\alpha_{H,0}^2 + h_0^2)^{1/2}$ and $\boldsymbol{w}_{\pm,0} = \boldsymbol{h}_{\pm,0}$. These conditions are preserved by the dynamics, and the logit matrix takes the form $\boldsymbol{Z}_t = \boldsymbol{U}\,\mathrm{diag}_{+1}\,(\alpha_t, a_t, \boldsymbol{a}_{+,t}, \boldsymbol{a}_{-,t})\,\widetilde{\boldsymbol{V}}^\top$ for all $t \geq 0$. With (γ_t, δ_t) defined as in Eq. (6) (a one-to-one transformation of (α_t, a_t)), the logit parameters evolve as*

$$\dot{\gamma} = \frac{(2R\gamma^2 + \delta^2)\,X_+ + \gamma\delta\,X_-}{\sqrt{2(R\gamma^2 + \delta^2)}}, \quad \dot{\delta} = \frac{(R\gamma^2 + 2\delta^2)\,X_- + R\gamma\delta\,X_+}{\sqrt{2(R\gamma^2 + \delta^2)}}, \quad \dot{\boldsymbol{a}}_\pm = 2\,\boldsymbol{a}_\pm \odot \boldsymbol{b}_\pm. \tag{11}$$

We henceforth focus on the ODE in (11) with (strictly) positive initialization for all parameters. Because $\boldsymbol{b}_\pm \geq \boldsymbol{0}$ and $X_\pm \geq 0$, this directly yields that all the parameters are increasing in time.

5.2 EMERGENCE

In analogy with the RP, one expects the modes to appear sequentially, with majority emerging first. This is indeed the case, although the statement is necessarily different. This is because, here, modes are not exactly zero before activation (indeed, the origin is a saddle point!), but rather start at a small scale $O(\varepsilon)$. We therefore compare the times at which the different modes first reach a fixed scale $\eta > \varepsilon$. The analysis is also more delicate than for the RP, since the reduced ODE does not admit an explicit solution. We give some insights below after formally stating the result.

Theorem 6 (Majority first)**.** *Fix $R > 1$. Consider the reduced logit dynamics (11) from positive initialization of size $O(\varepsilon)$, with at least one majority coordinate initialized at order ε. Then, for every sufficiently small fixed threshold $\eta > 0$, and all sufficiently small $\varepsilon \in (0, \eta)$, the majority mode $\boldsymbol{a}_+$ reaches scale η before both the minority mode $\boldsymbol{a}_-$ and the maj-min mode $a_0 := \sqrt{(R\gamma^2 + \delta^2)/2}$.*

See App. F.3 for complete statement and proof. We now give the main intuition. Note that the majority and minority evolution $\dot{\boldsymbol{a}}_\pm = 2\boldsymbol{a}_\pm \odot \boldsymbol{b}_\pm$ is coupled through γ, δ, since $\boldsymbol{b}_\pm$ depends on $\boldsymbol{x}_\pm$, which in turn depends on γ, δ, and γ, δ are coupled in (11). For emergence, however, it suffices to control $\boldsymbol{b}_\pm$ uniformly in the small-coordinate regime. A key observation is that by setting $\boldsymbol{z}_+ := (M\sqrt{R})^{-1}\overline{\boldsymbol{\Phi}}\boldsymbol{a}_+$ and $\boldsymbol{z}_- := M^{-1}\overline{\boldsymbol{\Phi}}\boldsymbol{a}_-$, it holds $|\frac{X_\pm}{K}\overline{\boldsymbol{\Phi}}^\top e^{\boldsymbol{x}_\pm}| \le \frac{1}{\mathbf{1}^\top e^{\boldsymbol{z}_\pm}}|\overline{\boldsymbol{\Phi}}^\top e^{\boldsymbol{z}_\pm}|$; thus we avoid the dependence on γ, δ. Hence, when $\boldsymbol{a}_\pm = O(\varepsilon)$, $\boldsymbol{b}_+ = \sqrt{R}\,\mathbf{1} + O(\varepsilon)$ and $\boldsymbol{b}_- = \mathbf{1} + O(\varepsilon)$, implying that, near the origin, $\dot{\boldsymbol{a}}_+ \approx 2\sqrt{R}\,\boldsymbol{a}_+$ while $\dot{\boldsymbol{a}}_- \approx 2\boldsymbol{a}_-$, so the majority mode grows faster than the minority mode for $R > 1$. It remains to compare with the majority–minority mode $a_0 := \sqrt{(R\gamma^2 + \delta^2)/2}$. From (11), $\dot{a}_0 = R\gamma X_+ + \delta X_-$. Thus, in the $O(\varepsilon)$-regime, where $X_\pm = 1 + O(\varepsilon)$, it holds $\dot{a}_0 \approx R\gamma + \delta$. Unlike the majority/minority equations, this cannot be written in terms of a_0 alone without controlling the relative sizes of γ and δ. A simple Cauchy–Schwarz bound gives $\dot{a}_0 \approx R\gamma + \delta \le 2\sqrt{(R+1)/2} \cdot a_0$. Thus a_0 grows at rate at most $2\sqrt{(R+1)/2}$, strictly below the majority rate $2\sqrt{R}$ for $R > 1$. This upper bound is enough to prove Thm. 6, but it is not tight: the actual growth of a_0 depends on the evolving ratio of γ and δ. Thus, the majority–minority mode is not governed by a closed scalar growth equation even in the linearized regime. This is the ODE analogue of the RP picture: the majority and minority activation thresholds are fixed to $\sqrt{R}$ and 1, while $\lambda_{K,R}$ depends nontrivially on K and R. The GF setting is even subtler. The crude lower bound $\dot{a}_0 \approx R\gamma + \delta \ge \sqrt{2}\, a_0$ does not imply that the majority–minority mode grows faster than the minority mode near the origin; in this sense, it does not provide a GF analogue of the regularization-path fact $\lambda_{K,R} > 1$. Moreover, even a sharper local lower bound would not by itself prove that the majority–minority mode reaches scale η before the minority mode, since the majority mode will have already left the linearized regime.

5.3 CONVERGENCE

To further relate the GF equations to the RP, it is useful to view (11) as a preconditioned gradient flow over the loss $\mathcal{E}(\boldsymbol{a}) := \mathcal{L}(\boldsymbol{U}\mathrm{diag}_{+1}(\boldsymbol{a})\widetilde{\boldsymbol{V}}^\top)$ for $\boldsymbol{a} = (\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-)$. Indeed, one readily checks that $\dot{\boldsymbol{a}} = -\mathcal{P}(\boldsymbol{a})\nabla\mathcal{E}(\boldsymbol{a})$, where $\mathcal{P}(\boldsymbol{a})$ is block diagonal: the (α, a)-block is $\mathcal{P}_{\alpha a}(\alpha, a) = \frac{1}{\sqrt{\alpha^2+a^2}}\begin{pmatrix} a^2+2\alpha^2 & \alpha a \\ \alpha a & 2a^2+\alpha^2 \end{pmatrix}$ and the $\boldsymbol{a}_+, \boldsymbol{a}_-$ blocks are $2\,\mathrm{diag}\,(\boldsymbol{a}_+)$ and $2\,\mathrm{diag}\,(\boldsymbol{a}_-)$. Since we work in the positive regime, the regularizer $\|\boldsymbol{a}\|_{+1} := \sqrt{a^2 + \alpha^2} + \|\boldsymbol{a}_+\|_1 + \|\boldsymbol{a}_-\|_1$ in (7) is differentiable and a direct calculation verifies $\nabla\|\boldsymbol{a}\|_{+1} = 2\mathcal{P}(\boldsymbol{a})^{-1}\boldsymbol{a}$. Thus the RP optimality condition $\nabla\mathcal{E}(\boldsymbol{a}_\lambda) + \lambda\nabla\|\boldsymbol{a}_\lambda\|_{+1} = 0$ is equivalent to $-\mathcal{P}(\boldsymbol{a}_\lambda)\nabla\mathcal{E}(\boldsymbol{a}_\lambda) = 2\lambda\boldsymbol{a}_\lambda$. We conclude that RP solutions are exactly *radial points* of the GF.

Proposition 2. *A positive point $\boldsymbol{a} > 0$ is radial for the ODE in (11), that is, satisfies $\dot{\boldsymbol{a}} = \mu\boldsymbol{a}$ for some $\mu \in (0, 2)$, if and only if it is the unique RP solution of (7) with $\lambda = \mu/2$.*

For (11), radiality requires $\boldsymbol{b}_\pm = (\mu/2)\mathbf{1} = \lambda\mathbf{1}$ and since positive radial points coincide with RP solutions, the symmetry argument of Thm. 2 further implies that radial points satisfy $\boldsymbol{a}_\pm \propto \mathbf{1}$. Although this need not generically hold for GF (unlike for the RP), we can prove that the symmetric subspace is invariant (Fig. 10) and that deviations from uniformity, measured in KL divergence, decrease monotonically along the flow.

Theorem 7 (Maj. and min. tend to uniformity)**.** *If the GF (11) is initialized so that $\boldsymbol{a}_\pm \propto \mathbf{1}$ then this remains true. More generally, for positive initialization, $\dot{D}_{KL}(\overline{\mathbf{1}}\|\overline{\boldsymbol{a}}_\pm) \le 0$, with equality iff $\boldsymbol{a}_\pm \propto \mathbf{1}$. Moreover, $D_{\mathrm{KL}}(\overline{\mathbf{1}}\|\overline{\boldsymbol{a}}_{\pm,t}) \to 0$ as $t \to \infty$. Thus, $\overline{\boldsymbol{a}}_{\pm,t} \to \overline{\mathbf{1}}$.*

We prove Thm. 7 in App. F.2. Since $\boldsymbol{a}_\pm$ approach uniformity in direction from general positive initialization, we focus onwards on initialization $\boldsymbol{a}_{\pm,0} = a_{\pm,0}\mathbf{1}$, for which $\boldsymbol{a}_{\pm,t} = a_{\pm,t}\mathbf{1}$ for all t and, thus, the dynamics in Eq. (11) reduce to four scalar variables $(\gamma, \delta, a_+, a_-)$, just like the RP.

Together with Prop. 2, these reinforce the conjecture that GF approaches the same limiting direction as the vanishing RP (Thm. 3). In the balanced case $R = 1$, Garrod et al. (2025) prove this under their spectral initialization, but the diagonal-plus-rank-one parameterization needed here makes the argument significantly more delicate. As evidence for this added difficulty, App. F.4 shows that even when $R = 1$, the KL divergence to the ETF direction, which is a Lyapunov function under the exact diagonalization of Garrod et al. (2025), need not be a Lyapunov function in our coordinates.

By the transformation in (6), the conjectured convergence of the four-variable dynamics to the RP limit is equivalent to

$$\tfrac{a_{+,t}}{\sqrt{R}} \sim \gamma_t \sim \delta_t \sim a_{-,t}\,, \qquad \text{as } t \to \infty, \tag{12}$$

where $\sim$ denotes that the ratio tends to one. Indeed, since $a_t\sqrt{2(R+1)} = R\gamma_t + \delta_t$ and $\alpha_t\sqrt{2(R+1)} = \sqrt{R}(\gamma_t - \delta_t)$, Eq. (12) implies $\overline{\boldsymbol{a}}_t \to \overline{\boldsymbol{a}}_\infty$, where $\boldsymbol{a}_t \coloneqq (\alpha_t, a_t, a_{+,t}, a_{-,t})$ and $\boldsymbol{a}_\infty$ is precisely the RP limiting direction in Thm. 3. The following result proves the desired relation between the majority and minority modes, together with their growth rate.

Theorem 8 (Majority and minority asymptotics)**.** *Consider the GF (11) with $\gamma_0, \delta_0 > 0$ and $\boldsymbol{a}_{\pm,0} = a_{\pm,0}\mathbf{1}$ for $a_{\pm,0} > 0$. Then, $\lim_{t\to\infty} a_{+,t}/\sqrt{R}\,a_{-,t} = 1$. Moreover, $a_{\pm,t} = \Theta(\log t)$.*

Proving the theorem (see App. G.3) boils down to substantial refinement of the bounds on the coefficients $\boldsymbol{b}_{\pm,t}$ discussed in Sec. 5.2 as part of the emergence study. Using only positivity of γ_t, δ_t, we first establish logarithmic growth of $a_\pm$, but the resulting bounds only guarantee the limit of $a_{+,t}/(\sqrt{R}\,a_{-,t})$ between $(M-1)/M$ and $M/(M-1)$. To close this gap, it is necessary to study the relative growth of δ_t and $a_{-,t}$, which we do using $\log a_{-,t}$ as the time variable (similarly for γ_t and $a_{+,t}/\sqrt{R}$). As t grows, whenever δ_t falls below $(1-\epsilon)a_{-,t}$ for some ϵ, its relative growth is faster, bringing again the ratio above $1-\epsilon$. Consequently, for every $\epsilon > 0$, asymptotically

$$\delta_t \geq (1-\epsilon)a_{-,t}, \qquad \gamma_t \geq (1-\epsilon)\tfrac{a_{+,t}}{\sqrt{R}}. \tag{13}$$

Leveraging these lower bounds and plugging back into the bounds on $\boldsymbol{b}_{\pm,t}$ tightens the growth estimates and gives the desired by letting $\epsilon \to 0$.

6 Discussion

Our setting is deliberately minimal: Step-imbalanced and orthogonal data exhibit non-trivial mode structure, and the two-layer linear architecture allows this structure to be learned sequentially. Given the lack of CE trajectory-level results we deem this a reasonable first step. The regularization and gradient-flow paths reveal both quantitative and qualitative differences from the MSE dynamics most commonly studied in the literature. Methodologically, the main obstacle in extending the MSE theory to CE is the loss of closed-form decoupled dynamics. By extending the softmax diagonalization perspective of (Garrod et al., 2025) to a non-symmetric setting, we further showcase that this obstacle can be overcome. In the imbalanced UFM, closed-form expressions can be replaced by low-dimensional scalar equations for the regularization path and low-dimensional ODEs for gradient flow, both of which remain amenable to analysis. A natural next step is to identify richer data settings and mode structures for which similar suitable spectral parameterizations, without the possibly strict requirement of exact diagonalization of prior works, makes CE dynamics tractable.

Our GF analysis uses a spectral initialization that keeps the trajectory in the subspace identified by the regularization path. This is an exact CE trajectory, and empirically resembles the dynamics from small random initialization (Fig. 9), but proving the latter connection remains open (as is open even in the balanced case). Under initialization uniform within each group, we prove that the majority-to-minority mode ratio converges to its RP limit. Our asymptotic lower bounds (Eq. (13)) further show that the majority–minority coordinates attain at least the relative scale predicted by the RP. Matching upper bounds would thus settle convergence of the full logit direction to the RP limit. More ambitious future directions include extending the analysis to nonlinear and deeper models. See App. B.2 for a discussion of specific technical restrictions and possible ways to relax them.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the text of the paper, for assistance with specific steps within proofs of our results, and for writing code for experiments. We have not used generative AI tools for drafting the paper, for proposing or refining hypotheses, interpreting results, developing conceptual frameworks, formulating mathematical statements, or designing or providing feedback on research methodology or experiments and other cases to assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, generate synthetic data sets, or qualitative data analysis are not applicable to this work. Additionally, we used generative AI tools for checking the consistency of stated claims against the corresponding proofs and for numerical sanity checks of analytical results. We have reviewed all AI-assisted work. All AI-assisted proof steps were checked line by line by at least one author, all AI-assisted code was verified and tested for correctness by at least one author. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI. The tools used were general-purpose chat assistants (Claude Opus 5 and ChatGPT 5.6).

REFERENCES

Alnur Ali, Edgar Dobriban, and Ryan Tibshirani. The implicit regularization of stochastic gradient flow for least squares. In *International conference on machine learning*, pp. 233–244. PMLR, 2020.

George Andriopoulos, Zixuan Dong, Li Guo, Zifan Zhao, and Keith Ross. The prevalence of neural collapse in neural multivariate regression. *arXiv preprint arXiv:2409.04180*, 2024.

Sanjeev Arora, N Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In *35th International Conference on Machine Learning*, 2018a.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018b.

Francis Bach. Closed-form dynamics beyond quadratics, mar 2026. URL https://francisbach.com/closed-form-dynamics/.

Kiril Bangachev, Guy Bresler, Iliyas Noman, and Yury Polyanskiy. Global minimizers of sigmoid contrastive loss. *arXiv preprint arXiv:2509.18552*, 2025.

Ioannis Bantzis, James B Simon, and Arthur Jacot. Saddle-to-saddle dynamics in deep relu networks: Low-rank bias in the first saddle escape. *arXiv preprint arXiv:2505.21722*, 2025.

Amir Beck. *First-Order Methods in Optimization*. Society for Industrial and Applied Mathematics, Philadelphia, PA, 2017. doi: 10.1137/1.9781611974997.

Tina Behnia, Ganesh Ramachandra Kini, Vala Vakilian, and Christos Thrampoulidis. On the implicit geometry of cross-entropy parameterizations for label-imbalanced data. In *International Conference on Artificial Intelligence and Statistics*, pp. 10815–10838. PMLR, 2023.

Raphaël Berthier. Diagonal linear networks and the lasso regularization path. *arXiv preprint arXiv:2509.18766*, 2025.

Rajendra Bhatia. *Matrix analysis*, volume 169. Springer Science & Business Media, 2013.

Jian-Feng Cai, Emmanuel J Candès, and Zuowei Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

Nadav Cohen and Noam Razin. Lecture notes on linear neural networks: A tale of optimization and generalization in deep learning. *arXiv preprint arXiv:2408.13767*, 2024.

Hien Dang, Tan Nguyen, Tho Tran, Hung Tran, and Nhat Ho. Neural collapse in deep linear network: From balanced to imbalanced data. *arXiv preprint arXiv:2301.00437*, 2023.

Hien Dang, Tho Tran, Tan Nguyen, and Nhat Ho. Neural collapse for cross-entropy class-imbalanced learning with unconstrained relu feature model. *arXiv preprint arXiv:2401.02058*, 2024.

Cong Fang, Hangfeng He, Qi Long, and Weijie J Su. Exploring deep neural networks via layer-peeled model: Minority collapse in imbalanced training. *Proceedings of the National Academy of Sciences*, 118(43), 2021.

Quinn LeBlanc Fisher, Haoming Meng, and Vardan Papyan. Pushing boundaries: Mixup's influence on neural collapse. In *The Twelfth International Conference on Learning Representations*, 2024.

Connall Garrod and Jonathan P Keating. The persistence of neural collapse despite low-rank bias: An analytic perspective through unconstrained features. *arXiv preprint arXiv:2410.23169*, 2024.

Connall Garrod, Jonathan P Keating, and Christos Thrampoulidis. Diagonalizing the softmax: Hadamard initialization for tractable cross-entropy dynamics. *arXiv preprint arXiv:2512.04006*, 2025.

Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Suriya Gunasekar, Blake E. Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–10, 2017. URL https://api.semanticscholar.org/CorpusID:3909231.

Li Guo, George Andriopoulos, Zifan Zhao, Shuyang Ling, Zixuan Dong, and Keith Ross. Cross entropy versus label smoothing: A neural collapse perspective. *arXiv preprint arXiv:2402.03979*, 2024.

XY Han, Vardan Papyan, and David L Donoho. Neural collapse under mse loss: Proximity to and dynamics on the central path. *arXiv preprint arXiv:2106.02073*, 2021.

Wanli Hong and Shuyang Ling. Neural collapse for unconstrained feature model under cross-entropy loss with imbalanced data. *arXiv preprint arXiv:2309.09725*, 2023.

Wanli Hong and Shuyang Ling. Beyond unconstrained features: Neural collapse for shallow neural networks with general data. *arXiv preprint arXiv:2409.01832*, 2024.

Arthur Jacot, François Ged, Berfin Şimşek, Clément Hongler, and Franck Gabriel. Saddle-to-saddle dynamics in deep linear networks: Small initialization training, symmetry, and sparsity. *arXiv preprint arXiv:2106.15933*, 2021.

Arthur Jacot, Peter Súkeník, Zihan Wang, and Marco Mondelli. Wide neural networks trained with weight decay provably exhibit neural collapse. *arXiv preprint arXiv:2410.04887*, 2024.

Wenlong Ji, Yiping Lu, Yiliang Zhang, Zhun Deng, and Weijie J Su. An unconstrained layer-peeled perspective on neural collapse. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WZ3yjh8coDg.

Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032*, 2018.

Ziwei Ji, Miroslav Dudík, Robert E Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Conference on Learning Theory*, pp. 2109–2136. PMLR, 2020.

Jiachen Jiang, Jinxin Zhou, Peng Wang, Qing Qu, Dustin Mixon, Chong You, and Zhihui Zhu. Generalized neural collapse for a large number of classes. *arXiv preprint arXiv:2310.05351*, 2023.

Ganesh Ramachandra Kini, Orestis Paraskevas, Samet Oymak, and Christos Thrampoulidis. Label-imbalanced and group-sensitive classification under overparameterization. *Advances in Neural Information Processing Systems*, 34:18970–18983, 2021.

Ganesh Ramachandra Kini, Vala Vakilian, Tina Behnia, Jaidev Gill, and Christos Thrampoulidis. Symmetric neural-collapse representations with supervised contrastive loss: The impact of relu and batching. In *The Twelfth International Conference on Learning Representations*, 2024.

Pengyu Li, Xiao Li, Yutong Wang, and Qing Qu. Neural collapse in multi-label learning with pick-all-label loss. *arXiv preprint arXiv:2310.15903*, 2023.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. *arXiv preprint arXiv:2012.09839*, 2020.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. In *International Conference on Learning Representations*, 2020.

Dustin G Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *arXiv preprint arXiv:2011.11619*, 2020.

Tiberiu Musat. The geometry of grokking: Norm minimization on the zero-loss manifold. *arXiv preprint arXiv:2511.01938*, 2025.

Vardan Papyan, XY Han, and David L Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Mee Young Park and Trevor Hastie. L 1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 69(4):659–677, 2007.

Saharon Rosset, Ji Zhu, and Trevor J. Hastie. Margin maximizing loss functions. In *NIPS*, 2003.

Saharon Rosset, Ji Zhu, and Trevor Hastie. Boosting as a regularized path to a maximum margin classifier. *Journal of Machine Learning Research*, 5(Aug):941–973, 2004.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1312.6120.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23): 11537–11546, 2019.

Arun Suggala, Adarsh Prasad, and Pradeep K Ravikumar. Connecting optimization and regularization paths. *Advances in Neural Information Processing Systems*, 31, 2018.

Peter Súkeník, Marco Mondelli, and Christoph H Lampert. Deep neural collapse is provably optimal for the deep unconstrained features model. In *Advances in Neural Information Processing Systems*, volume 36, pp. 52991–53024. Curran Associates, Inc., 2023.

Peter Súkeník, Marco Mondelli, and Christoph Lampert. Neural collapse versus low-rank bias: Is deep neural collapse really optimal? *arXiv preprint arXiv:2405.14468*, 2024.

Peter Súkeník, Christoph H Lampert, and Marco Mondelli. Neural collapse is globally optimal in deep regularized resnets and transformers. *arXiv preprint arXiv:2505.15239*, 2025.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism, 2023.

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint 2308.16898*, 2024.

Christos Thrampoulidis, Ganesh R Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *arXiv preprint arXiv:2208.05512*, 2022.

Zhenfeng Tu, Santiago Aranguri, and Arthur Jacot. Mixed dynamics in linear networks: Unifying the lazy and active regimes. In *Advances in Neural Information Processing Systems*, volume 37, pp. 106059–106104. Curran Associates, Inc., 2024.

Gal Vardi, Ohad Shamir, and Nati Srebro. On margin maximization in linear and relu networks. *Advances in Neural Information Processing Systems*, 35:37024–37036, 2022.

Bhavya Vasudeva, Puneesh Deora, Yize Zhao, Vatsal Sharan, and Christos Thrampoulidis. How muon's spectral design benefits generalization: A study on imbalanced data. *arXiv preprint arXiv:2510.22980*, 2025.

Diyuan Wu and Marco Mondelli. Neural collapse beyond the unconstrained features model: Landscape, dynamics, and generalization in the mean-field regime. *arXiv preprint arXiv:2501.19104*, 2025.

Robert Wu and Vardan Papyan. Linguistic collapse: Neural collapse in (large) language models. *arXiv preprint arXiv:2405.17767*, 2024.

Yedi Zhang, Aaditya K Singh, Peter E Latham, and Andrew Saxe. Training dynamics of in-context learning in linear attention. *arXiv preprint arXiv:2501.16265*, 2025.

Yize Zhao and Christos Thrampoulidis. Geometry of semantics in next-token prediction: How optimization implicitly organizes linguistic representations. *arXiv preprint 2505.08348*, 2025.

Yize Zhao, Tina Behnia, Vala Vakilian, and Christos Thrampoulidis. Implicit geometry of next-token prediction: From language sparsity patterns to model representations. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=qyilOnIRHI.

Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. *arXiv preprint arXiv:2210.02192*, 2022.

Zhihui Zhu, Tianyu Ding, Jinxin Zhou, Xiao Li, Chong You, Jeremias Sulam, and Qing Qu. A geometric analysis of neural collapse with unconstrained features. *Advances in Neural Information Processing Systems*, 34, 2021.

CONTENTS

1 Introduction **1**

2 Problem setup and background **3**

3 Technical tool: Approximate softmax diagonalization **4**

4 Regularization path **5**

5 Gradient-flow Path **7**

5.1 Logit dynamics 7

5.2 Emergence 8

5.3 Convergence 8

6 Discussion **9**

A Background on MSE loss **15**

A.1 Gradient-flow 15

A.2 Regularization-path 15

A.3 On their relation 16

B Related Work and Limitations **17**

B.1 Detailed Related Work 17

B.2 Limitations 19

C Notation and basic facts **20**

D Proof of Theorem 1 **21**

D.1 Permutation structure of the softmax residual 21

D.2 Spectral factorization and rank-one correction 23

D.3 Adding the rank-one term 25

D.4 Diagonal specialization 25

D.5 Simplified formulas 26

E Proofs for Regularization-path **27**

E.1 Proof of Theorem 2 and basic properties of the solution 27

E.1.1 Reduction to four variables 30

E.2 Proof of proposition 1: Analytic solution 31

E.3 Analytic solution of Proposition 1 34

E.4 Additional characteristics of the RP solution 36

E.5 Behavior of the threshold $\lambda_{K,R}$ 40

E.6 Asymptotic behavior of RP solution and Proof of Theorem 3 41

E.7 Alternative proof of regularization-path limit 43

F Proofs for gradient-flow path **44**

F.1 Proof of Theorem 5 and Corollary 1 44

F.2 Proof of Theorem 7 45

F.3 Proof of Theorem 6 48

F.3.1 Auxiliary lemmas 49

F.3.2 Proof of Proposition 13 50

F.3.3 Proof of Proposition 14 51

F.4 Lyapunov function counterexample 52

G Proof of Theorem 8 **53**

G.1 Auxiliary results 53

G.2 Logarithmic growth 55

G.3 Asymptotics 55

G.3.1 Simplified equations 55

G.3.2 Growth bounds for majority and minority logits 56

G.3.3 Majority and minority logits converge in direction to their SEL limit 57

H Numerical results **59**

H.1 Regularization-path experiments 59

H.1.1 Regularization-path comparison for different K and R 59

H.2 Gradient-flow experiments 62

H.2.1 Setup 62
H.2.2 Results 64
H.2.3 Preservation of block-constant spectral coordinates 65

A Background on MSE loss

Our motivation for this work stems in part from the gap between the understanding of CE and MSE training dynamics. The latter is well studied for GF (Saxe et al., 2014; Gidel et al., 2019). The relation between the ODE and the RP, however, is less explored. We therefore include such a comparison in this section. In the MSE setting, this comparison can be made explicit thanks to the closed-form solutions available for both the GF dynamics and the RP. We highlight that this relation differs from the convex case, where RP and GF relate *pointwise* (Suggala et al., 2018; Ali et al., 2020). In the two-layer nonconvex setup considered here, the RP solution instead connects to the time-averaged GF solution. This finding is a special case of the more general and elegant result of Berthier (2025).

In Sec. A.1, we review the closed-form GF solution. In Sec. A.2, we record the regularization-path solution, which is straightforward but included for completeness. Finally, in Sec. A.3, we relate the two.

A.1 Gradient-flow

We consider the two-layer linear MSE loss with target matrix the SEL matrix $\widehat{\boldsymbol{Z}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$:

$$\mathcal{L}_{\text{mse}}(\boldsymbol{W}\boldsymbol{H}) := \frac{1}{2}\|\widehat{\boldsymbol{Z}} - \boldsymbol{W}\boldsymbol{H}\|_F^2. \tag{14}$$

Under the spectrally aligned initialization of Saxe et al. (2014), $\boldsymbol{W}_0 = \boldsymbol{U}\,\mathrm{diag}\,(\boldsymbol{w}_0)\,\boldsymbol{R}^\top$, $\boldsymbol{H}_0 = \boldsymbol{R}\,\mathrm{diag}\,(\boldsymbol{h}_0)\,\boldsymbol{V}^\top$, the singular subspace is invariant for the GF. Concretely, further assuming balancedness ($\boldsymbol{w}_0 = \boldsymbol{h}_0$), the logit singular values

$$\boldsymbol{W}_t\boldsymbol{H}_t = \boldsymbol{U}\,\mathrm{diag}\,(\boldsymbol{a}_t)\,\boldsymbol{V}^\top, \qquad a_{i,t} = w_{i,t}h_{i,t},\ i \in [K-1]$$

evolve in a decoupled manner, i.e., for all $i \in [K-1]$:

$$\dot{a}_{i,t} = 2a_{i,t}(\sigma_i - a_{i,t})\,, \tag{15}$$

where σ_i are the diagonal entries of $\boldsymbol{\Sigma}$. Thus, for initialization $a_i(0) = e^{-2\delta}$, the modes evolve as (Gidel et al., 2019, Theorem 1)

$$a_{i,t} = \sigma_i \cdot \frac{1}{1 + (\sigma_i e^{2\delta} - 1)e^{-2\sigma_i t}}. \tag{16}$$

Hence each mode grows monotonically from initialization value to the finite SEL singular value σ_i. Moreover, reparameterizing time as $t \mapsto \delta t$ and taking the infinitesimal-initialization limit $\delta \to \infty$ gives (Gidel et al., 2019, Theorem 2)

$$a_{i,\delta t} \to \frac{\sigma_i}{1+\sigma_i} \cdot \mathbf{1}\{t = 1/\sigma_i\} + \sigma_i \cdot \mathbf{1}\{t > 1/\sigma_i\}, \tag{17}$$

Thus, in the rescaled time variable and with vanishing initialization, mode i activates at time $1/\sigma_i$; equivalently, in the original time variable, its activation time is of order δ/σ_i. This explains why, in our step-imbalanced setting, when MSE loss is used, the three modes (majority, majority–minority, and minority) are learned sequentially as t increases (Zhao & Thrampoulidis, 2025).

A.2 Regularization-path

We next study the ridge-regularized counterpart of (14):

$$(\boldsymbol{W}_\lambda, \boldsymbol{H}_\lambda) \in \arg\min_{\boldsymbol{W},\boldsymbol{H}} \mathcal{L}_{\text{mse}}(\boldsymbol{W}\boldsymbol{H}) + \frac{\lambda}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{H}\|_F^2. \tag{18}$$

As in the CE case, the factorized problem admits an equivalent logit-space form (recall $d \geq K$):

$$\boldsymbol{Z}_\lambda \in \arg\min_{\boldsymbol{Z}} \frac{1}{2}\|\widehat{\boldsymbol{Z}} - \boldsymbol{Z}\|_F^2 + \lambda\|\boldsymbol{Z}\|_*. \tag{19}$$

Since $\widehat{\boldsymbol{Z}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, the solution is given by singular-value soft-thresholding (e.g., (Cai et al., 2010, Thm. 2.1)):

$$\boldsymbol{Z}_\lambda = \boldsymbol{U}\,\mathrm{diag}\left((\boldsymbol{\sigma} - \lambda)_+\right)\boldsymbol{V}^\top, \tag{20}$$

where $\boldsymbol{\sigma}$ is the diagonal of $\boldsymbol{\Sigma}$ and the positive part $(\cdot)_+ := \max\{\cdot, 0\}$ is applied entrywise. Equivalently, the logit singular values along the RP are

$$a_{i,\lambda} = (\sigma_i - \lambda)_+, \qquad i \in [K-1]. \tag{21}$$

Thus, mode i is active exactly when $\lambda < \sigma_i$. Unlike CE, the MSE RP has finite mode-wise targets:

$$a_{i,\lambda} \uparrow \sigma_i \qquad \text{as} \qquad \lambda \downarrow 0,$$

for every SEL mode $i \in [K-1]$.

For comparison, the one-layer MSE ridge path is

$$\arg\min_{\boldsymbol{Z}} \frac{1}{2}\|\widehat{\boldsymbol{Z}} - \boldsymbol{Z}\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{Z}\|_F^2 = \frac{1}{1+\lambda}\widehat{\boldsymbol{Z}}. \tag{22}$$

Thus all SEL modes are present for every finite λ in the one-layer MSE model, and there is no thresholded mode emergence.

A.3 On their relation

Since we have closed-form solutions for both GF (in(16)) and RP (in(21)) we can explicitly relate them. The proposition below is a special case of a more general finding by Berthier (2025).

Proposition 3 (RP to GF for MSE loss)**.** *Let $a_{i,\lambda}$ denote the logit singular values of the RP in Eq. (21). Further, let $a_{i,t}^{(\delta)}$ denote the logit singular values of the GF solution to the MSE loss Eq. (16) with balanced factor initialization $w_i(0) = h_i(0) = e^{-\delta}$, so that $a_i(0) = e^{-2\delta}$. Then, for all $i \in [K-1]$ and all $\lambda > 0$,*

$$a_{i,\lambda} = \lim_{\delta\to\infty} \hat{a}_{i,\frac{\delta}{\lambda}}^{(\delta)}, \qquad \textit{where} \qquad \hat{a}_{i,t}^{(\delta)} := \frac{1}{t}\int_0^t a_{i,\tau}^{(\delta)}\, d\tau. \tag{23}$$

Proof. We have for any A, B that $\int_0^t \frac{1}{1+Ae^{-B\tau}}\mathrm{d}\tau = t + \frac{1}{B}\log\left(\frac{1+Ae^{-Bt}}{1+A}\right)$. Thus, from (16), for all t and all δ:

$$\hat{a}_{i,t}^{(\delta)} = \sigma_i + \frac{1}{2t}\log\left(\sigma_i^{-1}e^{-2\delta} + (1 - \sigma_i^{-1}e^{-2\delta})e^{-2\sigma_i t}\right).$$

Set $t \leftarrow \frac{\delta}{\lambda}$, this gives

$$\hat{a}_{i,\frac{\delta}{\lambda}}^{(\delta)} = \sigma_i + \frac{\lambda}{2\delta}\log\left(\sigma_i^{-1}e^{-2\delta} + \left(1 - \sigma_i^{-1}e^{-2\delta}\right)e^{-2\sigma_i\delta/\lambda}\right). \tag{24}$$

Taking the limits and invoking the log-sum-exp asymtptotic, it can be shown that

$$\frac{1}{\delta}\log\left(\sigma_i^{-1}e^{-2\delta} + \left(1 - \sigma_i^{-1}e^{-2\delta}\right)e^{-2\sigma_i\delta/\lambda}\right) \overset{\delta\to\infty}{\longrightarrow} \max\{-2\,,\, -\frac{2\sigma_i}{\lambda}\} = -2\,\min\{1\,,\,\frac{\sigma_i}{\lambda}\}.$$

Put together, completes the proof as follows:

$$\hat{a}_{i,\frac{\delta}{\lambda}}^{(\delta)} \overset{\delta\to\infty}{\longrightarrow} \sigma_i - \min\{\lambda, \sigma_i\} = (\sigma_i - \lambda)_+ = a_{i,\lambda}\,.$$

□

This proposition makes explicit the connection between training time and regularization strength in the MSE setting. As expected, time plays the role of an inverse regularization parameter. More precisely, under infinitesimal balanced initialization, the relevant correspondence is $t \leftrightarrow \delta/\lambda$, so the time scale is also modulated by the logarithm of inverse the initialization scale (i.e., $e^{-2\delta}$).

More interestingly, the relation between GF and the RP is not pointwise in time. Instead, in the infinitesimal-initialization limit, the RP solution is obtained from a time average of the GF trajectory.

The established identity can be viewed as a concrete special case of the elegant and more general phenomenon recently proved by Berthier (2025) for diagonal linear networks trained on convex quadratic objectives, of which the standard MSE loss studied here is a special case. In that more general setting, the coordinates are coupled, and closed-form expressions for the trajectory are not available, making their proof substantially more involved. Our proposition therefore connects Berthier's result to the classical literature on closed-form dynamics of two-layer linear networks and provides a simpler, more direct proof, albeit in a (much more) special case.

B Related Work and Limitations

B.1 Detailed Related Work

Our work connects three lines of literature: neural collapse and the UFM, mode emergence in two-layer models, and regularization paths for cross-entropy objectives. While we point out specific connections throughout the main text, we collect and discuss them in more detail here.

Neural collapse, UFM, and data-dependent geometry. Our primary motivation comes from the neural-collapse geometry phenomenon, first empirically identified and formalized by Papyan et al. (2020). This now rich literature supports the idea that sufficiently large neural networks, trained for sufficiently long, learn training-data representations that converge to specific geometries reflecting the information encoded in the label matrix. This picture is supported both theoretically (Súkeník et al., 2025; Súkeník et al., 2023; Jacot et al., 2024; Hong & Ling, 2024; Wu & Mondelli, 2025) and empirically (Papyan et al., 2020; Fang et al., 2021; Thrampoulidis et al., 2022; Zhao et al., 2024; Guo et al., 2024; Zhou et al., 2022; Andriopoulos et al., 2024; Fisher et al., 2024; Kini et al., 2024; Dang et al., 2024; 2023). A central theoretical route for identifying these geometries is the Unconstrained Features Model (UFM), a two-layer linear model with standard-basis inputs and one output per class. In the balanced one-hot setting, the UFM correctly predicts that learned representations form a simplex equiangular tight frame (ETF) (Mixon et al., 2020; Han et al., 2021; Zhu et al., 2021). Its predictions extend to imbalanced classes (Fang et al., 2021; Thrampoulidis et al., 2022; Behnia et al., 2023; Dang et al., 2024), multilabel data (Li et al., 2023), language data (Zhao et al., 2024; Wu & Papyan, 2024), and other losses (Bangachev et al., 2025; Jiang et al., 2023; Fisher et al., 2024).

Deep networks exhibiting neural-collapse geometries are typically trained with cross-entropy (CE), so the faithful UFM reduction is also CE training. This introduces features that are absent from the MSE theory. The softmax map adds a nonlinear coupling across classes on top of the nonconvexity coming from the bilinear factorization $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{H}$, and unregularized CE drives the optima, and hence the training trajectory, to infinity. This is different from MSE, where the loss structure often enables explicit decompositions and closed-form or nearly closed-form dynamics (Han et al., 2021). In the symmetric balanced one-hot setting, the limiting neural-collapse geometry is the same for MSE-UFM and CE-UFM (Mixon et al., 2020; Han et al., 2021). Moreover, in that setting, regularization does not change the qualitative conclusion (Zhu et al., 2021; Ji et al., 2022; Garrod et al., 2025); subtleties start to appear, for example, in deeper UFM variants (Garrod & Keating, 2024; Súkeník et al., 2024). Once one departs from the balanced symmetric setting, however, the geometry can become sensitive to whether regularization is present (Fang et al., 2021; Papyan et al., 2020; Dang et al., 2023; 2024). In this paper, we study both the entire ridge-regularization path and the unregularized GF trajectory.

Most of the neural-collapse/UFM literature focuses on convergence: either identifying the global minimizer of a regularized UFM objective (Zhu et al., 2021; Dang et al., 2024; Hong & Ling, 2023; Dang et al., 2023), or characterizing the limiting direction of GD/GF in the unregularized case (Ji et al., 2022; Thrampoulidis et al., 2022; Garrod et al., 2025). Our focus is different: in addition to convergence, we study *mode emergence* along the training path. Our motivation on this is Zhao & Thrampoulidis (2025), who identify the step-imbalanced one-hot setting studied here as the simplest

UFM setup with nontrivial mode structure. In this setting, the representation geometry is organized around three semantic axes: a majority mode, a majority–minority mode, and a minority mode. This interpretation is visible directly from the columns of the basis $\boldsymbol{U}$: there are $M-1$ columns supported on the majority classes, one column separating majorities from minorities with opposite signs, and $M-1$ columns supported on the minority classes. Thus, the step-imbalanced UFM provides a minimal setting in which one can ask not only where CE training converges, but also which semantic directions are learned first.

Mode emergence and the MSE framework. The mode interpretation above connects the neural-collapse/UFM literature to the long line of work on mode emergence in two-layer linear networks. This line was pioneered by Saxe et al. (2014) in terms of technical tools and by Saxe et al. (2019) in terms of semantic interpretation. In that framework, the singular modes of the target matrix represent semantic directions, and training under MSE learns these modes sequentially. The resulting picture is mathematically clean and has proved robust across many extensions (e.g., Gidel et al., 2019; Tu et al., 2024; Bantzis et al., 2025; Bach, 2026); see also the recent discussion in Bach (2026). However, this framework is largely restricted to MSE. One of the drivers of the present work is to push this mode-emergence program toward the CE loss.

The difficulty is that the MSE framework relies on two special properties. First, under spectral initialization, the MSE residual $\boldsymbol{Y}-\boldsymbol{W}\boldsymbol{H}$, the logits $\boldsymbol{W}\boldsymbol{H}$, and the label matrix $\boldsymbol{Y}$ are simultaneously diagonalizable in any singular basis of $\boldsymbol{Y}$. This fixes the parameter subspaces along the trajectory and reduces training to the evolution of singular values. Second, the resulting singular-value dynamics often admit closed-form or nearly closed-form descriptions. For CE, even the first property fails in general because the softmax nonlinearity couples the logits across classes.

Towards a CE analogy. A recent step toward overcoming this obstacle was taken very recently by Garrod et al. (2025). Their work inspired ours in its technical core. In the balanced one-hot setting, the authors of (Garrod et al., 2025) showed that if one chooses a particular Sylvester–Hadamard singular basis of the label matrix, then the CE residual $\boldsymbol{Y}-\mathbb{S}\left(\boldsymbol{W}\boldsymbol{H}\right)$ remains diagonalizable together with $\boldsymbol{W}\boldsymbol{H}$ and $\boldsymbol{Y}$. This preserves the parameter subspaces under spectral initialization and reduces GF to singular-value dynamics. Although these dynamics are not closed form as in MSE, they are still tractable. However, their setting is balanced one-hot data. From the perspective of semantic mode emergence, this case is limited because there is only one nontrivial mode. More importantly, even the smallest departure from balance, namely the step-imbalanced setting studied here, breaks pure Hadamard diagonalization: the imbalanced target matrix is no longer diagonalized by the Sylvester–Hadamard basis in the same way.

Our contribution on this front is to show that the program remains viable beyond the balanced setting. We identify a decomposition of the step-imbalanced label matrix involving Sylvester–Hadamard components, which lets us leverage the same interaction between Hadamard structure, entrywise products, and softmax equivariance that underlies the proof of Garrod et al. (2025). However, exact diagonalization is no longer preserved. Instead, Theorem 1 identifies the correct replacement: a diagonal-plus-rank-one decomposition of the CE residual. We then show that this structure is still sufficient to reduce the full matrix ODE to a vector ODE. Thus, while pure diagonalization breaks under imbalance, the analysis can continue after enlarging the spectral structure in the right way.

Regularization paths and sparse mode activation. A further distinction of our work is that the same spectral structure also solves the regularization path. This is not an initialization assumption: for every regularization strength, we prove that the global minimizer of the logit-space nuclear-norm problem lies in the diagonal-plus-rank-one subspace identified by the softmax analysis. A priori, this is not obvious. The regularization-path problem is a convex logit-space problem, and it is not clear in advance why its solution should admit a tight reduction to an ℓ_1-style CE minimization, let alone to a four-dimensional problem with analytic phase formulas.

There is related work on regularization paths for logistic regression and other exponentially tailed losses (Rosset et al., 2003; 2004). In particular, Park & Hastie (2007) develop an algorithm for numerically producing the entire trajectory of solutions to ℓ_1-regularized logistic regression for general input data. There, the ℓ_1-style penalty induces sparsity and sequential entry of coordinates. In our specific data setting, the coordinates are singular modes of the imbalanced label geometry, and we can analyze the path explicitly. Thus, the sparsity induced by the nuclear-norm/logit formulation becomes a mode-emergence phenomenon: majority first, then majority–minority, and finally minority.

Comparison with finite-regularization CE-UFM results. The closest finite-regularization CE-UFM results we are aware of are Dang et al. (2024); Hong & Ling (2023). Dang et al. (2024) characterize global minimizers of a CE objective under an unconstrained ReLU-feature model, i.e., a UFM variant with entrywise nonnegative features. Their Theorem 4.1 gives a class-wise thresholding result: when the effective regularization is large relative to the size of a class, the corresponding class mean collapses to zero. Thus their path-like structure is a class-mean collapse/activation phenomenon in a nonnegative-feature model. Our result concerns a different object: we study the standard unconstrained UFM through the logit-space nuclear-norm formulation and characterize the regularization path spectrally, mode by mode. Modes are not classes. In our path, the majority singular mode activates first, then a majority–minority mode, and finally the minority mode. By contrast, the thresholding in (Dang et al., 2024) is class-wise: class means become nonzero sequentially, and when active they satisfy the angular structure specified by their theorem. The model and mechanism are also different: entrywise nonnegativity changes the feasible geometry and enables a lower-bound/equality-condition proof strategy, whereas our analysis works in the standard unconstrained UFM through a logit-space spectral reduction. As detailed in the paper's main body, the result closest to our RP analysis is by Hong & Ling (2023). Finally, (Thrampoulidis et al., 2022) introduced SELI as the limiting imbalanced neural-collapse geometry and showed that the UFM converges to this geometry in the vanishing-regularization limit. This was the first work to explicitly characterize the imbalanced limiting geometry through explicit analysis of the UFM and connect it to minority collapse of Fang et al. (2021) as a special case. Their analysis identifies the endpoint of the path. We study the path itself: the finite-regularization phases, mode-wise singular-value evolution, convergence rate to the SELI direction, and gradient-flow emergence.

B.2 Limitations

We outline possible extensions and remaining challenges for three restrictions of our analysis: the number of classes, balanced initialization, and the form of class imbalance.

In particular, we mention several concrete possible directions of future work:

Number of classes. Our analysis assumes that K is a power of two, which enables the use of Sylvester-Hadamard matrices to establish the diagonal-plus-rank-one factorization. However, with the small exception of auxiliary bounds on $\boldsymbol{b}_+, \boldsymbol{b}_-$, our subsequent results do not utilize the specific structure of Sylvester-Hadamard matrices beyond the existence of an all-ones row and column and orthogonality up to a constant factor. More importantly, (26) along with (27) in Lemma 4 are the main identities enabling this factorization. Consequently, the factorization and the subsequent results should extend to general M in $K = 2M$, given an appropriate replacement for Sylvester-Hadamard matrices. In particular, the discrete Fourier transform (DFT) matrix satisfies these conditions with minor modifications. This suggests a path to extend our results to general $K = 2M$. However, we leave this to future work for two major reasons. First, using the DFT matrices requires additional care to handle the arising complex-valued dynamics and to ensure the logit matrix remains real-valued throughout the dynamics. Second, our current setting already establishes the phenomena of interest, such as the four-parameter structure and mode emergence, which we expect to persist beyond the power-of-two setting.

Balanced initialization in gradient flow. Theorem 5 does *not* require balanced initialization. We impose $\boldsymbol{W}_0^\top \boldsymbol{W}_0 = \boldsymbol{H}_0 \boldsymbol{H}_0^\top$ from Sec. 5.1 onward to simplify the logit dynamics. The difference $\boldsymbol{W}^\top \boldsymbol{W} - \boldsymbol{H}\boldsymbol{H}^\top$ is conserved along GF (Cohen & Razin, 2024; Saxe et al., 2014; 2019; Gidel et al., 2019; Garrod et al., 2025); see App. F.1. Within our spectral parameterization, this gives conserved quantities $c = w^2 - \alpha_H^2 - h^2$ and $\boldsymbol{c}_\pm = \boldsymbol{w}_\pm^{\odot 2} - \boldsymbol{h}_\pm^{\odot 2}$. Specifically for the majority and minority modes, the logit dynamics become $\dot{\boldsymbol{a}}_\pm = \left(4\boldsymbol{a}_\pm^{\odot 2} + \boldsymbol{c}_\pm^{\odot 2}\right)^{\odot 1/2} \odot \boldsymbol{b}_\pm$ instead of $\dot{\boldsymbol{a}}_\pm = 2\,\boldsymbol{a}_\pm \odot \boldsymbol{b}_\pm$. But since under CE the modes diverge, the two dynamics coincide asymptotically. Thus, we expect the convergence picture to hold for any finite imbalance at initialization. For emergence, consider $\|\boldsymbol{c}_\pm\|_\infty = O(\|\boldsymbol{a}_{\pm,0}\|_\infty)$. The local estimates in Lemma 18 motivate replacing $\boldsymbol{b}_+$ and $\boldsymbol{b}_-$ by their values at zero logits, $\sqrt{R}\,\mathbf{1}$ and $\mathbf{1}$, respectively. Under this approximation, the majority and minority modes have the same leading logarithmic escape times as under balancedness, up to additive $O(1)$ terms. We therefore expect the majority-first conclusion of Theorem 6 to persist when all initial layer imbalances, including c, are at this scale. Howver, for large imbalance $\|\boldsymbol{c}_\pm\| \gg \|\boldsymbol{a}_{\pm,0}\|_\infty$ and while the logits remain small, the initial growth of majority and minority modes is approximately

linear at rates $\sqrt{R}\,|\boldsymbol{c}_+|$ and $|\boldsymbol{c}_-|$. Thus, here, the emergence order may indeed depend on the level of initialization imbalance itself. We leave a full analysis without balancedness to future work.

Beyond step-imbalance. Our analysis considers two equally sized groups of classes, with a common sample count within each group. A natural extension is to consider $K = kM$ classes partitioned into k groups of M classes, with R_i samples per class in group $i \in [k]$. We expect Theorem 1, which has enabled our analysis of the step-imbalanced case, to extend to this more general setting as well. Extensions to long-tail imbalances are more challenging and would likely require substantially new ideas.

C Notation and basic facts

For simplicity, we introduce additional necessary notation here. All matrix indexing starts from 0 instead of 1, and $\boldsymbol{\Phi}_2$ is commonly denoted as $\boldsymbol{\Phi}$, with columns $[\boldsymbol{\phi}_0, \boldsymbol{\phi}_1]$. In addition, $\boldsymbol{\Psi}_k$ is defined as $\boldsymbol{\Psi}_K \coloneqq \mathbf{1}_{K-1}\mathbf{1}_{K-1}^\top - \boldsymbol{\Phi}_{[1:,1:]}$. Note that $\boldsymbol{\Psi}_K \in \{0,2\}^{(K-1)\times(K-1)}$ is invertible and satisfies $\boldsymbol{\Psi}\mathbf{1}_{K-1} = K\mathbf{1}_{K-1},\ \boldsymbol{\Psi}^{-1}\mathbf{1}_{K-1} = \frac{1}{K}\mathbf{1}_{K-1}$.

The properties of Hadamard matrices shown in the following lemmas will be useful later in the proofs.

Lemma 1 (Hadamard identities)**.** *For integers m, i, j, consider their $l = \log_2(M)$-bit binary representations $m_1, \ldots, m_l$, $i_1, \ldots, i_l$ and $j_1, \ldots, j_l$, respectively. Accordingly, define the permutation matrices*

$$\boldsymbol{\Pi}_m \coloneqq \boldsymbol{J}^{m_1} \otimes \ldots \otimes \boldsymbol{J}^{m_l}, \quad \boldsymbol{J} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Then, any Sylvester-Hadamard matrix of order M satisfies

$$\boldsymbol{\Pi}_m \overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_i = (-1)^{m.(i+1)} \overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_i, \tag{25}$$

$$\boldsymbol{\Pi}_{i\oplus j}\left(\boldsymbol{\Phi}_M \boldsymbol{e}_m \boldsymbol{e}_m^\top \boldsymbol{\Phi}_M^\top \boldsymbol{e}_j\right) = \boldsymbol{\Phi}_M \boldsymbol{e}_m \boldsymbol{e}_m^\top \boldsymbol{\Phi}_M^\top \boldsymbol{e}_i, \tag{26}$$

where $i \oplus j$ is the xor operation.

In addition, the permutation matrices $\boldsymbol{\Pi}_m$ are symmetric and satisfy

$$\sum_{m=0}^{M-1} \boldsymbol{\Pi}_m = (\boldsymbol{J} + \boldsymbol{I}_2) \otimes \ldots \otimes (\boldsymbol{J} + \boldsymbol{I}_2) = \mathbf{1}_M \mathbf{1}_M^\top. \tag{27}$$

Proof. The key to prove (25) is to expand the RHS as a Kronecker product by using the fact that $\overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_i = \boldsymbol{\Phi}_M \boldsymbol{e}_{i+1}$:

$$\boldsymbol{\Pi}_m \overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_i = \bigotimes_{p=0}^{l} \boldsymbol{J}^{m_p} \boldsymbol{\phi}_{(i+1)_p} = \bigotimes_{p=0}^{l} (-1)^{m_p (i+1)_p} \boldsymbol{\phi}_{(i+1)_p} = (-1)^{m(i+1)} \overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_i.$$

The second identity can be proved from the previous one as follows:

$$\begin{aligned} \boldsymbol{\Pi}_{i\oplus j}(\boldsymbol{\Phi}_M \boldsymbol{e}_m \boldsymbol{e}_m^\top \boldsymbol{\Phi}_M^\top)\boldsymbol{e}_j &= (-1)^{m.j} \boldsymbol{\Pi}_{i\oplus j}(\boldsymbol{\Phi}_M \boldsymbol{e}_m) = (-1)^{m.j}(-1)^{(i+j).m}(\boldsymbol{\Phi}_M \boldsymbol{e}_m) \\ &= (-1)^{m.i}(\boldsymbol{\Phi}_M \boldsymbol{e}_m) = (\boldsymbol{\Phi}_M \boldsymbol{e}_m \boldsymbol{e}_m^\top \boldsymbol{\Phi}_M^\top)\boldsymbol{e}_i. \end{aligned}$$

□

Lemma 2. *The matrix $\boldsymbol{Z} \in \mathbb{R}^{K\times n}$ which is defined below, can be diagonalized as follows:*

$$\boldsymbol{Y} = \boldsymbol{I}_M \otimes \begin{bmatrix} \mathbf{1}_R^\top & 0 \\ \mathbf{0}_R^\top & 1 \end{bmatrix}, \quad \widehat{\boldsymbol{Z}} = (\boldsymbol{I}_k - \frac{1}{K}\mathbf{1}_k\mathbf{1}_k^\top)\boldsymbol{Y}$$

$$\boldsymbol{U} \coloneqq \begin{bmatrix} \frac{1}{\sqrt{2M}}\mathbf{1}_M \otimes \begin{pmatrix} 1 \\ -1 \end{pmatrix} & \frac{1}{\sqrt{M}}\overline{\boldsymbol{\Phi}}_M \otimes \boldsymbol{I}_2 \end{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{Z}} \coloneqq \begin{bmatrix} \sqrt{\frac{R+1}{2}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{M-1} \otimes \boldsymbol{\Sigma}_2 \end{bmatrix} \tag{28}$$

$$\boldsymbol{\Sigma}_2 \coloneqq \begin{bmatrix} \sqrt{R} & 0 \\ 0 & 1 \end{bmatrix} \quad \boldsymbol{V} \coloneqq \begin{bmatrix} \frac{1}{\sqrt{M(R+1)}}\mathbf{1}_M \otimes \begin{pmatrix} \mathbf{1}_R \\ -1 \end{pmatrix} & \frac{1}{\sqrt{M}}\overline{\boldsymbol{\Phi}}_M \otimes \begin{pmatrix} \frac{\mathbf{1}_R}{\sqrt{R}} & \frac{\mathbf{0}_R}{\sqrt{R}} \\ 0 & 1 \end{pmatrix} \end{bmatrix}.$$

Proof. The diagonalization in (28) can be proved by direct matrix multiplication and the proof is omitted for brevity. Alternatively, $\boldsymbol{Z}\boldsymbol{Z}^\top$ can be diagonalized by noting that its eigenvectors are of the form of Kronecker products, which follows from standard linear algebra. □

Lemma 3. *For any two coordinates $r, s \in [M-1]$, there exist permutation matrices*

$$\boldsymbol{P} \in \mathbb{R}^{M\times M}, \qquad \boldsymbol{C} \in \mathbb{R}^{(M-1)\times(M-1)}$$

such that

$$\boldsymbol{P}\overline{\boldsymbol{\Phi}}_M = \overline{\boldsymbol{\Phi}}_M \boldsymbol{C}, \qquad \boldsymbol{C}\boldsymbol{e}_r = \boldsymbol{e}_s.$$

Proof. Let $\ell = \log_2 M$. We use the standard representation of the Sylvester–Hadamard matrix $(\boldsymbol{\Phi}_M)_{\boldsymbol{x},\boldsymbol{y}} = (-1)^{\boldsymbol{x}^\top \boldsymbol{y}}$, where $\boldsymbol{x}, \boldsymbol{y} \in \{0,1\}^\ell$ are binary encodings: $\boldsymbol{x}$ indexes rows and $\boldsymbol{y}$ indexes columns. The column $\boldsymbol{y} = \boldsymbol{0}$ is the all-ones column, so the columns of $\overline{\boldsymbol{\Phi}}_M$ are indexed by the nonzero vectors $\boldsymbol{y} \neq \boldsymbol{0}$.

Let $\boldsymbol{y}_r$ and $\boldsymbol{y}_s$ be the nonzero binary vectors corresponding to coordinates r and s. We choose an invertible binary matrix $\boldsymbol{A}$ such that

$$\boldsymbol{A}^\top \boldsymbol{y}_r = \boldsymbol{y}_s.$$

Such an $\boldsymbol{A}$ exists because any nonzero binary vector can be mapped to any other nonzero binary vector by an invertible linear change of coordinates.

Since $\boldsymbol{A}$ is invertible, the map $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$ is a bijection of the row index set $\{0,1\}^\ell$. Therefore it corresponds to a row permutation matrix $\boldsymbol{P}$.

Now compute the entries of the row-permuted Hadamard matrix:

$$(\boldsymbol{P}\boldsymbol{\Phi}_M)_{\boldsymbol{x},\boldsymbol{y}} = (\boldsymbol{\Phi}_M)_{\boldsymbol{A}\boldsymbol{x},\boldsymbol{y}} = (-1)^{(\boldsymbol{A}\boldsymbol{x})^\top \boldsymbol{y}} = (-1)^{\boldsymbol{x}^\top \boldsymbol{A}^\top \boldsymbol{y}}.$$

But the last expression is exactly the Hadamard column indexed by $\boldsymbol{A}^\top \boldsymbol{y}$. Therefore, after the row permutation $\boldsymbol{P}$, the column indexed by $\boldsymbol{y}$ becomes the column indexed by $\boldsymbol{A}^\top \boldsymbol{y}$.

Since $\boldsymbol{A}^\top$ is invertible, it permutes all nonzero binary vectors and fixes $\boldsymbol{0}$. Hence the all-ones column stays fixed, and the remaining columns are merely permuted. Therefore there exists a permutation matrix $\boldsymbol{C}$ such that

$$\boldsymbol{P}\overline{\boldsymbol{\Phi}}_M = \overline{\boldsymbol{\Phi}}_M \boldsymbol{C}.$$

Moreover, because we chose $\boldsymbol{A}$ so that $\boldsymbol{A}^\top \boldsymbol{y}_r = \boldsymbol{y}_s$, the induced column permutation sends coordinate r to coordinate s, i.e. $\boldsymbol{C}\boldsymbol{e}_r = \boldsymbol{e}_s$. □

D Proof of Theorem 1

In this section, we prove that the diagonal-plus-rank-one factorization is in fact a special case of a more general block-diagonal-plus-rank-one factorization which is made possible by Sylvester-Hadamard matrices. After that, we provide simplified expressions on the parameters and elementary upper and lower bounds on them, which will be used in other sections.

To highlight the appearance of the rank-one term, we first prove the general permutation structure *without* having this term inside the softmax and later show that the results are retained after adding the rank-one term, with minor changes to the expressions. Recall the definitions of the singular bases $\boldsymbol{U}, \boldsymbol{V}$ and the extended basis $\widehat{\boldsymbol{V}}$ in Sec. 3. Direct computation confirms that $\widehat{\boldsymbol{Z}} = \boldsymbol{U}\,\mathrm{diag}\,(\boldsymbol{\sigma})\,\boldsymbol{V}^\top$ where $\boldsymbol{\sigma} = [\sqrt{\frac{R+1}{2}}, \sqrt{R}\boldsymbol{1} \otimes \boldsymbol{e}_+ + \boldsymbol{1} \otimes \boldsymbol{e}_-]$.

D.1 Permutation structure of the softmax residual

Lemma 4. *For any $\boldsymbol{A} \in \mathbb{R}^{(K-1)\times(K-1)}$ having the block structure*

$$\boldsymbol{A} = \mathrm{diag}\,(a, \boldsymbol{A}_0, \ldots \boldsymbol{A}_{M-2}) = \mathrm{diag}\left(a, \sum_{m=0}^{M-2} (\boldsymbol{e}_m \boldsymbol{e}_m^\top) \otimes \boldsymbol{A}_m\right), \quad \boldsymbol{A}_m \in \mathbb{R}^{2\times 2},$$

the i-th block column of $\boldsymbol{Y} - \mathbb{S}\left(\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}^\top\right)$, denoted as $(\boldsymbol{Y} - \mathbb{S}\left(\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}^\top\right))_i \in \mathbb{R}^{K\times(R+1)}$ satisfies

$$(\boldsymbol{Y} - \mathbb{S}\left(\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}^\top\right))_i = (\boldsymbol{\Pi}_i \otimes \boldsymbol{I}_2)\begin{bmatrix}\boldsymbol{p}\boldsymbol{1}_R^\top & \boldsymbol{q}\end{bmatrix}, \qquad \boldsymbol{p} = \boldsymbol{e}_0 - \mathbb{S}(\boldsymbol{x}),\ \boldsymbol{q} = \boldsymbol{e}_1 - \mathbb{S}(\boldsymbol{y}),$$

where $\boldsymbol{x}, \boldsymbol{y}$ are defined as

$$\begin{aligned}\boldsymbol{x} &\coloneqq \frac{1}{2M}\left(\sqrt{\frac{2}{R+1}}a\boldsymbol{1}_M \otimes \begin{pmatrix}1\\-1\end{pmatrix} + 2\sum_{m=0}^{M-2}\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_m \otimes \frac{\boldsymbol{a}_{m,0}}{\sqrt{R}}\right)\\ \boldsymbol{y} &\coloneqq \frac{1}{2M}\left(-\sqrt{\frac{2}{R+1}}a\boldsymbol{1}_M \otimes \begin{pmatrix}1\\-1\end{pmatrix} + 2\sum_{m=0}^{M-2}\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_m \otimes \boldsymbol{a}_{m,1}\right).\end{aligned} \tag{29}$$

Proof. The first step is to calculate $\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}^\top$, to reach

$$\begin{aligned}\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}^\top = &\frac{a}{\sqrt{R+1}}\left(\frac{1}{M}\boldsymbol{1}_M\boldsymbol{1}_M^\top\right) \otimes \left[\tfrac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}\boldsymbol{1}_R^\top \quad -\tfrac{1}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}\right]\\ &+ \frac{1}{2}\sum_{m=0}^{M-2}\left(\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_m\boldsymbol{e}_m^\top\overline{\boldsymbol{\Phi}}_M^\top\right) \otimes \left(\boldsymbol{A}_m\left[\begin{pmatrix}1\\0\end{pmatrix}\tfrac{\boldsymbol{1}_R^\top}{\sqrt{R}} \quad \begin{pmatrix}0\\1\end{pmatrix}\right]\right).\end{aligned} \tag{30}$$

Next, we analyze the structure of $\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_m\boldsymbol{e}_m^\top\overline{\boldsymbol{\Phi}}_M^\top$. Using $\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_m = \boldsymbol{\Phi}\boldsymbol{e}_{m+1}$ and defining $\boldsymbol{a}_{m,0} \coloneqq \boldsymbol{A}_m\boldsymbol{e}_0, \boldsymbol{a}_{m,1} \coloneqq \boldsymbol{A}_m\boldsymbol{e}_1$, the summation can be changed to

$$\sum_{m=0}^{M-2}\left(\boldsymbol{\Phi}_M\boldsymbol{e}_{m+1}\boldsymbol{e}_{m+1}^\top\boldsymbol{\Phi}_M^\top\right) \otimes \left[\boldsymbol{a}_{m,0}\tfrac{\boldsymbol{1}_R^\top}{\sqrt{R}} \quad \boldsymbol{a}_{m,1}\right].$$

The result of Lemma (26) shows that all columns of $\boldsymbol{\Phi}_M\boldsymbol{e}_m\boldsymbol{e}_m^\top\boldsymbol{\Phi}_M^\top$ are permutations of each other. Consequently, we fix the first column as reference, to expand the first term in the Kronecker product as

$$\boldsymbol{\Phi}_M\boldsymbol{e}_{m+1}\boldsymbol{e}_{m+1}^\top\boldsymbol{\Phi}_M^\top = [\boldsymbol{\Phi}_M\boldsymbol{e}_{m+1} \quad \dots \quad \boldsymbol{\Pi}_{M-1}(\boldsymbol{\Phi}_M\boldsymbol{e}_{m+1})] \tag{31}$$

where we have used the fact that $\boldsymbol{\Pi}_{i\oplus 0} = \boldsymbol{\Pi}_i$. Calculating the above summation and combining it with the term involving a would yield

$$\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}^\top = \begin{bmatrix}\boldsymbol{x}\boldsymbol{1}_R^\top & \boldsymbol{y} & (\widetilde{\boldsymbol{\Pi}}_1\boldsymbol{x})\boldsymbol{1}_R^\top & \widetilde{\boldsymbol{\Pi}}_1\boldsymbol{y} & \dots & (\widetilde{\boldsymbol{\Pi}}_{M-1}\boldsymbol{x})\boldsymbol{1}_R^\top & \widetilde{\boldsymbol{\Pi}}_{M-1}\boldsymbol{y}\end{bmatrix}, \tag{32}$$

where $\widetilde{\boldsymbol{\Pi}}_m \coloneqq \boldsymbol{\Pi}_m \otimes \boldsymbol{I}_2$ and $\boldsymbol{x}, \boldsymbol{y}$ are defined as above.

Since for any permutation matrix $\boldsymbol{\Pi}$, the softmax operator satisfies $\mathbb{S}(\boldsymbol{\Pi}\boldsymbol{x}) = \boldsymbol{\Pi}\mathbb{S}(\boldsymbol{x})$, the expression $\mathbb{S}\left(\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}^\top\right)$ will also have the same block column structure. In addition, using $(\boldsymbol{\Pi}_i \otimes \boldsymbol{I}_2)\boldsymbol{e}_0 = \boldsymbol{e}_{2i}$ and $(\boldsymbol{\Pi}_i \otimes \boldsymbol{I}_2)\boldsymbol{e}_1 = \boldsymbol{e}_{2i+1}$, it can be shown that $(\boldsymbol{Y} - \mathbb{S}\left(\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}^\top\right))_i$ satisfies

$$(\boldsymbol{Y} - \mathbb{S}\left(\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}^\top\right))_i = (\boldsymbol{\Pi}_i \otimes \boldsymbol{I}_2)\left[(\boldsymbol{e}_0 - \mathbb{S}(\boldsymbol{x}))\boldsymbol{1}_R^\top \quad \boldsymbol{e}_1 - \mathbb{S}(\boldsymbol{y})\right]$$

□

Corollary 2. *If the matrix $\boldsymbol{A}$ is diagonal, with each block $\boldsymbol{A}_m = \operatorname{diag}([a_{m,0},\ a_{m,1}])$, the softmax expressions simplify to*

$$\begin{aligned}\widetilde{\boldsymbol{x}} &\coloneqq \frac{1}{M}\left(\sqrt{\frac{2}{R+1}}a\boldsymbol{1}_M + \frac{1}{\sqrt{R}}\overline{\boldsymbol{\Phi}}_M\boldsymbol{a}_+\right), \quad \mathbb{S}(\boldsymbol{x}) = \mathbb{S}\left(\widetilde{\boldsymbol{x}} \otimes \begin{pmatrix}1\\0\end{pmatrix}\right)\\ \widetilde{\boldsymbol{y}} &\coloneqq \frac{1}{M}\left(\sqrt{\frac{2}{R+1}}a\boldsymbol{1}_M + \overline{\boldsymbol{\Phi}}_M\boldsymbol{a}_-\right) \quad \mathbb{S}(\boldsymbol{y}) = \mathbb{S}\left(\widetilde{\boldsymbol{y}} \otimes \begin{pmatrix}0\\1\end{pmatrix}\right),\end{aligned} \tag{33}$$

where $\boldsymbol{a}_+, \boldsymbol{a}_- \in \mathbb{R}^{M-2}$ and $(\boldsymbol{a}_+)_m = a_{m,0}$, $(\boldsymbol{a}_-)_m = a_{m,1}$.

Proof. Direct substitution of diagonal $\boldsymbol{A}_m$ in (29) leads to

$$\begin{aligned}\boldsymbol{x} &= \frac{1}{2M}\left(\sqrt{\frac{2}{R+1}}a\mathbf{1}_M \otimes \begin{pmatrix}1\\-1\end{pmatrix} + 2\sum_{m=0}^{M-2}\frac{a_{m,0}}{\sqrt{R}}\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_m \otimes \begin{pmatrix}1\\0\end{pmatrix}\right)\\ \boldsymbol{y} &= \frac{1}{2M}\left(-\sqrt{\frac{2}{R+1}}a\mathbf{1}_M \otimes \begin{pmatrix}1\\-1\end{pmatrix} + 2\sum_{m=0}^{M-2}a_{m,1}\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_m \otimes \begin{pmatrix}0\\1\end{pmatrix}\right).\end{aligned} \tag{34}$$

At this point, notice that adding or subtracting multiples of $\mathbf{1}_K$ from $\boldsymbol{x}, \boldsymbol{y}$ does not change the softmax value. As a result, by adding appropriate constant vectors to $\boldsymbol{x}, \boldsymbol{y}$ to change their first term, we reach the desired result. □

D.2 Spectral factorization and rank-one correction

The following theorem establishes that for any matrix $\boldsymbol{Q}$ with the specified permutation structure, the projection $\boldsymbol{U}^\top\boldsymbol{Q}\boldsymbol{V}$ is block diagonal with the same pattern as the argument $\boldsymbol{A}$.

Theorem 9. *Suppose a matrix $\boldsymbol{Q}$ has the block structure $\boldsymbol{Q}_i = (\boldsymbol{\Pi}_i \otimes \boldsymbol{I}_2)\begin{bmatrix}\boldsymbol{p}\mathbf{1}_R^\top & \boldsymbol{q}\end{bmatrix}$. Then, for $\boldsymbol{U}, \boldsymbol{V}$ as defined in (28)*

$$\boldsymbol{U}^\top\boldsymbol{Q}\boldsymbol{V} = \boldsymbol{B} = \operatorname{diag}(b, \boldsymbol{B}_0, \dots \boldsymbol{B}_{M-2}), \quad \boldsymbol{B}_i \in \mathbb{R}^{2\times 2}, \tag{35}$$

$$b = \frac{1}{\sqrt{2(R+1)}}(\overline{\boldsymbol{\Phi}}_K\boldsymbol{e}_0)^\top(R\boldsymbol{p}-\boldsymbol{q}), \quad \boldsymbol{B}_i = \left((\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_i)^\top \otimes \boldsymbol{I}_2\right)\begin{bmatrix}\sqrt{R}\boldsymbol{p} & \boldsymbol{q}\end{bmatrix} \tag{36}$$

Proof. The matrix $\boldsymbol{Q}$ can be expressed as $\boldsymbol{Q} = \sum_{m=0}^{M-1}\boldsymbol{e}_m^\top \otimes \left((\boldsymbol{\Pi}_m \otimes \boldsymbol{I}_2)\begin{bmatrix}\boldsymbol{p}\mathbf{1}_R^\top & \boldsymbol{q}\end{bmatrix}\right)$.

Using this expansion and (27), the first column of $\boldsymbol{B}$ can be established as

$$\begin{aligned}\boldsymbol{Q}\boldsymbol{v}_0 &= \frac{1}{\sqrt{M(R+1)}}\sum_{m=0}^{M-1}(\boldsymbol{\Pi}_m \otimes \boldsymbol{I}_2)(R\boldsymbol{p}-\boldsymbol{q}) = \frac{1}{\sqrt{M(R+1)}}\left((\mathbf{1}_M\mathbf{1}_M^\top) \otimes \boldsymbol{I}_2\right)(R\boldsymbol{p}-\boldsymbol{q})\\ \boldsymbol{B}\boldsymbol{e}_0 &= \boldsymbol{U}^\top\boldsymbol{Q}\boldsymbol{v}_0 = \frac{1}{\sqrt{2(R+1)}}(\overline{\boldsymbol{\Phi}}_K\boldsymbol{e}_0)^\top(R\boldsymbol{p}-\boldsymbol{q})\boldsymbol{e}_0 = b\boldsymbol{e}_0.\end{aligned} \tag{37}$$

Next, we will show that the first row is also the same multiple b of $\boldsymbol{e}_0^\top$:

$$\begin{aligned}\boldsymbol{u}_0^\top\boldsymbol{Q} &= \sum_{m=0}^{M-1}\boldsymbol{e}_m^\top \otimes (\underbrace{\boldsymbol{u}_0^\top(\boldsymbol{\Pi}_m \otimes \boldsymbol{I}_2)}_{\boldsymbol{u}_0^\top}\begin{bmatrix}\boldsymbol{p}\mathbf{1}_R^\top & \boldsymbol{q}\end{bmatrix}) = \mathbf{1}_M^\top \otimes \begin{bmatrix}(\boldsymbol{u}_0^\top\boldsymbol{p})\mathbf{1}_R^\top & \boldsymbol{u}_0^\top\boldsymbol{q}\end{bmatrix})\\ \boldsymbol{u}_0^\top\boldsymbol{Q}\boldsymbol{V} &= \frac{\sqrt{2M}\boldsymbol{u}_0^\top(R\boldsymbol{p}-\boldsymbol{q})}{\sqrt{2(R+1)}}\boldsymbol{e}_0^\top = \frac{(\overline{\boldsymbol{\Phi}}_K\boldsymbol{e}_0)^\top(R\boldsymbol{p}-\boldsymbol{q})}{\sqrt{2(R+1)}}\boldsymbol{e}_0^\top = b\boldsymbol{e}_0^\top.\end{aligned} \tag{38}$$

The final phase would be to compute each block of the sub-matrix $\boldsymbol{B}_{[1:,1:]}$. Recall that the matrix $\boldsymbol{Q}$ can be expressed as $\boldsymbol{Q} = \sum_{m=0}^{M-1}\boldsymbol{e}_m^\top \otimes \left((\boldsymbol{\Pi}_m \otimes \boldsymbol{I}_2)\begin{bmatrix}\boldsymbol{p}\mathbf{1}_R^\top & \boldsymbol{q}\end{bmatrix}\right)$. Expanding the expression for each block would result in

$$\begin{aligned}(\boldsymbol{B}_{[1:,1:]})_{ij} &= (\boldsymbol{U}_{[:,1:]}^\top\boldsymbol{Q}\boldsymbol{V}_{[:,1:]})_{ij} = (\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_i \otimes \boldsymbol{I}_2)^\top\boldsymbol{Q}\left((\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_j) \otimes \frac{1}{\sqrt{2}}\begin{bmatrix}\frac{\mathbf{1}_R}{\sqrt{R}} & \frac{\mathbf{0}_R}{\sqrt{R}}\\ 0 & 1\end{bmatrix}\right)\\ &= \frac{1}{M}(\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_i \otimes \boldsymbol{\Phi})^\top\sum_{m=0}^{M-1}(\boldsymbol{e}_m^\top\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_j)(\boldsymbol{\Pi}_m \otimes \boldsymbol{I}_2)\begin{bmatrix}\sqrt{R}\boldsymbol{p} & \boldsymbol{q}\end{bmatrix}\\ &= \frac{1}{M}\sum_{m=0}^{M-1}(-1)^{m.(j+1)}\left((\boldsymbol{\Pi}_m\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_i)^\top \otimes \boldsymbol{I}_2\right)\begin{bmatrix}\sqrt{R}\boldsymbol{p} & \boldsymbol{q}\end{bmatrix},\end{aligned}$$

where we have used the symmetry of $\boldsymbol{\Pi}_m, \boldsymbol{\Phi}$, and the mixed-product Kronecker identity. Next, we use the identity (25) for $\boldsymbol{\Pi}_m$ to simplify the expression to

$$\begin{aligned}(\boldsymbol{B}_{[1:,1:]})_{ij} &= \frac{1}{M}\sum_{m=0}^{M-1}(-1)^{m.(j+i)}\left((\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_i)^\top \otimes \boldsymbol{I}_2\right)\begin{bmatrix}\sqrt{R}\boldsymbol{p} & \boldsymbol{q}\end{bmatrix}\\ &= \delta_{ij}\left((\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_i)^\top \otimes \boldsymbol{I}_2\right)\begin{bmatrix}\sqrt{R}\boldsymbol{p} & \boldsymbol{q}\end{bmatrix},\end{aligned}$$

and the desired result is proved. □

The following lemma proves the rank-one difference for matrices $\boldsymbol{Q}$ with zero-sum columns and the specific permutation structure of this section.

Lemma 5. *The matrices $\boldsymbol{U}, \boldsymbol{V}$ as defined in (28), satisfy*

$$\begin{aligned} \boldsymbol{U}\boldsymbol{U}^\top &= \boldsymbol{I}_k - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top \\ \boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{I}_n &= \boldsymbol{I}_M \otimes \begin{bmatrix} \frac{\mathbf{1}_R\mathbf{1}_R^\top}{R} - \boldsymbol{I}_R & \mathbf{0}_R \\ \mathbf{0}_R^\top & 0 \end{bmatrix} - \frac{1}{M(R+1)}(\mathbf{1}_M\mathbf{1}_M^\top) \otimes \begin{bmatrix} \frac{\mathbf{1}_R\mathbf{1}_R^\top}{R} & \mathbf{1}_R \\ \mathbf{1}_R^\top & R \end{bmatrix}. \end{aligned} \tag{39}$$

Furthermore, it can be shown that for zero-sum vectors $\boldsymbol{p}$ and $\boldsymbol{q}$, and the matrix $\boldsymbol{Q} \in \mathbb{R}^{K\times n}$ defined such that its i-th column block $\boldsymbol{Q}_i = (\boldsymbol{\Pi}_i \otimes \boldsymbol{I}_2)\begin{bmatrix}\boldsymbol{p}\mathbf{1}_R^\top & \boldsymbol{q}\end{bmatrix}$, the following hold:

$$\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{Q} = \boldsymbol{Q}, \quad \boldsymbol{Q}(\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{I}_k) = -\beta\boldsymbol{u}_0\widetilde{\boldsymbol{v}}^\top, \tag{40}$$

$$\beta = \sqrt{\frac{R}{2(R+1)}}(\overline{\boldsymbol{\Phi}}_K\boldsymbol{e}_0)^\top(\boldsymbol{p}+\boldsymbol{q}), \quad \boldsymbol{u}_0 = \boldsymbol{U}\boldsymbol{e}_0 = \overline{\boldsymbol{\Phi}}_K\boldsymbol{e}_0, \quad \widetilde{\boldsymbol{v}} = \frac{\mathbf{1}_M \otimes \begin{pmatrix}\mathbf{1}_R \\ R\end{pmatrix}}{\sqrt{MR(R+1)}} \tag{41}$$

Proof. Equation (39) can be proved by direct calculation. The identity $\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{Q} = \boldsymbol{Q}$ can be proved that every column of $\boldsymbol{Q}$ consists of permutations of zero-sum vectors $\boldsymbol{p}, \boldsymbol{q}$ which do not change under the projection $\boldsymbol{U}\boldsymbol{U}^\top$.

Hence, we focus on proving the rank-1 difference between $\boldsymbol{Q}$ and $\boldsymbol{Q}\boldsymbol{V}\boldsymbol{V}^\top$:

$$\begin{aligned} \boldsymbol{Q}(\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{I}_n) &= \sum_{m=0}^{M-1}\left(\boldsymbol{e}_m^\top \otimes \mathbf{0}_{K\times(R+1)} - \frac{1}{M(R+1)}\mathbf{1}_M^\top \otimes \left((\boldsymbol{\Pi}_m \otimes \boldsymbol{I}_2)(\boldsymbol{p}+\boldsymbol{q})\begin{bmatrix}\mathbf{1}_R^\top & R\end{bmatrix}\right)\right) \\ &= -\frac{1}{M(R+1)}\mathbf{1}_M^\top \otimes \left(\left((\mathbf{1}_M\mathbf{1}_M^\top) \otimes \boldsymbol{I}_2\right)(\boldsymbol{p}+\boldsymbol{q})\begin{bmatrix}\mathbf{1}_R^\top & R\end{bmatrix}\right) \\ &= -\frac{1}{M(R+1)}\left(\mathbf{1}_M \otimes \left((\mathbf{1}_M^\top \otimes \boldsymbol{I}_2)(\boldsymbol{p}+\boldsymbol{q})\right)\right)\left(\mathbf{1}_M^\top \otimes \begin{bmatrix}\mathbf{1}_R^\top & R\end{bmatrix}\right). \end{aligned} \tag{42}$$

In addition, the fact that $\boldsymbol{p}, \boldsymbol{q}$ are zero-sum can be used to reach the following simplification:

$$\begin{aligned} (\mathbf{1}_M^\top \otimes \boldsymbol{I}_2)(\boldsymbol{p}+\boldsymbol{q}) &= (\mathbf{1}_M^\top \otimes (\boldsymbol{I}_2 - \frac{1}{2}\mathbf{1}_2\mathbf{1}_2^\top))(\boldsymbol{p}+\boldsymbol{q}) = \frac{1}{2}\left(\mathbf{1}_M^\top \otimes \begin{bmatrix}1 & -1 \\ -1 & 1\end{bmatrix}\right)(\boldsymbol{p}+\boldsymbol{q}) \\ &= \frac{1}{2}\begin{pmatrix}1 \\ -1\end{pmatrix}\left(\mathbf{1}_M^\top \otimes \begin{pmatrix}1 & -1\end{pmatrix}\right)(\boldsymbol{p}+\boldsymbol{q}) = \frac{1}{2}\begin{pmatrix}1 \\ -1\end{pmatrix}(\overline{\boldsymbol{\Phi}}_K\boldsymbol{e}_0)^\top(\boldsymbol{p}+\boldsymbol{q}). \end{aligned} \tag{43}$$

Normalizing each row and column vector using the above simplification in (43) would lead to the desired result. □

Remark 1. *Even in the case of $R = 1$, the rank-1 difference does not vanish if a block diagonal $\boldsymbol{A}$ is used, and it only vanishes for a diagonal matrix inside the softmax structure (Garrod et al., 2025).*

The following lemma will show how Theorem 9 and Lemma 5 can be used to give a factorization of the matrix $\boldsymbol{Q}$:

Lemma 6. *Suppose a matrix $\boldsymbol{Q}$, for given $\boldsymbol{U}, \boldsymbol{V}, \widetilde{\boldsymbol{v}}$ and $\boldsymbol{B} \coloneqq \boldsymbol{U}^\top\boldsymbol{Q}\boldsymbol{V}$ satisfies*

$$\boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{Q} = \boldsymbol{Q}, \ \boldsymbol{Q}(\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{I}_k) = -\beta(\boldsymbol{U}\boldsymbol{e}_0)\widetilde{\boldsymbol{v}}^\top.$$

then, the matrix $\boldsymbol{Q}$ can be decomposed as $\boldsymbol{Q} = \boldsymbol{U}\begin{bmatrix}\beta\boldsymbol{e}_0 & \boldsymbol{B}\end{bmatrix}\begin{bmatrix}\widetilde{\boldsymbol{v}} & \boldsymbol{V}\end{bmatrix}^\top$.

Proof. This lemma can be proved by direct calculation based on the assumptions:

$$\boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^\top = \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{Q}\boldsymbol{V}\boldsymbol{V}^\top = \boldsymbol{Q} - \beta\boldsymbol{u}_0\widetilde{\boldsymbol{v}}^\top = \boldsymbol{Q} - \boldsymbol{U}(\beta\boldsymbol{e}_0)\widetilde{\boldsymbol{v}}^\top,$$

after which rearranging and rewriting in block form will prove the result. □

D.3 Adding the rank-one term

We have shown that a block-diagonal spectral argument produces a residual with a block-diagonal component and an additional rank-one term. We now include this rank-one term in the input as well. The following theorem shows that the resulting family is preserved by the residual map.

Theorem 10. *Suppose the matrix $\boldsymbol{A} \in \mathbb{R}^{(K-1)\times(K-1)}$ has the block structure*

$$\boldsymbol{A} = \operatorname{diag}\left(a, \boldsymbol{A}_0, \dots \boldsymbol{A}_{M-2}\right), \quad \boldsymbol{A}_m \in \mathbb{R}^{2\times 2}$$

Then, for $\boldsymbol{U}, \boldsymbol{V}$ defined as in (28), the block columns of the matrix $\boldsymbol{Q} = \boldsymbol{Y} - \mathbb{S}\left(\boldsymbol{U}\begin{bmatrix}\alpha \boldsymbol{e}_0 & \boldsymbol{A}\end{bmatrix}\widetilde{\boldsymbol{V}}^\top\right)$ satisfy

$$\boldsymbol{Q}_i = \left(\boldsymbol{\Pi}_i \otimes \boldsymbol{I}_2\right)\begin{bmatrix}\boldsymbol{p}\boldsymbol{1}_R^\top & \boldsymbol{q}\end{bmatrix}, \quad \boldsymbol{p} = \boldsymbol{e}_0 - \mathbb{S}\left(\boldsymbol{x}\right), \quad \boldsymbol{q} = \boldsymbol{e}_1 - \mathbb{S}\left(\boldsymbol{y}\right),$$

where $\boldsymbol{x}, \boldsymbol{y}$ are defined as

$$\begin{aligned}
\boldsymbol{x} &\coloneqq \frac{1}{2M}\left(\sqrt{\frac{2}{R+1}}(a + \frac{\alpha}{\sqrt{R}})\boldsymbol{1}_M \otimes \begin{pmatrix}1\\-1\end{pmatrix} + 2\sum_{m=0}^{M-2}\overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_m \otimes \frac{\boldsymbol{a}_{m,0}}{\sqrt{R}}\right)\\
\boldsymbol{y} &\coloneqq \frac{1}{2M}\left(\sqrt{\frac{2}{R+1}}(-a + \alpha\sqrt{R})\boldsymbol{1}_M \otimes \begin{pmatrix}1\\-1\end{pmatrix} + 2\sum_{m=0}^{M-2}\overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_m \otimes \boldsymbol{a}_{m,1}\right).
\end{aligned} \tag{44}$$

In addition, it also holds that $\boldsymbol{Q} = \boldsymbol{U}\begin{bmatrix}\beta\boldsymbol{e}_0 & \boldsymbol{B}\end{bmatrix}\begin{bmatrix}\widetilde{\boldsymbol{v}} & \boldsymbol{V}\end{bmatrix}^\top$, with parameters

$$\boldsymbol{B} = \operatorname{diag}\left(b, \boldsymbol{B}_0, \dots \boldsymbol{B}_{M-2}\right), \quad \boldsymbol{B}_i \in \mathbb{R}^{2\times 2}, \tag{45}$$

$$b = \frac{1}{\sqrt{2(R+1)}}(\overline{\boldsymbol{\Phi}}_K \boldsymbol{e}_0)^\top(R\boldsymbol{p} - \boldsymbol{q}), \quad \beta = \sqrt{\frac{R}{2(R+1)}}(\overline{\boldsymbol{\Phi}}_K \boldsymbol{e}_0)^\top(\boldsymbol{p} + \boldsymbol{q}), \tag{46}$$

$$\boldsymbol{B}_i = \left((\overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_i)^\top \otimes \boldsymbol{I}_2\right)\begin{bmatrix}\sqrt{R}\boldsymbol{p} & \boldsymbol{q}\end{bmatrix} \quad \widetilde{\boldsymbol{v}} = \frac{\boldsymbol{1}_M \otimes \begin{pmatrix}\boldsymbol{1}_R\\ R\end{pmatrix}}{\sqrt{MR(R+1)}} \tag{47}$$

Proof. All parts directly follow from Lemma 4, Theorem 9, Lemma 5, Lemma 6, and Corollary 2. Note that the additional $\alpha\boldsymbol{u}_0\widetilde{\boldsymbol{v}}_0^\top$ introduces an additional

$$\frac{\alpha}{\sqrt{R+1}}\left(\frac{1}{M}\boldsymbol{1}_M\boldsymbol{1}_M^\top\right) \otimes \left[\frac{1}{\sqrt{2R}}\begin{pmatrix}1\\-1\end{pmatrix}\boldsymbol{1}_R^\top \quad -\frac{\sqrt{R}}{\sqrt{2}}\begin{pmatrix}1\\-1\end{pmatrix}\right]$$

into the expansion (30), which has been accounted for in the adjusted $\boldsymbol{x}, \boldsymbol{y}$ expressions. □

D.4 Diagonal specialization

Finally, we obtain Theorem 1 by specializing the block-diagonal theorem above to diagonal $\boldsymbol{A}$. For convenience, we re-state the result with its explicit coefficients below.

Theorem 11 (Restatement of Theorem 1)**.** *Fix any $\alpha, a \in \mathbb{R}$ and $\boldsymbol{a}_+, \boldsymbol{a}_- \in \mathbb{R}^{M-1}$. With these, define vectors $\boldsymbol{p}, \boldsymbol{q} \in \mathbb{R}^K$ as follows*

$$\begin{aligned}
\boldsymbol{p}_+ &\coloneqq \boldsymbol{e}_0 - \mathbb{S}\left(\boldsymbol{x}_+ \otimes \begin{pmatrix}1\\0\end{pmatrix}\right), \qquad \boldsymbol{x}_+ \coloneqq \frac{1}{M}\left(\sqrt{\frac{2}{R+1}}(a + \frac{\alpha}{\sqrt{R}})\boldsymbol{1}_M + \frac{1}{\sqrt{R}}\overline{\boldsymbol{\Phi}}_M \boldsymbol{a}_+\right),\\
\boldsymbol{p}_- &\coloneqq \boldsymbol{e}_1 - \mathbb{S}\left(\boldsymbol{x}_- \otimes \begin{pmatrix}0\\1\end{pmatrix}\right), \qquad \boldsymbol{x}_- \coloneqq \frac{1}{M}\left(\sqrt{\frac{2}{R+1}}(a - \alpha\sqrt{R})\boldsymbol{1}_M + \overline{\boldsymbol{\Phi}}_M \boldsymbol{a}_-\right).
\end{aligned} \tag{48}$$

Finally, for diagonal matrix $\boldsymbol{A} = \operatorname{diag}\left(a, \boldsymbol{a}_+ \otimes \boldsymbol{e}_+ + \boldsymbol{a}_- \otimes \boldsymbol{e}_-\right) \in \mathbb{R}^{(K-1)\times(K-1)}$, define the matrix

$$\boldsymbol{Q} = \boldsymbol{Y} - \mathbb{S}\left(\boldsymbol{U}\begin{bmatrix}\alpha\boldsymbol{e}_0 & \boldsymbol{A}\end{bmatrix}\tilde{\boldsymbol{V}}^\top\right) \in \mathbb{R}^{K\times(M)(R+1)}.$$

Then, the following two statements are true about $\boldsymbol{Q}$:

1. *The M block columns of the matrix $\boldsymbol{Q}$ satisfy*

$$\boldsymbol{Q}_i = (\Pi_i \otimes \boldsymbol{I}_2)\begin{bmatrix}\boldsymbol{p}_+ \mathbf{1}_R^\top & \boldsymbol{p}_-\end{bmatrix} \in \mathbb{R}^{K\times(R+1)},\ i \in [M], \qquad \boldsymbol{p}_+ = \boldsymbol{e}_0 - \mathbb{S}(\boldsymbol{x}),\ \boldsymbol{p}_- = \boldsymbol{e}_1 - \mathbb{S}(\boldsymbol{y})$$

2. *$\boldsymbol{Q}$ admits a decomposition as follows*

$$\boldsymbol{Q} = \boldsymbol{U}\begin{bmatrix}\beta \boldsymbol{e}_0 & \boldsymbol{B}\end{bmatrix}\tilde{\boldsymbol{V}}^\top$$

where $\beta \in \mathbb{R}$ and the diagonal matrix $\boldsymbol{B} \coloneqq \operatorname{diag}(b, \boldsymbol{b}_+ \otimes \boldsymbol{e}_+ + \boldsymbol{b}_- \otimes \boldsymbol{e}_-)$ with $\boldsymbol{b}_+ \in \mathbb{R}^{M-1}$, and $\boldsymbol{b}_- \in \mathbb{R}^{M-1}$ are given in terms of $\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-$ as follows

$$\begin{aligned} \beta &= \sqrt{\frac{R}{2(R+1)}}(\overline{\boldsymbol{\Phi}}_K \boldsymbol{e}_0)^\top(\boldsymbol{p}_+ + \boldsymbol{p}_-), \qquad b = \frac{1}{\sqrt{2(R+1)}}(\overline{\boldsymbol{\Phi}}_K \boldsymbol{e}_0)^\top(R\boldsymbol{p}_+ - \boldsymbol{p}_-), \\ \boldsymbol{b}_+ &= \sqrt{R}(\overline{\boldsymbol{\Phi}}_M \otimes \boldsymbol{e}_+)^\top \boldsymbol{p}_+, \qquad\qquad \boldsymbol{b}_- = (\overline{\boldsymbol{\Phi}}_M \otimes \boldsymbol{e}_-)^\top \boldsymbol{p}_-\,. \end{aligned} \tag{49}$$

D.5 Simplified formulas

Lemma 7. *For $\beta, b, \boldsymbol{b}_+$, and $\boldsymbol{b}_-$ defined as in (49), the following simplified expressions hold:*

$$\begin{aligned} \beta &= \sqrt{\frac{R}{2(R+1)}}\left(\frac{2M}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_+) + M} - \frac{2M}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_-) + M}\right) \\ b &= \frac{1}{\sqrt{2(R+1)}}\left(\frac{2RM}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_+) + M} + \frac{2M}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_-) + M}\right) \\ \boldsymbol{b}_+ &= \sqrt{R}\left(\mathbf{1}_{M-1} - \frac{\overline{\boldsymbol{\Phi}}_M^\top \exp(\boldsymbol{x}_+)}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_+) + M}\right), \quad \boldsymbol{b}_- = \mathbf{1}_{M-1} - \frac{\overline{\boldsymbol{\Phi}}_M^\top \exp(\boldsymbol{x}_-)}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_-) + M}. \end{aligned}$$

In addition, $\boldsymbol{b}_\pm \geq \mathbf{0}$, $\boldsymbol{b}_+ = \sqrt{R}\mathbf{1}_{M-1}$ if $\boldsymbol{a}_+ = \mathbf{0}$ and $\boldsymbol{b}_- = \mathbf{1}_{M-1}$ if $\boldsymbol{a}_- = \mathbf{0}$. It also holds that $\boldsymbol{b}_+ \leq \sqrt{R}\mathbf{1}_{M-1}$ if $\boldsymbol{a}_+ \geq \mathbf{0}$, and $\boldsymbol{b}_- \leq \mathbf{1}_{M-1}$ if $\boldsymbol{a}_- \geq \mathbf{0}$.

Proof. Notice the fact that the vectors $\boldsymbol{p}_+, \boldsymbol{p}_-$ can be written in terms of $\boldsymbol{x}_+, \boldsymbol{x}_-$ as

$$\boldsymbol{p} = \boldsymbol{e}_0 \otimes \boldsymbol{e}_+ - \frac{\exp(\boldsymbol{x}_+) \otimes \boldsymbol{e}_+ + \mathbf{1}_M \otimes \boldsymbol{e}_-}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_+) + M}, \quad \boldsymbol{q} = \boldsymbol{e}_0 \otimes \boldsymbol{e}_- - \frac{\exp(\boldsymbol{x}_-) \otimes \boldsymbol{e}_- + \mathbf{1}_M \otimes \boldsymbol{e}_+}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_-) + M}.$$

Using (49), for $\boldsymbol{b}_+$ and $\boldsymbol{b}_-$ we have,

$$\boldsymbol{b}_+ = \sqrt{R}\left(\mathbf{1}_{M-1} - \frac{\overline{\boldsymbol{\Phi}}_M^\top \exp(\boldsymbol{x}_+)}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_+) + M}\right), \quad \boldsymbol{b}_- = \mathbf{1}_{M-1} - \frac{\overline{\boldsymbol{\Phi}}_M^\top \exp(\boldsymbol{x}_-)}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_-) + M}.$$

In order to do the same for β and b, we first write the simplified form of $(\overline{\boldsymbol{\Phi}}\boldsymbol{e}_0)^\top \boldsymbol{p}_+$ and $(\overline{\boldsymbol{\Phi}}\boldsymbol{e}_0)^\top \boldsymbol{p}_-$:

$$\begin{aligned} (\overline{\boldsymbol{\Phi}}\boldsymbol{e}_0)^\top \boldsymbol{p} &= \left(\mathbf{1}_M^\top \otimes \begin{pmatrix}1 & -1\end{pmatrix}\right)\boldsymbol{p}_+ = 1 - \frac{\mathbf{1}_M^\top \exp(\boldsymbol{x}_+) - M}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_+) + M} = \frac{2M}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_+) + M}, \\ (\overline{\boldsymbol{\Phi}}\boldsymbol{e}_0)^\top \boldsymbol{p}_- &= \left(\mathbf{1}_M^\top \otimes \begin{pmatrix}1 & -1\end{pmatrix}\right)\boldsymbol{q} = -1 - \frac{-\mathbf{1}_M^\top \exp(\boldsymbol{x}_-) + M}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_-) + M} = \frac{-2M}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_-) + M}. \end{aligned}$$

By substituting the above in the equations for b, β, the desired relations are proved.

Our next step is to prove the nonnegativity of $\boldsymbol{b}_\pm$. It becomes evident after simplifying as

$$\boldsymbol{b}_+ = \sqrt{R}\left(\frac{\begin{bmatrix}\mathbf{0}_{M-1} & \boldsymbol{\Psi}_M\end{bmatrix}^\top \exp(\boldsymbol{x}_+)}{\mathbf{1}_M^\top \exp(\boldsymbol{x}_+) + M}\right) \geq \mathbf{0},$$

since $\boldsymbol{\Psi}_M$ only has entries in $\{0, 2\}^{(M-1)\times(M-1)}$. The same reasoning applies to $\boldsymbol{b}_-$ to show $\boldsymbol{b}_- \geq \mathbf{0}$. Our next step is to prove the auxiliary equations and inequalities. Note that $\boldsymbol{a}_\pm = \mathbf{0}$ would result in $\exp(\boldsymbol{x}_+) = \exp(\gamma/M)\mathbf{1},\ \exp(\boldsymbol{x}_-) = \exp(\delta/M)\mathbf{1}$. In both cases, $\overline{\boldsymbol{\Phi}}_M^\top \mathbf{1} = \mathbf{0}$, and thus

$\boldsymbol{b}_+ = \sqrt{R}\mathbf{1}_{M-1}$, $\boldsymbol{b}_- = \mathbf{1}_{M-1}$. Finally, to prove that $\boldsymbol{b}_- \leq \mathbf{1}_{M-1}$ for $\boldsymbol{a}_- \geq \mathbf{0}$, we proceed to show that $\overline{\boldsymbol{\Phi}}_M^\top \exp(\boldsymbol{x}_-) \geq 0$. Expanding the exponential would yield

$$\exp(\boldsymbol{x}_-) = e^{\delta/M} \exp(\overline{\boldsymbol{\Phi}}_M \boldsymbol{a}_-) = e^{\delta/M} \bigodot_{i=0}^{M-2} \left(\cosh(\frac{a_{-,i}}{M}) + \sinh(\frac{a_{-,i}}{M})\boldsymbol{\Phi}\boldsymbol{e}_{i+1}\right).$$

Since $\boldsymbol{a}_- \geq \mathbf{0}$, all hyperbolic functions are positive. Also, given that $(\boldsymbol{\Phi}\boldsymbol{e}_i) \odot (\boldsymbol{\Phi}\boldsymbol{e}_j) = \boldsymbol{\Phi}\boldsymbol{e}_{i\oplus j}$, we conclude that $\exp(\boldsymbol{x}_-) = e^{\delta/M}(\overline{\boldsymbol{\Phi}}_M \boldsymbol{r}_- + r_0 \mathbf{1}_M)$ for some $\boldsymbol{r}_- \geq \mathbf{0}$. As a result, $\overline{\boldsymbol{\Phi}}_M^\top \exp(\boldsymbol{x}_-) = e^{\delta/M} M \boldsymbol{r}_- \geq \mathbf{0}$. The same argument applies to $\boldsymbol{b}_+$ to show that $\boldsymbol{b}_+ \leq \sqrt{R}\mathbf{1}$ for $\boldsymbol{a}_+ \geq \mathbf{0}$. □

E Proofs for Regularization-path

E.1 Proof of Theorem 2 and basic properties of the solution

We begin with a proof sketch before giving the formal argument. The theorem splits into three claims. The theorem can be understood by splitting in three claims. The first—that *if* $\boldsymbol{Z}_\lambda$ has the form given in the theorem statement, then (Z-UFM$_\lambda$) reduces to the vector problem (7)—is direct. The second, and substantive, claim is that this restriction is *tight*: the global minimizer of (Z-UFM$_\lambda$) lies on the $(\boldsymbol{U}, \widetilde{\boldsymbol{V}})$ subspace $\forall \lambda > 0$. This is where Theorem 1 enters. The optimality condition for (Z-UFM$_\lambda$) is $-\nabla_{\boldsymbol{Z}} \mathcal{L}(\boldsymbol{Z}) = \boldsymbol{Y} - \mathbb{S}(\boldsymbol{Z}) \in \lambda \partial \|\boldsymbol{Z}\|_*$. For any candidate $\boldsymbol{Z} = \boldsymbol{U} \operatorname{diag}_{+1}(\boldsymbol{a}) \widetilde{\boldsymbol{V}}^\top$, Theorem 1 rewrites the loss gradient as $\boldsymbol{U} \operatorname{diag}_{+1}(\mathcal{F}(\boldsymbol{a})) \widetilde{\boldsymbol{V}}^\top$, so the condition becomes $\boldsymbol{U} \operatorname{diag}_{+1}(\mathcal{F}(\boldsymbol{a})) \widetilde{\boldsymbol{V}}^\top \in \lambda \partial \|\boldsymbol{U} \operatorname{diag}_{+1}(\boldsymbol{a}) \widetilde{\boldsymbol{V}}^\top\|_*$. By orthogonality of $\boldsymbol{U}, \widetilde{\boldsymbol{V}}$, this is equivalent to $\mathcal{F}(\boldsymbol{a}) \in \lambda \partial \|\boldsymbol{a}\|_{+1}$, where $\|\cdot\|_{+1}$ denotes the regularizer in (7). Since this is precisely the optimality condition for (7), each of its solutions certifies a solution of (Z-UFM$_\lambda$). Uniqueness on both sides (due to strict convexity of the CE loss on the subspace orthogonal to $\mathbf{1}$) closes the loop. Finally, the third claim, the reduction to four variables follows by symmetry: By leveraging the structure of the $\boldsymbol{U}, \widetilde{\boldsymbol{V}}$ subspace with respect to the Sylvester-Hadamard matrix, we show that the objective in (7) is invariant to swapping the coordinates of $\boldsymbol{a}_\pm$. Because the minimizer is unique, this symmetry forces all elements within these vectors to be identical.

We now formalize each claim in turn. After that, at the end of this section, we further derive basic but important properties of the solution, which we use in proofs of later stages.

Our first step is to characterize the regularization path solution, while assuming it is constrained to the Hadamard factorization discussed in previous sections.

Theorem 12 (KKT conditions for regularized CE in the diagonal basis)**.** *Consider the optimization problem*

$$\min_{\alpha \in \mathbb{R},\ \boldsymbol{A} \in \mathbb{R}^{(k-1)\times(k-1)}} \operatorname{CE}(\boldsymbol{L}) + \lambda \|\boldsymbol{L}\|_*, \qquad \text{s.t. } \boldsymbol{L} = \boldsymbol{U} \begin{bmatrix} \alpha \boldsymbol{e}_0 & \boldsymbol{A} \end{bmatrix} \widetilde{\boldsymbol{V}}^\top, \tag{50}$$

where $\lambda > 0$, the matrix $\boldsymbol{A}$ is restricted to the diagonal form $\boldsymbol{A} = \operatorname{diag}(a, \boldsymbol{a}_+ \otimes \boldsymbol{e}_+ + \boldsymbol{a}_- \otimes \boldsymbol{e}_-)$. Then, the unique minimizer of (50) satisfies

$$\boldsymbol{Y} - \mathbb{S}(\boldsymbol{L}) = \boldsymbol{U} \begin{bmatrix} \beta \boldsymbol{e}_0 & \boldsymbol{B} \end{bmatrix} \widetilde{\boldsymbol{V}}^\top, \qquad \boldsymbol{B} = \operatorname{diag}(b, \boldsymbol{b}_+ \otimes \boldsymbol{e}_+ + \boldsymbol{b}_- \otimes \boldsymbol{e}_-), \tag{51}$$

where

$$(\beta, b) \in \lambda \partial \sqrt{\alpha^2 + a^2}, \tag{52}$$

$$\boldsymbol{b}_\pm \in \lambda \partial \|\boldsymbol{a}_\pm\|_1, \tag{53}$$

Explicitly, this yields

$$(\beta, b) = \lambda \frac{(\alpha, a)}{\sqrt{\alpha^2 + a^2}} \ \text{ if } (\alpha, a) \neq (0, 0), \quad \beta^2 + b^2 \leq \lambda^2 \text{ otherwise}, \tag{54}$$

$$\boldsymbol{b}_\pm[j] = \lambda \operatorname{sign}(\boldsymbol{a}_\pm[j]) \ \text{ if } \boldsymbol{a}_\pm[j] \neq 0, \quad |\boldsymbol{b}_\pm[j]| \leq \lambda \text{ otherwise}. \tag{55}$$

Proof. Theorem 11 guarantees that (51) for any $\alpha, \boldsymbol{A}$, without imposing any constraints on $\beta, \boldsymbol{B}$ yet. Denote the linear mapping $\mathcal{T}(\boldsymbol{X}) = \boldsymbol{U}\boldsymbol{X}\widetilde{\boldsymbol{V}}^\top$ and its adjoint map as $\mathcal{T}^*(\boldsymbol{X}) = \boldsymbol{U}^\top \boldsymbol{X} \widetilde{\boldsymbol{V}}$. Since

$\boldsymbol{U}, \widetilde{\boldsymbol{V}}$ are fixed, the objective in (50) is convex in the free parameters $(\alpha, \boldsymbol{A})$, and the KKT condition reduces to first-order subgradient optimality:

$$\mathbf{0} \in \nabla_{\alpha,\boldsymbol{A}} \operatorname{CE}(\mathcal{T}([\alpha \boldsymbol{e}_0 \quad \boldsymbol{A}])) + \lambda\, \partial_{\alpha,\boldsymbol{A}} \|\mathcal{T}([\alpha \boldsymbol{e}_0 \quad \boldsymbol{A}])\|_*. \tag{56}$$

The nuclear norm in the above expression is invariant with respect to the transformation $\mathcal{T}$, and using the subgradient chain rule for affine transformations (Beck, 2017, Theorem 3.43 (b)), the above expression can be simplified to

$$\boldsymbol{U}^\top (\boldsymbol{Y} - \mathbb{S}\left(\mathcal{T}([\alpha \boldsymbol{e}_0 \quad \boldsymbol{A}])\right) \widetilde{\boldsymbol{V}} = [\beta \boldsymbol{e}_0 \quad \boldsymbol{B}] \in \lambda\, \partial \left\|[\alpha \boldsymbol{e}_0 \quad \boldsymbol{A}]\right\|_*. \tag{57}$$

Given the diagonal structure of $\boldsymbol{A}$, the nuclear norm expression simplifies as

$$\left\|[\alpha \boldsymbol{e}_0 \quad \boldsymbol{A}]\right\|_* = \sqrt{\alpha^2 + a^2} + \|\boldsymbol{a}_+\|_1 + \|\boldsymbol{a}_-\|_1, \tag{58}$$

Applying the subgradient formula to (57) yields

$$(\beta, b) \in \lambda\, \partial \sqrt{\alpha^2 + a^2}, \qquad \boldsymbol{b}_\pm \in \lambda \partial \|\boldsymbol{a}_\pm\|_1,$$

which are exactly (52) and (53). Uniqueness follows from the convexity of the nuclear norm and the strict convexity of the cross-entropy over the subspace of logit matrices with zero-sum columns, a constraint that our factorization automatically satisfies. □

Our next step is to show that removing the factorization constraint $\boldsymbol{L} = \boldsymbol{U}\,[\alpha \boldsymbol{e}_0 \quad \boldsymbol{A}]\,\widetilde{\boldsymbol{V}}^\top$ does not change the solution via a KKT certificate proof, and thus the solution of (Z-UFM$_\lambda$) admits the Hadamard decomposition. Before proceeding with the proof, two utility lemmas are introduced:

Lemma 8. *Let $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times s}$ be matrices with orthonormal columns, i.e., they satisfy*

$$\boldsymbol{U}^\top \boldsymbol{U} = I_r, \qquad \boldsymbol{V}^\top \boldsymbol{V} = I_s.$$

Then, for any $\boldsymbol{C} \in \mathbb{R}^{r\times s}$ and any $\boldsymbol{D} \in \lambda\, \partial \|\boldsymbol{C}\|_*$,

$$\|\boldsymbol{U}\boldsymbol{C}\boldsymbol{V}^\top\|_* = \|\boldsymbol{C}\|_*, \qquad \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \in \lambda\, \partial \|\boldsymbol{U}\boldsymbol{C}\boldsymbol{V}^\top\|_*.$$

Proof. Let $\boldsymbol{D} \in \lambda\, \partial \|\boldsymbol{C}\|_*$. By the definition of the subdifferential, for every $\boldsymbol{Z} \in \mathbb{R}^{r\times s}$,

$$\lambda \|\boldsymbol{Z}\|_* \geq \lambda \|\boldsymbol{C}\|_* + \langle \boldsymbol{D}, \boldsymbol{Z} - \boldsymbol{C} \rangle.$$

Fix any $\mathbf{Y} \in \mathbb{R}^{m\times n}$, and set

$$\boldsymbol{Z} = \boldsymbol{U}^\top \mathbf{Y} \boldsymbol{V}.$$

Using the fact that multiplication by matrices with orthonormal columns does not increase the nuclear norm (see (Bhatia, 2013, Proposition IV.2.4)),

$$\|\mathbf{Y}\|_* \geq \|\boldsymbol{U}^\top \mathbf{Y} \boldsymbol{V}\|_*,$$

we obtain

$$\lambda \|\mathbf{Y}\|_* \geq \lambda \|\boldsymbol{U}^\top \mathbf{Y} \boldsymbol{V}\|_* \geq \lambda \|\boldsymbol{C}\|_* + \langle \boldsymbol{D}, \boldsymbol{U}^\top \mathbf{Y} \boldsymbol{V} - \boldsymbol{C} \rangle$$

Since $\boldsymbol{U}$ and $\boldsymbol{V}$ have orthonormal columns, the matrices $\boldsymbol{U}\boldsymbol{C}\boldsymbol{V}^\top$ and $\boldsymbol{C}$ have the same nonzero singular values. Therefore,

$$\|\boldsymbol{U}\boldsymbol{C}\boldsymbol{V}^\top\|_* = \|\boldsymbol{C}\|_*,$$

and hence,

$$\lambda \|\mathbf{Y}\|_* \geq \lambda \|\boldsymbol{U}\boldsymbol{C}\boldsymbol{V}^\top\|_* + \langle \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top, \mathbf{Y} - \boldsymbol{U}\boldsymbol{C}\boldsymbol{V}^\top \rangle.$$

Since this holds for every $\mathbf{Y}$, it follows that

$$\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top \in \lambda\, \partial \|\boldsymbol{U}\boldsymbol{C}\boldsymbol{V}^\top\|_*.$$

□

Lemma 9. *Let $\boldsymbol{L} \in \mathbb{R}^{k\times n}$ be a matrix satisfying $\boldsymbol{L}^\top \mathbf{1}_k = \mathbf{0}$. Then, $\boldsymbol{c} = \mathbf{0}$ is the unique solution to*

$$\min_{\boldsymbol{c}\in\mathbb{R}^n} \|\boldsymbol{L} + \mathbf{1}_k \boldsymbol{c}^\top\|_*.$$

Proof. The nuclear norm expression can be simplified using $\boldsymbol{L}^\top \mathbf{1} = \mathbf{0}$ to reach

$$\|\boldsymbol{L} + \mathbf{1}_k \boldsymbol{c}^\top\|_* = \operatorname{tr}\sqrt{\boldsymbol{L}^\top \boldsymbol{L} + k\boldsymbol{c}\boldsymbol{c}^\top} \geq \operatorname{tr}\sqrt{\boldsymbol{L}^\top \boldsymbol{L}} = \|\boldsymbol{L}\|_*,$$

which establishes $\boldsymbol{c} = \mathbf{0}$ as a minimizer. Next, Weyl's monotonicity theorem (Bhatia, 2013, Corollary 3.2.3) results in $\lambda_i(\boldsymbol{L}^\top \boldsymbol{L} + k\boldsymbol{c}\boldsymbol{c}^\top) \geq \lambda_i(\boldsymbol{L}^\top \boldsymbol{L})$, where λ_i denotes the i-th eigenvalue of each matrix. This, alongside the fact that $\operatorname{tr}(\boldsymbol{L}^\top \boldsymbol{L} + k\boldsymbol{c}\boldsymbol{c}^\top) > \operatorname{tr}(\boldsymbol{L}^\top \boldsymbol{L})$ for nonzero $\boldsymbol{c}$, results in at least one eigenvalue to strictly increase. Consequently, for $\boldsymbol{c} \neq \mathbf{0}$

$$\|\boldsymbol{L} + \mathbf{1}_k \boldsymbol{c}^\top\|_* = \operatorname{tr}\sqrt{\boldsymbol{L}^\top \boldsymbol{L} + k\boldsymbol{c}\boldsymbol{c}^\top} > \operatorname{tr}\sqrt{\boldsymbol{L}^\top \boldsymbol{L}} = \|\boldsymbol{L}\|_*,$$

which proves the uniqueness. □

Proposition 4 (KKT certificate for the nuclear-norm-regularized CE problem)**.** *Let*

$$\boldsymbol{L}^\star = \boldsymbol{U}\begin{bmatrix}\alpha \boldsymbol{e}_0 & \boldsymbol{A}\end{bmatrix}\widetilde{\boldsymbol{V}}^\top, \qquad \boldsymbol{A} = \operatorname{diag}\left(a, \boldsymbol{a}_+ \otimes \boldsymbol{e}_+ + \boldsymbol{a}_- \otimes \boldsymbol{e}_-\right),$$

and

$$\boldsymbol{Y} - \mathbb{S}\left(\boldsymbol{L}^\star\right) = \boldsymbol{U}\begin{bmatrix}\beta \boldsymbol{e}_0 & \boldsymbol{B}\end{bmatrix}\widetilde{\boldsymbol{V}}^\top, \qquad \boldsymbol{B} = \operatorname{diag}\left(b, \boldsymbol{b}_+ \otimes \boldsymbol{e}_+ + \boldsymbol{b}_- \otimes \boldsymbol{e}_-\right),$$

be the decomposition given by Theorem 11. If

$$(\beta, b) \in \lambda\, \partial\sqrt{\alpha^2 + a^2}, \tag{59}$$

$$\boldsymbol{b}_\pm[j] \in \lambda\, \partial|\boldsymbol{a}_\pm[j]|, \qquad j \in [k/2] \tag{60}$$

then $\boldsymbol{L}^\star$ is the unique minimizer of the convex optimization problem

$$\min_{\boldsymbol{L}\in\mathbb{R}^{k\times n}} \operatorname{CE}(\boldsymbol{L}) + \lambda\|\boldsymbol{L}\|_*.$$

Proof. Consider the objective

$$f(\boldsymbol{L}) \coloneqq \operatorname{CE}(\boldsymbol{L}) + \lambda\|\boldsymbol{L}\|_*.$$

Notice that the cross-entropy term is invariant to centering each column; that is, $\operatorname{CE}(\boldsymbol{L}) = \operatorname{CE}(\widehat{\boldsymbol{L}})$, where $\widehat{\boldsymbol{L}}$ is the $\boldsymbol{L}$ with all columns centered around zero. Moreover, the strict convexity of cross-entropy over this subspace alongside the overall objective convexity requires that $\widehat{\boldsymbol{L}}_1 = \widehat{\boldsymbol{L}}_2$ for any two solutions $\boldsymbol{L}_1, \boldsymbol{L}_2$.

Therefore, all solutions are of the form $\boldsymbol{L} = \boldsymbol{L}^* + \mathbf{1}_k \boldsymbol{r}^\top$ for a unique $\widehat{\boldsymbol{L}}$ that satisfies $\widehat{\boldsymbol{L}}^\top \mathbf{1}_k = \mathbf{0}$, and all possible $\boldsymbol{r}$ are determined from $\min_{\boldsymbol{r}\in\mathbb{R}^n} \|\widehat{\boldsymbol{L}} + \mathbf{1}_k \boldsymbol{r}^\top\|_*$. This minimization is shown in Lemma 9 to have $\boldsymbol{r} = \mathbf{0}$ as its only solution, which proves the solution is unique and satisfies $\boldsymbol{L}^\top \mathbf{1} = \mathbf{0}$. The proposed solution automatically satisfies this requirement; therefore, it suffices to verify the first-order optimality condition

$$\mathbf{0} \in \nabla \operatorname{CE}(\boldsymbol{L}^\star) + \lambda\, \partial\|\boldsymbol{L}^\star\|_* = \mathbb{S}\left(\boldsymbol{L}^\star\right) - \boldsymbol{Y} + \lambda\, \partial\|\boldsymbol{L}^\star\|_*$$

Define

$$\boldsymbol{C}^\star \coloneqq \begin{bmatrix}\alpha \boldsymbol{e}_0 & \boldsymbol{A}\end{bmatrix}, \qquad \boldsymbol{D}^\star \coloneqq \begin{bmatrix}\beta \boldsymbol{e}_0 & \boldsymbol{B}\end{bmatrix}.$$

Then, by definition in the statement of the proposition,

$$\boldsymbol{L}^\star = \boldsymbol{U}\boldsymbol{C}^\star\widetilde{\boldsymbol{V}}^\top, \qquad \boldsymbol{Y} - \mathbb{S}\left(\boldsymbol{L}^\star\right) = \boldsymbol{U}\boldsymbol{D}^\star\widetilde{\boldsymbol{V}}^\top, \quad \boldsymbol{D}^\star \in \lambda\, \partial\|\boldsymbol{C}^\star\|_*,$$

Since $\boldsymbol{U}$ and $\widetilde{\boldsymbol{V}}$ have orthonormal columns, the nuclear norm is invariant under multiplication by these matrices, and its subdifferential transforms equivariantly. Hence, using Lemma 8, we obtain

$$\boldsymbol{D}^\star \in \lambda\, \partial\|\boldsymbol{C}^\star\|_* \quad \Longrightarrow \quad \boldsymbol{U}\boldsymbol{D}^\star\widetilde{\boldsymbol{V}}^\top \in \lambda\, \partial\|\boldsymbol{U}\boldsymbol{C}^\star\widetilde{\boldsymbol{V}}^\top\|_*.$$

Thus $\boldsymbol{L}^\star$ satisfies the KKT condition for the convex problem

$$\min_{\boldsymbol{L}\in\mathbb{R}^{k\times n}} \operatorname{CE}(\boldsymbol{L}) + \lambda\|\boldsymbol{L}\|_*,$$

□

Now that the Hadamard factorization for the regularized problem (Z-UFM$_\lambda$) has been proven, basic properties of the solution are discussed. These properties include sparsity patterns and useful nonnegativity results for variables in the optimal solution.

Lemma 10 (Basic properties of regulaization path solution)**.** *The optimal solution to* (50) *has the following properties:*

1. *The following quantities are nonnegative:* $a_\lambda \geq 0, \boldsymbol{a}_+ \geq \mathbf{0}, \boldsymbol{a}_- \geq \mathbf{0}$*, and* $\gamma_\lambda, \delta_\lambda \geq 0$*.*
2. $\boldsymbol{a}_{\lambda,+} = \mathbf{0} \iff \lambda \geq \sqrt{R}$ *and* $\boldsymbol{a}_{\lambda,-} = \mathbf{0} \iff \lambda \geq 1$*.*
3. *A threshold* $\lambda_{K,R} \leq \sqrt{\frac{R+1}{2}}$ *exists such that* $(\alpha_\lambda, a_\lambda) = (0,0)$ *for* $\lambda \geq \lambda_{K,R}$*.*

Proof. We prove the statements in the order in which they appear.

1. Notice that $b \geq 0, \boldsymbol{b}_\pm \geq \mathbf{0}$ regardless of the problem variables, and $b + \frac{\beta}{\sqrt{R}} = \sqrt{\frac{R+1}{2}} X_+ > 0$ and $b - \beta\sqrt{R} = \sqrt{\frac{R+1}{2}} X_- > 0$. The KKT condition for (α, a) is either $(\alpha, a) = (0,0)$ or $(\beta, b) = \frac{\lambda}{\sqrt{a^2+\alpha^2}}(\alpha, a)$. It follows that in either case, $a \geq 0$, and the definition of (γ, δ) in (6) implies $\gamma, \delta \geq 0$.

   Likewise, the conditions for $\boldsymbol{a}_\pm$ state either $\boldsymbol{b}_\pm[j] = \lambda \operatorname{sign}(\boldsymbol{a}_\pm[j])$ or $\boldsymbol{a}_\pm[j] = 0 \; \forall j \in [M]$. Since $\boldsymbol{b}_\pm \geq \mathbf{0}$, this implies $\boldsymbol{a}_\pm \geq 0$.

2. We shall prove the second statement by contradiction. Note that $\boldsymbol{a}_\pm \geq \mathbf{0}$ yields $\boldsymbol{b}_+ \leq \sqrt{R}\mathbf{1}$, $\boldsymbol{b}_- \leq \mathbf{1}$. Suppose for $\lambda > \sqrt{R}$, there exists $j \in [M]$ such that $\boldsymbol{a}_+[j] \neq 0$. The KKT conditions would give $\lambda = \boldsymbol{b}_+[j] \leq \sqrt{R}$, which leads to a contradiction. To prove the other direction, suppose that $\boldsymbol{a}_+ = \mathbf{0}$ for some $\lambda < \sqrt{R}$, which leads to $\boldsymbol{b}_+ = \sqrt{R}\mathbf{1}$. However, the KKT conditions would require $\boldsymbol{b}_+ \leq \lambda\mathbf{1}$, leading to a contradiction. An analogous argument proves that $\boldsymbol{a}_- = \mathbf{0} \iff \lambda \leq 1$.

3. Based on the KKT equations for (α, a) either $(\alpha, a) = 0, \beta^2 + b^2 \leq \lambda^2$ or $\beta^2 + b^2 = \lambda^2$, where $\beta^2 + b^2 = \frac{1}{2}(RX_+^2 + X_-^2)$. Jensen's inequality on the strictly convex exponential function gives $\mathbf{1}^\top \exp(\boldsymbol{x}_+) > Me^\gamma$, and likewise $\mathbf{1}^\top \exp(\boldsymbol{x}_-) > Me^\delta$. Since $\gamma, \delta \geq 0$, the definition of $X_\pm$ implies $X_\pm < 1 \Rightarrow \beta^2 + b^2 < \frac{R+1}{2}$. This in turn establishes $\beta^2 + b^2 = \lambda^2$ is not possible for $\lambda \geq \sqrt{\frac{R+1}{2}}$.

□

E.1.1 Reduction to four variables

We prove that the vectorized regularization path reduces to an optimization over four variables.

Proposition 5 (Reduction to four variables)**.** *There exist scalars* $a_{+,\lambda}$ *and* $a_{-,\lambda}$ *such that the minimizer of* (7) *satisfies* $\boldsymbol{a}_{\pm,\lambda} = a_{\pm,\lambda}\mathbf{1}_{M-1}$*.*

Proof. Suppose $(\alpha_\lambda, a_\lambda, \boldsymbol{a}_{+,\lambda}, \boldsymbol{a}_{-,\lambda})$ is a minimizer. Let $r, s \in [M-1]$. By Proposition 6 (see also 3), there exist $\boldsymbol{P}, \boldsymbol{C}$ such that $\boldsymbol{P}\overline{\boldsymbol{\Phi}}_M = \overline{\boldsymbol{\Phi}}_M\boldsymbol{C}$ and $\boldsymbol{C}\boldsymbol{e}_r = \boldsymbol{e}_s$, and, $(\alpha_\lambda, a_\lambda, \boldsymbol{C}\boldsymbol{a}_{+,\lambda}, \boldsymbol{C}\boldsymbol{a}_{-,\lambda})$ is also a minimizer. By uniqueness of the minimizer of (7) it must be that

$$\boldsymbol{C}\boldsymbol{a}_{\pm,\lambda} = \boldsymbol{a}_{\pm,\lambda} \qquad \Longrightarrow \qquad (\boldsymbol{a}_{\pm,\lambda})_s = (\boldsymbol{a}_{\pm,\lambda})_r \,.$$

Since r, s were arbitrary, all coordinates of $\boldsymbol{a}_{\pm,\lambda}$ are equal completing the proof □

The key observation behind the proof is showing invariance of the optimization objective under admissible Hadamard permutations.

Proposition 6 (Reduced-objective invariance). *Let $\boldsymbol{C} \in \mathbb{R}^{(M-1)\times(M-1)}$ be a permutation matrix for which there exists a permutation matrix $\boldsymbol{P} \in \mathbb{R}^{M\times M}$ satisfying*

$$\boldsymbol{P}\overline{\boldsymbol{\Phi}}_M = \overline{\boldsymbol{\Phi}}_M \boldsymbol{C}. \tag{61}$$

For example, this is guaranteed to exist by Lemma 3. Then the vectorized objective in (7)

$$f_\lambda(\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-) \coloneqq \mathcal{L}\Big(\boldsymbol{U}\mathrm{diag}_{+1}\left(\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-\right)\widetilde{\boldsymbol{V}}^\top\Big) + \lambda\left(\sqrt{\alpha^2+a^2} + \|\boldsymbol{a}_+\|_1 + \|\boldsymbol{a}_-\|_1\right)$$

satisfies

$$f_\lambda(\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-) = f_\lambda(\alpha, a, \boldsymbol{C}\boldsymbol{a}_+, \boldsymbol{C}\boldsymbol{a}_-).$$

Proof. Recall that the cross-entropy loss $\mathcal{L}(\boldsymbol{Z}) = -\langle \boldsymbol{Y}, \log \mathbb{S}(\boldsymbol{Z})\rangle$, where $\langle\cdot,\cdot\rangle$ denotes the Frobenius inner product between matrices and the log function is applied element-wise. Statement 1 of Theorem 11 only requires the structure of

$$\boldsymbol{A} = \mathrm{diag}\left(a, \boldsymbol{a}_+ \otimes \boldsymbol{e}_+ + \boldsymbol{a}_- \otimes \boldsymbol{e}_-\right) \in \mathbb{R}^{(K-1)\times(K-1)}, \quad \boldsymbol{Q}' \coloneqq \mathbb{S}\left(\boldsymbol{U}\begin{bmatrix}\alpha\boldsymbol{e}_0 & \boldsymbol{A}\end{bmatrix}\tilde{\boldsymbol{V}}^\top\right) \in \mathbb{R}^{K\times(M)(R+1)}.$$

instead of the full matrix $\boldsymbol{Q}$. Consequently, the block columns $\boldsymbol{Q}'_i$, defined in similar fashion to Theorem 11 satisfy

$$\boldsymbol{Q}'_i = (\Pi_i \otimes \boldsymbol{I}_2)\begin{bmatrix}\boldsymbol{p}_+\mathbf{1}_R^\top & \boldsymbol{p}_-\end{bmatrix} \in \mathbb{R}^{K\times(R+1)},\ i \in [M], \qquad \boldsymbol{p}'_+ = \mathbb{S}\left(\boldsymbol{x}_+ \otimes \boldsymbol{f}_+\right),\ \boldsymbol{p}'_- = \mathbb{S}\left(\boldsymbol{x}_- \otimes \boldsymbol{f}_-\right),$$

and the loss can be simplified to

$$\begin{aligned}
\mathcal{L}(\boldsymbol{Q}') &= -\langle \boldsymbol{Y}, \log(\boldsymbol{Q})'\rangle = -\sum_{m\in[M]} \langle \widetilde{\boldsymbol{\Pi}}_m \begin{bmatrix}\boldsymbol{e}_0\mathbf{1}_R^\top & \boldsymbol{e}_1\end{bmatrix}, \widetilde{\boldsymbol{\Pi}}_m \begin{bmatrix}\log(\boldsymbol{p}'_+)\mathbf{1}_R^\top & \log(\boldsymbol{p}'_-)\end{bmatrix}\rangle \\
&= -M\langle \begin{bmatrix}\boldsymbol{e}_0\mathbf{1}_R^\top & \boldsymbol{e}_1\end{bmatrix}, \begin{bmatrix}\log(\boldsymbol{p}'_+)\mathbf{1}_R^\top & \log(\boldsymbol{p}'_-)\end{bmatrix}\rangle = -M(R\boldsymbol{e}_0^\top \log(\boldsymbol{p}'_+) + \boldsymbol{e}_1^\top \log(\boldsymbol{p}'_-)) \\
&= -M\left(R\left(\boldsymbol{e}_0^\top \boldsymbol{x}_+ - \log(\mathbf{1}_M^\top \exp(\boldsymbol{x}_+) + M)\right) + \boldsymbol{e}_1^\top \boldsymbol{x}_- - \log(\mathbf{1}_M^\top \exp(\boldsymbol{x}_-) + M)\right).
\end{aligned}$$

Substituting the expressions for $\boldsymbol{x}_\pm$ and noting that $\boldsymbol{e}_0^\top \overline{\boldsymbol{\Phi}}_M = \mathbf{1}_{M-1}^\top$ further simplifies the loss to

$$\begin{aligned}
\mathcal{L}(\boldsymbol{Q}') =& R\left(M\log(\mathbf{1}_M^\top \exp(\boldsymbol{x}_+) + M) - (\gamma + \frac{1}{\sqrt{R}}\mathbf{1}_{M-1}^\top \boldsymbol{a}_+)\right) \\
&+ \left(M\log(\mathbf{1}_M^\top \exp(\boldsymbol{x}_-) + M) - (\delta + \mathbf{1}_{M-1}^\top \boldsymbol{a}_-)\right).
\end{aligned}$$

The terms $\mathbf{1}_{M-1}^\top \boldsymbol{a}_\pm$ in the cross-entropy and $\|\boldsymbol{a}_\pm\|_1$ in the regularizer are permutation-invariant, and the terms $\mathbf{1}_M^\top \exp(\boldsymbol{x}_\pm)$ remain unchanged given that the permutation matrix $\boldsymbol{C}$ maps $\boldsymbol{x}_\pm$ to $\boldsymbol{P}\boldsymbol{x}_\pm$. As a result, the loss function is invariant to any matrix $\boldsymbol{C}$ with the property at the beginning of the theorem statement.

□

E.2 Proof of proposition 1: Analytic solution

In this subsection we solve the four-dimensional convex problem from (7) explicitly. The argument has three steps. First, we isolate the scalar relations that encode the active KKT system. Second, we show that these scalar equations admit a unique solution in the relevant λ-ranges. Third, we substitute these scalar solutions back into the KKT conditions and verify the phase-wise formulas stated in the main theorem. This organization separates the existence of the scalar roots from the algebra needed to recover the optimization variables.

First, we prove a utility lemma, which shows that whenever $\boldsymbol{b}_\pm \propto \mathbf{1}_{M-1}$, $\boldsymbol{a}_\pm \propto \mathbf{1}_{M-1}$, and determine the multiplication coefficient as well:

Lemma 11. *Let $0 \le t \le 1$, $\gamma \in \mathbb{R}$, and suppose $\boldsymbol{r} \in \mathbb{R}^M$ satisfies*

$$\boldsymbol{r} = \exp\left(\frac{1}{M}(\gamma\mathbf{1}_M + \overline{\boldsymbol{\Phi}}_M\boldsymbol{\eta})\right), \qquad \overline{\boldsymbol{\Phi}}_M^\top \boldsymbol{r} = t\big(M + \mathbf{1}_M^\top \boldsymbol{r}\big)\mathbf{1}_{M-1}, \tag{62}$$

Then, the solution to this equation satisfies

$$\boldsymbol{\eta} = (\gamma - M\log(s))\,\mathbf{1}_{M-1}, \quad s^{M-1}((1+\theta)s + \theta) = e^\gamma, \quad \theta \coloneqq \frac{Mt}{1-t}, \tag{63}$$

where $0 < s \le \exp(\gamma/M)$ is the unique solution of the scalar equation in (63)

Proof. The first step is to multiply (62) by $\overline{\mathbf{\Phi}}_M$ given that $\overline{\mathbf{\Phi}}_M\overline{\mathbf{\Phi}}_M^\top = M\boldsymbol{I}_M - \mathbf{1}_M\mathbf{1}_M^\top$ to reach

$$(M\boldsymbol{I}_M - \mathbf{1}_M\mathbf{1}_M^\top)\boldsymbol{r} = t\big(M + \mathbf{1}_M^\top\boldsymbol{r}\big)(\overline{\mathbf{\Phi}}_M\mathbf{1}_{M-1})$$
$$\Rightarrow \boldsymbol{r} = \frac{\mathbf{1}_M^\top\boldsymbol{r}}{M}\mathbf{1}_M + t(1+\frac{\mathbf{1}_M^\top\boldsymbol{r}}{M})(M\boldsymbol{e}_0 - \mathbf{1}_M)$$

After defining $\kappa \coloneqq \frac{\mathbf{1}_M^\top\boldsymbol{r}}{M}$ and using that $\mathbf{1}_M^\top\log(\boldsymbol{r}) = \gamma$, $\overline{\mathbf{\Phi}}_M^\top\log(\boldsymbol{r}) = \boldsymbol{\eta}$, standard algebra leads to

$$\log(\boldsymbol{r}) = \log(\frac{\kappa + t(1+\kappa)(M-1)}{\kappa - t(1+\kappa)})\boldsymbol{e}_0 + \log(\kappa - t(1+\kappa))\mathbf{1}_M, \tag{64}$$

$$\boldsymbol{\eta} = \overline{\mathbf{\Phi}}_M^\top\log(\boldsymbol{r}) = \log(\frac{\kappa + t(1+\kappa)(M-1)}{\kappa - t(1+\kappa)})\mathbf{1}_M, \tag{65}$$

$$\gamma = \mathbf{1}_M^\top\log(\boldsymbol{r}) = \log\Big(\big(\kappa + t(1+\kappa)(M-1)\big)\big(\kappa - t(1+\kappa)\big)^{M-1}\Big), \tag{66}$$

after which the desired result can be reached by defining $s \coloneqq \kappa - t(1+\kappa)$ and rewriting the above equations in terms of s. Note that the uniqueness of s for a given γ is the result the monotonicity of $s^{M-1}((1+\theta)s+\theta)$ for $s > 0$, and $s \le e^{\gamma/M}$ follows by $e^\gamma = s^{M-1}((1+\theta)s+\theta) \ge s^M$. □

Auxiliary scalar functions. For $\lambda > 0$, define the following variables and functions:

$$\Psi_\lambda(x) \coloneqq \left(\frac{M}{\lambda(\theta_{-,\lambda}+M)}\right)\sqrt{\frac{2(1+x)^2}{(1+x)^2-2}} - 1, \quad x > \sqrt{2}-1 \tag{67a}$$

$$H_\lambda(x) \coloneqq Q(\Psi_\lambda(x), \theta_{-,\lambda}) - \left(\frac{M\sqrt{R}}{\lambda(\theta_{-,\lambda}+M)}\right)\frac{1+x}{1+\Psi_\lambda(x)}Q(x,\theta_{+,\lambda}) \tag{67b}$$

The role of Ψ_λ is to express the minority-side scalar in terms of the majority-side scalar once the first KKT relation is imposed, while H_λ packages the remaining compatibility condition. In the mixed and fully active phases, solving the regularization path therefore reduces to finding a root of H_λ and then setting $r_\lambda = \Psi_\lambda(s_\lambda)$.

The following lemma determines that the equation governing $\lambda_{K,R}$, as will be derived in the main theorem, has a unique solution.

Lemma 12. *For any $M > 1$ and $R > 1$, there exists a unique $\lambda_{K,R} \in [1, \sqrt{\frac{R+1}{2}}]$ such that*

$$Q\left(\frac{2\lambda_{K,R}}{\sqrt{2\lambda_{K,R}^2 - 1}} - 1, \theta_{+,\lambda_{K,R}}\right) = 0.$$

Proof. Let $\xi_\lambda \coloneqq \frac{2\lambda}{\sqrt{2\lambda^2-1}} - 1$ and $\theta_\lambda \coloneqq \theta_{+,\lambda}$. Then, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda}Q(\xi_\lambda,\theta_\lambda) &= \frac{\partial Q}{\partial \xi_\lambda}\frac{\mathrm{d}\xi_\lambda}{\mathrm{d}\lambda} + \frac{\partial Q}{\partial\theta_\lambda}\frac{\mathrm{d}\theta_\lambda}{\mathrm{d}\lambda} \\ &= \frac{\partial Q}{\partial\xi_\lambda}\frac{-2}{(2x^2-1)^{3/2}} + \frac{\partial Q}{\partial\theta_\lambda}\frac{-m\sqrt{R}}{\lambda^2} < 0,\end{aligned} \tag{68}$$

where we have used the fact that $\frac{\partial Q}{\partial\theta_\lambda} > 0$ and $\frac{\partial Q}{\partial\xi_\lambda} > 0$. Therefore, the function $Q(\xi_\lambda,\theta_\lambda)$ is strictly decreasing in λ. Moreover,

$$Q(\xi_1,\theta_1) = Q(1,\theta_1) = \log(1+2\theta_1) > 0,$$

and at $\lambda^* = \sqrt{\frac{R+1}{2}}$ the relation $\theta_{\lambda^*} = M\frac{1-\xi_{\lambda^*}}{1+\xi_{\lambda^*}}$ would result in

$$Q(\xi_{\lambda^*},\theta_{\lambda^*}) = \log(\xi_{\lambda^*}^{M-1}M - (M-1)\xi_{\lambda^*}^M) < \log(1) = 0,$$

which implies the existence of a unique root in $[1, \sqrt{\frac{R+1}{2}}]$. □

The previous lemma identifies the transition point between phases (ii) and (iii). Below this threshold, the mixed mode becomes active, and one must solve a genuinely coupled scalar system. The next two propositions show that this system is well posed in the two remaining nontrivial regimes and record the positivity properties needed later when reconstructing the active coordinates.

Proposition 7. *Let $\lambda_{K,R} \in [1, \sqrt{\frac{R+1}{2}}]$ be the point defined in Lemma 12. Then, for any $\lambda \in [1, \lambda_{K,R})$, the function $H_\lambda(\cdot)$ has a root $s_\lambda \in (\sqrt{2}-1, \frac{2\lambda}{\sqrt{2\lambda^2-1}} - 1)$. Furthermore, if $r_\lambda = \Psi_\lambda(s_\lambda)$, then*

$$0 < Q(s_\lambda, \theta_{+,\lambda}), \quad 0 < Q(r_\lambda, \theta_{-,\lambda}), \quad 0 < s_\lambda \exp\left(\frac{-Q(s_\lambda, \theta_{+,\lambda})}{M}\right) < 1\,.$$

Proof. Let $\xi_\lambda := \frac{2\lambda}{\sqrt{2\lambda^2-1}} - 1$. We apply the intermediate value theorem by establishing that $H_\lambda(x) \to +\infty$ as $x \downarrow (\sqrt{2}-1)$ and $H_\lambda(\xi_\lambda) < 0$.

For the first claim, note that $\Psi_\lambda(x) \to +\infty$ as $x \downarrow (\sqrt{2}-1)$, which implies $Q(\Psi_\lambda(x), \theta_{-,\lambda}) \to +\infty$. Since $\frac{1+x}{1+\Psi_\lambda(x)} \to 0$ in the same limit, the second term of H_λ vanishes, and hence $H_\lambda(x) \to +\infty$.

For the second claim, since $\lambda \geq 1$ we have $\theta_{-,\lambda} = 0$, and substituting $\Psi_\lambda(\xi_\lambda) = 1$ into the definition of H_λ yields

$$H_\lambda(\xi_\lambda) = -\frac{\sqrt{R}(1+\xi_\lambda)}{2\lambda}\, Q(\xi_\lambda, \theta_{+,\lambda}).$$

Since $\lambda \leq \lambda_{K,R}$, (68) implies $Q(\xi_\lambda, \theta_{+,\lambda}) \geq 0$, with equality only at $\lambda = \lambda_{K,R}$, so $H_\lambda(\xi_\lambda) < 0$ for all $\lambda < \lambda_{K,R}$. The intermediate value theorem then guarantees a root $s_\lambda \in (\sqrt{2}-1, \xi_\lambda)$ of H_λ.

It remains to verify the sign conditions. Since $\theta_{-,\lambda} = 0$, and Ψ_λ is strictly decreasing with $\Psi_\lambda(\xi_\lambda) = 1$. Because $s_\lambda < \xi_\lambda$, it follows that $r_\lambda = \Psi_\lambda(s_\lambda) > 1$, and therefore

$$Q(r_\lambda, \theta_{-,\lambda}) = Q(r_\lambda, 0) = M \log r_\lambda > 0.$$

From $H_\lambda(s_\lambda) = 0$ with $\theta_{-,\lambda} = 0$ and $\frac{M}{\lambda(\theta_{-,\lambda}+M)} = \frac{1}{\lambda}$, we obtain

$$Q(s_\lambda, \theta_{+,\lambda}) = \frac{\lambda(1+r_\lambda)}{\sqrt{R}(1+s_\lambda)}\, Q(r_\lambda, 0) > 0.$$

Finally, writing $\gamma := Q(s_\lambda, \theta_{+,\lambda})$ and using the definition of Q,

$$\exp(\gamma) = s_\lambda^{M-1}\big((1+\theta_{+,\lambda})s_\lambda + \theta_{+,\lambda}\big) > s_\lambda^M,$$

since $\theta_{+,\lambda} > 0$ (as $\lambda < \lambda_{K,R} \leq \sqrt{R}$). Hence $0 < s_\lambda \exp(-\gamma/M) < 1$. □

Proposition 8. *For any $\lambda \in (0,1)$, the function $H_\lambda(\cdot)$ has a root $s_\lambda \in (\sqrt{2}-1, 1)$. Furthermore, if $r_\lambda = \Psi_\lambda(s_\lambda)$, then,*

$$0 < Q(s_\lambda, \theta_{+,\lambda}), \quad 0 < Q(r_\lambda, \theta_{-,\lambda}),$$
$$0 < s_\lambda \exp\left(\frac{-Q(s_\lambda, \theta_{+,\lambda})}{M}\right) < 1, \quad 0 < r_\lambda \exp\left(\frac{-Q(r_\lambda, \theta_{-,\lambda})}{M}\right) < 1\,.$$

Proof. Fix $\lambda \in (0,1)$. We drop the subscript λ for better readability. We apply the intermediate value theorem by showing that $H(x) \to +\infty$ as $x \downarrow \sqrt{2}-1$ and $H(1) < 0$.

Since $\lambda < 1$, we have $\theta_- = M(1/\lambda - 1) > 0$ and $\frac{M}{\lambda(\theta_- + M)} = 1$, so the definitions of Ψ and H reduce to

$$\Psi(x) = \sqrt{\frac{2(1+x)^2}{(1+x)^2 - 2}} - 1, \qquad H(x) = Q(\Psi(x), \theta_-) - \sqrt{R}\,\frac{1+x}{1+\Psi(x)}\, Q(x, \theta_+).$$

For the first claim, as $x \downarrow \sqrt{2}-1$ we have $\Psi(x) \to +\infty$, which implies $Q(\Psi(x), \theta_-) \to +\infty$. Since $Q(x, \theta_+)$ remains bounded while $\frac{1+x}{1+\Psi(x)} \to 0$, the second term vanishes and hence $H(x) \to +\infty$.

For the second claim, since $\Psi(1) = 1$ we have

$$H(1) = Q(1, \theta_-) - \sqrt{R}\, Q(1, \theta_+) = \log(1 + 2\theta_-) - \sqrt{R}\, \log(1 + 2\theta_+).$$

Since $R > 1$, we have $\theta_+ = M(\sqrt{R}/\lambda - 1) > M(1/\lambda - 1) = \theta_- > 0$, and therefore

$$\sqrt{R}\, \log(1 + 2\theta_+) > \log(1 + 2\theta_+) > \log(1 + 2\theta_-),$$

which gives $H(1) < 0$. The intermediate value theorem then guarantees a root $s_\lambda \in (\sqrt{2} - 1, 1)$ of H.

It remains to verify the sign conditions. Since Ψ is strictly decreasing and $s_\lambda < 1$, we have $r_\lambda = \Psi(s_\lambda) > \Psi(1) = 1$. Because $Q(\cdot, \theta_-)$ is strictly increasing and $Q(1, \theta_-) = \log(1 + 2\theta_-) > 0$, it follows that

$$Q(r_\lambda, \theta_-) > Q(1, \theta_-) > 0.$$

From $H(s_\lambda) = 0$ we then obtain

$$Q(s_\lambda, \theta_+) = \frac{(1 + r_\lambda)}{\sqrt{R}(1 + s_\lambda)}\, Q(r_\lambda, \theta_-) > 0.$$

Finally, writing $\gamma \coloneqq Q(s_\lambda, \theta_+)$ and using the definition of Q,

$$\exp(\gamma) = s_\lambda^{M-1}\big((1 + \theta_+)s_\lambda + \theta_+\big) = s_\lambda^M + \theta_+ s_\lambda^{M-1}(s_\lambda + 1) > s_\lambda^M,$$

since $\theta_+ > 0$. Combining with $s_\lambda > 0$, we imply that $0 < s_\lambda \exp(-\gamma/M) < 1$. An identical argument with $\delta \coloneqq Q(r_\lambda, \theta_-)$ and r_λ in place of θ_+ and s_λ gives $0 < r_\lambda \exp(-\delta/M) < 1$. □

Lemma 13. *For each λ, let z_λ denote the unique zero of $Q(x, \theta_{+,\lambda})$, defined in (69). Then, for any $0 < \lambda < \lambda' \le \sqrt{R}$, we have $z_\lambda < z_{\lambda'}$.*

Proof. Let $\xi_\lambda \coloneqq \theta_{+,\lambda}$ Notice that z_λ is the solution to $Q(z_\lambda, \xi_\lambda) = 0$. Differentiating it with respect to λ gives

$$\frac{\partial Q}{\partial z_\lambda}\frac{\mathrm{d}z_\lambda}{\mathrm{d}\lambda} + \frac{\partial Q}{\partial \xi_\lambda}\frac{\mathrm{d}\xi_\lambda}{\mathrm{d}\lambda} = 0$$

which implies that

$$\left(\frac{m-1}{z_\lambda} + \frac{1 + \xi_\lambda}{(1 + \xi_\lambda)z_\lambda + \xi_\lambda}\right)\frac{\mathrm{d}z_\lambda}{\mathrm{d}\lambda} = -\frac{1}{(1 + \xi_\lambda)z_\lambda + \xi_\lambda}\frac{\mathrm{d}\xi_\lambda}{\mathrm{d}\lambda}.$$

Since $\xi'_\lambda < 0$ it directly follows that $z'_\lambda > 0$, thereby proving the desired result. □

With the scalar existence theory in place, we can now return to the vector optimization problem. The theorem below is obtained by imposing the sparsity pattern appropriate to each phase, simplifying the KKT system under that pattern, and then invoking the previous scalar results to certify that the resulting formulas are well defined.

E.3 Analytic solution of Proposition 1

We first give the deferred definition of the pair (r_λ, s_λ). Recall from the main body the definitions

$$\theta_{+,\lambda} \coloneqq M[\sqrt{R}\,\lambda^{-1} - 1]_+, \quad \theta_{-,\lambda} \coloneqq M[\lambda^{-1} - 1]_+, \quad Q(x, \theta) \coloneqq \log\big(x^{M-1}((1 + \theta)x + \theta)\big). \tag{69}$$

For $\lambda \le \lambda_{K,R}$, let (s_λ, r_λ) denote the unique pair satisfying

$$\begin{cases} 1 + r_\lambda = \frac{M}{\lambda(\theta_{-,\lambda} + M)}\sqrt{\frac{2(1+s_\lambda)^2}{(1+s_\lambda)^2 - 2}}, \\ (1 + r_\lambda)\cdot Q(r_\lambda, \theta_{-,\lambda}) = \sqrt{R}\left(\frac{M}{\lambda(\theta_{-,\lambda} + M)}\right)(1 + s_\lambda)\cdot Q(s_\lambda, \theta_{+,\lambda}), \end{cases} \tag{70}$$

and set $\gamma_\lambda \coloneqq Q(s_\lambda, \theta_{+,\lambda})$ and $\delta_\lambda \coloneqq Q(r_\lambda, \theta_{-,\lambda})$ for all λ. Notice that the above equations are equivalent to $H_\lambda(s_\lambda) = 0, r_\lambda = \Psi(s_\lambda)$ with the definitions above.

Theorem 13 (Phase-wise analytic solution). *Let $\lambda_{K,R} \in [1, \sqrt{\frac{R+1}{2}}]$ be the unique root of*

$$Q(\frac{2\lambda}{\sqrt{2\lambda^2-1}} - 1, \theta_{+,\lambda}) \tag{71}$$

Then the unique solution $\boldsymbol{a}_\lambda$ of the minimization in (7) is given as follows:

(i) *For $\lambda \in [\sqrt{R}, \infty)$, $\boldsymbol{a}_\lambda = [0 \quad 0 \quad 0 \quad 0]^\top$*

(ii) *For $\lambda \in [\lambda_{K,R}, \sqrt{R})$, $\boldsymbol{a}_\lambda = \begin{bmatrix} 0 & 0 & -M\sqrt{R}\log s_\lambda & 0 \end{bmatrix}^\top$, where s_λ is the unique root of $Q(x, \theta_{+,\lambda})$.*

(iii) *For $\lambda \in [1, \lambda_{K,R})$,*

$$\boldsymbol{a}_\lambda = \begin{bmatrix} \frac{\sqrt{R}(\gamma_\lambda - \delta_\lambda)}{\sqrt{2(R+1)}} & \frac{R\gamma_\lambda + \delta_\lambda}{\sqrt{2(R+1)}} & \sqrt{R}(\gamma_\lambda - M\log s_\lambda) & 0 \end{bmatrix}^\top,$$

where for $s_\lambda \in (\sqrt{2}-1, 1)$ the root of $H_\lambda(x)$ and $r_\lambda := \Psi_\lambda(s_\lambda) \geq 1$, we have set

$$\gamma_\lambda = Q(s_\lambda, \theta_{+,\lambda}), \quad \delta_\lambda = Q(r_\lambda, \theta_{-,\lambda}). \tag{72}$$

(iv) *For $\lambda \in (0, 1)$,*

$$\boldsymbol{a}_\lambda = \begin{bmatrix} \frac{\sqrt{R}(\gamma_\lambda - \delta_\lambda)}{\sqrt{2(R+1)}} & \frac{R\gamma_\lambda + \delta_\lambda}{\sqrt{2(R+1)}} & \sqrt{R}(\gamma_\lambda - M\log s_\lambda) & \delta_\lambda - M\log r_\lambda \end{bmatrix}^\top,$$

where for $s_\lambda \in (\sqrt{2}-1, 1)$ the root of $H_\lambda(x)$, and $r_\lambda := \Psi(s_\lambda)$, we have set

$$\gamma_\lambda = Q(s_\lambda, \theta_{+,\lambda}), \quad \delta_\lambda = Q(r_\lambda, \theta_{-,\lambda}). \tag{73}$$

Proof. For clarity, we suppress the subscript λ throughout the proof. The existence of a unique $\lambda_{K,R} \in [1, \sqrt{R}]$ is guaranteed by Lemma 12.

We treat each phase separately, deriving the equations that arise from the KKT conditions under the appropriate sparsity pattern. In each case the resulting system admits a unique solution; the uniqueness of the overall minimizer from Theorem 12 then guarantees that the proposed formulas hold in every phase.

(i) This case would follow trivially from the sparsity features proved in Lemma 10.

(ii) Based on Lemma 10, any solution would have to satisfy $\boldsymbol{a}_+ \neq \boldsymbol{0}, \boldsymbol{a}_- = \boldsymbol{0}, \gamma = \delta = 0$. Assuming all entries in $\boldsymbol{a}_+$ are nonzero, we have $\boldsymbol{b}_+ = \lambda \boldsymbol{1}$, after which Lemma 11 (with $t = 1 - \frac{\lambda}{\sqrt{R}}$) would result in $\boldsymbol{a}_+ = -M\sqrt{R}\log s_\lambda$, where $Q(s, \theta_+) = 0$ and $s \in (0, 1)$

At this point, we only need to check the complementary KKT condition for $(\alpha, a) = (0, 0)$, which is $\beta^2 + b^2 \leq \lambda^2$. Substituting for this would lead to

$$X_+ = \frac{2M}{M + s^{1-M} + (M-1)s} = \frac{2\lambda}{\sqrt{R}(1+s)}, \quad X_- = 1$$
$$\frac{\beta^2 + b^2}{\lambda^2} = \frac{1}{2\lambda^2}(RX_+^2 + X_-^2) = \frac{2}{(1+s)^2} + \frac{1}{2\lambda^2}.$$

Moreover, using Lemma 13, we have $s \geq \frac{2\lambda_{K,R}}{\sqrt{2\lambda_{K,R}^2 - 1}} - 1$, which implies that $\beta^2 + b^2 \leq \lambda^2$. Thus, the unique solution in this phase has been found, validating the assumption of no zero entries in $\boldsymbol{a}_+$.

(iii) This is the first genuinely coupled regime: The majority coordinate and the mixed (α, a)-block are active, while the minority coordinate is still zero.

First, note that given $1 \leq \lambda \leq \sqrt{R}$, Lemma 10 would require $\boldsymbol{a}_+ \neq \boldsymbol{0}$ and $\boldsymbol{a}_- = \boldsymbol{0}$ for any solution. Assuming all entries in $\boldsymbol{a}_+$ are non-zero would again give $\boldsymbol{b}_+ = \lambda \boldsymbol{1}_{M-1}$,

after which Lemma 11 would solve yield $\boldsymbol{a}_+ = \sqrt{R}(\gamma - M\log s)\mathbf{1}_{M-1} \geq \mathbf{0}$, where $\gamma = Q(s, \theta_+)$ and s is unique for a given γ.

The KKT conditions for (α, a) in this phase are $(\beta, b) = \frac{\lambda}{\sqrt{a^2+\alpha^2}}(\alpha, a)$. Given that Lemma 10 guarantees $\gamma, \delta \geq 0$, they can be reformulated equivalently as

$$\lambda^2 = \beta^2 + b^2 = \frac{1}{2}(RX_+^2 + X_-^2), \quad \frac{\gamma}{X_+} = \frac{\delta}{X_-}. \tag{74}$$

Defining $r \coloneqq \exp(\delta/M)$, the above conditions would simplify to

$$\frac{1}{2} = \frac{1}{(1+s)^2} + \frac{1}{\lambda^2(1+r)^2}, \tag{75a}$$

$$\lambda\delta\,(1+r) = \sqrt{R}\gamma\,(1+s)\,. \tag{75b}$$

The existence of $s \in (\sqrt{2}-1, \xi_\lambda)$ and the fact that all terms in (72) are well-defined follow from Proposition 7. Consequently, the above system has a unique solution, which via uniqueness of the solution to the original problem, verifies the assumption of all-nonzero entries in $\boldsymbol{a}_+$.

Finally, we should note that by utilizing $\theta_- = 0$ in this phase, the above conditions can be written in the form of (70).

(iv) The final phase follows the same template as phase (iii), with the minority coordinate becoming active its only new feature.

Similar to the previous phase, Lemma 10 requires $\boldsymbol{a}_\pm \neq \mathbf{0}$. By assuming $\boldsymbol{a}_+$ and $\boldsymbol{a}_-$ have strictly positive entries, the KKT conditions give $\boldsymbol{b}_\pm = \lambda\mathbf{1}$, from which Lemma 11 would give

$$\begin{gathered}\boldsymbol{a}_+ = \sqrt{R}(\gamma - M\log s)\mathbf{1}_{M-1}, \quad \boldsymbol{a}_- = (\delta - M\log r)\mathbf{1}_{M-1}, \\ \gamma = Q(s, \theta_+), \quad \delta = Q(r, \theta_-).\end{gathered} \tag{76}$$

In addition, the reformulated KKT conditions for (α, a) of (74) would simplify to the slightly different form

$$\frac{1}{2} = \frac{1}{(1+s)^2} + \frac{1}{(1+r)^2}, \tag{77a}$$

$$\delta\,(1+r) = \sqrt{R}\gamma\,(1+s)\,. \tag{77b}$$

The above system of four equations can be reformulated as a scalar equation in s, where Proposition 8 guarantees the existence of a unique $s \in (\sqrt{2}-1, 1)$ and well-defined other quantities.

Finally, by using the fact that $\theta_- = 0$ for $\lambda \geq 1$, one can reach a unified form of equations for the third and fourth phase as in (70). □

E.4 Additional characteristics of the RP solution

In this subsection, we establish additional characteristics of the RP path. While properties such as Proposition 9 and Proposition 10 are structural inequalities characterizing the behavior of the RP solution, others are more technical and are introduced to facilitate the asymptotic analysis of the RP solution.

Proposition 9 (More characteristics of the regularization path solution)**.** *For all $\lambda \geq 0$, the regularization path solution parameters satisfy the following properties:*

$$a_{+,\lambda} \geq \sqrt{R}a_{-,\lambda}, \qquad \textit{equivalently,} \qquad \delta - \gamma \leq M\log(\frac{r}{s}) \tag{78a}$$

$$\alpha_\lambda \leq 0 \qquad \textit{equivalently,} \qquad 0 \leq \delta - \gamma \tag{78b}$$

We note that the bound (78a) is refined later. The refined bound requires a more technical argument and is only needed in a few specialized cases.

Proof. We first show that $\sqrt{R}\, a_- \le a_+$. This is immediate for $\lambda \ge 1$, since $a_- = 0$ and $a_+ \ge 0$. It remains to verify the claim for $\lambda \in (0, 1)$. In this regime, we have

$$\begin{aligned} a_+ = \sqrt{R}(\gamma - M\log s), \quad s^{M-1}((1+\theta_+)s + \theta_+) = e^{\gamma} \Rightarrow a_+ = \sqrt{R}\log(1 + \theta_+ + \frac{\theta_+}{s}) \\ a_- = (\delta - M\log r), \quad r^{M-1}((1+\theta_-)r + \theta_-) = e^{\delta} \Rightarrow a_- = \log(1 + \theta_- + \frac{\theta_-}{r}) \end{aligned} \tag{79}$$

We know that $s \le r$ from Proposition 8 and $\theta_+ = \sqrt{R}\theta_- + M(\sqrt{R} - 1) \ge \theta_-$. For $\lambda \le 1$, we therefore conclude that

$$\log(1 + \theta_- + \frac{\theta_-}{r}) \le \log(1 + \theta_+ + \frac{\theta_+}{s}) \Rightarrow a_- \le \frac{a_+}{\sqrt{R}},$$

after which substituting the expressions for $a_\pm$ results in the equivalent inequality $\delta - \gamma \le M\log(\frac{r}{s})$

For all values of λ, either $(\alpha, a) = (0, 0)$ or at least one of them is nonzero, which gives

$$\begin{aligned} \beta = \frac{\lambda\alpha}{\sqrt{\alpha^2 + a^2}} \Rightarrow \alpha &\propto [\mathbf{1}_M^\top \exp(\boldsymbol{x}_-) - \mathbf{1}_M^\top \exp(\boldsymbol{x}_+)] \\ &\propto e^{-\frac{\alpha\sqrt{R}}{M}} \left(e^{-\frac{a_-}{M\sqrt{R}}}(e^{\frac{a_-}{\sqrt{R}}} + M - 1)\right) - e^{\frac{\alpha}{M\sqrt{R}}} \left(e^{-\frac{a_+}{M\sqrt{R}}}(e^{\frac{a_+}{\sqrt{R}}} + M - 1)\right), \end{aligned}$$

where $\boldsymbol{a}_\pm = a_\pm \mathbf{1}$ was used to simplify the above equation. Now the claim can be proved by contradiction. Suppose that $\alpha > 0$. Since the expression $a^{M-1} + (M-1)a$ is increasing in a and $a_- \le \frac{a_+}{\sqrt{R}}$, the above expression would become negative, contradicting the prior assumption. Hence, $\alpha \le 0$, with strict inequality in its active phase.

□

The following lemma refines the bounds $s \in (\sqrt{2} - 1, 1]$ and $r \in [1, \infty)$ established in Theorem 13.

Lemma 14. *For the regularization path characterized (76) and (77), the following bounds can be established:*

$$s_\infty \le s_\lambda \le 1, \quad 1 \le r_\lambda \le r_\infty, \qquad s_\infty \coloneqq \sqrt{\frac{2(R+1)}{R}} - 1 \quad r_\infty \coloneqq \sqrt{2(R+1)} - 1 \tag{80}$$

$$(1 + r_\lambda) \le \sqrt{R}(1 + s_\lambda) \tag{81}$$

$$\delta_\lambda \le \sqrt{R}\gamma_\lambda \tag{82}$$

Proof. Note that the equations for phase (i) can be interpreted as $r = s = 1$ and $\gamma = \delta = 0$ where (81) holds trivially. This establishes the claim for all $\lambda \ge 0$. For phase (ii) we simliarly have $r = 1$ and $\gamma = \delta = 0$. For the remaining bound, using Lemma 13, we have

$$s \ge \frac{2\lambda_{K,R}}{\sqrt{2\lambda_{K,R}^2 - 1}} - 1 \ge \sqrt{\frac{2(R+1)}{R}} - 1 = s_\infty.$$

For other two phases, recall that $s \in (\sqrt{2} - 1, 1]$ and $r \in [1, \infty)$ has already been established in Proposition 8. Furthermore, notice that using $\delta \ge \gamma$ from (78b) in the phase-wise equations would result in

$$(iii): \quad \sqrt{R}\gamma(1+s) = \lambda(1+r)\delta \ge \lambda(1+r)\gamma \Rightarrow \sqrt{R}(1+s) \ge \lambda(1+r) \tag{83a}$$

$$(iv): \quad \sqrt{R}\gamma(1+s) = (1+r)\delta \ge (1+r)\gamma \Rightarrow \sqrt{R}(1+s) \ge (1+r), \tag{83b}$$

after which we reach (81) for phases (iii) and (iv) by $\lambda \ge 1$ in phase (iii).

To prove the final claim, we substitute $s \le 1 \le r$ from Theorem 13 in (75b) and (77b) to prove the claim:

$$(iii): \delta = \sqrt{R}\gamma\frac{1+s}{\lambda(1+r)} \le \sqrt{R}\gamma, \quad (iv): \delta = \sqrt{R}\gamma\frac{1+s}{1+r} \le \sqrt{R}\gamma.$$

Substituting (83a) in (75a) and (83b) in (77a) yields the desired results (80). □

Proposition 10 (More RP characteristics - part 2)**.** *The following inequality holds for* $\forall \lambda \geq 0$

$$\frac{a_+}{\sqrt{R}} \geq \sqrt{\frac{2}{R+1}}\sqrt{a^2+\alpha^2}. \tag{84}$$

Proof. The above inequality trivially holds true for phases (i) and (ii), and we need to prove it for phases (iii) and (iv). Squaring the above inequality and substituting the phase-wise equations leads to the equivalent inequality

$$(\gamma - M\log s)^2 \geq \frac{R\gamma^2+\delta^2}{R+1} = \gamma^2 + \frac{\delta^2-\gamma^2}{R+1}. \tag{85}$$

Simplifying (85) and using (82) yields the following equivalent claim on the left along with the bound established so far on the right hand:

$$\frac{(\delta-\gamma)(\delta+\gamma)}{R+1} \leq (M\log s)^2 + 2\gamma M\log(\frac{1}{s}), \quad \frac{(\delta-\gamma)(\delta+\gamma)}{R+1} \leq \gamma M\frac{\sqrt{R}+1}{R+1}\log(\frac{r}{s}).$$

Discarding $(M\log s)^2 \geq 0$, it is sufficient that

$$\forall \lambda \leq \lambda_{K,R} : 2\log(\frac{1}{s_\lambda}) \geq \frac{\sqrt{R}+1}{R+1}\log(\frac{r_\lambda}{s_\lambda}) \iff \left(\frac{2(R+1)}{\sqrt{R}+1}-1\right)\log s_\lambda + \log r_\lambda \leq 0.$$

As proved in Lemma 15, this bound is reduced to a scalar inequality in R, which is subsequently established in Lemma 17, which proves the desired result. □

The following lemma reduces checking the useful auxiliary inequality for the regularization path to checking a scalar inequality involving R

Lemma 15. *For any* $c_R \geq 0$ *and* s_λ, r_λ *being the regularization path parameters, it is sufficient to check any inequality of the form* $c_R\log s_\infty + \log r_\infty \leq 0$ *for the pair* s_∞, r_∞*; in other words,*

$$c_R\log s_\infty + \log r_\infty \leq 0 \Rightarrow [\forall \lambda \geq 0 : c_R\log s_\lambda + \log r_\lambda \leq 0]. \tag{86}$$

Proof. The above is trivially true for $\lambda \geq R$, as $r_\lambda = s_\lambda = 1$ in this phase. Similarly, $s_\lambda < 1, r_\lambda = 1$ for $\lambda_{k,R} \leq \lambda \leq \sqrt{R}$ establishes the above for phase (ii) as well. The rest of the proof focuses on proving the convexity of $\log s_\lambda$, $\log r_\lambda$ with a suitable reformulation, which means the inequality only needs to be checked at endpoints.

Notice that phase (iii) and phase (iv) equations give

$$(iii) : \frac{1}{2} = \frac{1}{(1+s)^2} + \frac{1}{\lambda^2(1+r)^2} \quad (iv) : \frac{1}{2} = \frac{1}{(1+s)^2} + \frac{1}{(1+r)^2}.$$

Define $\frac{1}{1+s} = \frac{1}{\sqrt{2}}\cos\phi$ to reach $s(\phi) = \frac{\sqrt{2}}{\cos\phi} - 1$ in both phases, and

$$(iii) : r = \frac{\sqrt{2}}{\lambda\sin\phi} - 1 \leq \frac{\sqrt{2}}{\sin\phi} - 1 \qquad (iv) : r(\phi) = \frac{\sqrt{2}}{\sin\phi} - 1,$$

where we used $\lambda \geq 1$ to reach the upper bound for phase (iii). As a result, it is sufficient to check

$$c_R\log s(\phi) + \log r(\phi) \leq 0$$

in the admissible range in phases (iii), (iv). We can obtain the following bounds on the range, using (78b) and (82)

$$(iii) : \tan\phi = \frac{1+s}{\lambda(1+r)} = \frac{\delta}{\sqrt{R}\gamma} \in [\frac{1}{\sqrt{R}}, 1], \quad (iv) : \tan\phi = \frac{1+s}{1+r} = \frac{\delta}{\sqrt{R}\gamma} \in [\frac{1}{\sqrt{R}}, 1].$$

Next, we differentiate both functions twice and simplify to reach

$$(\log s(\phi))'' = \frac{\sqrt{2}}{\cos\phi(\sqrt{2}-\cos\phi)^2}\left(\frac{\sqrt{2}}{\cos\phi} - 2 + \cos^2\phi\right),$$

$$(\log r(\phi))'' = \frac{\sqrt{2}}{\sin\phi(\sqrt{2}-\sin\phi)^2}\left(\frac{\sqrt{2}}{\sin\phi} - 2 + \sin^2\phi\right).$$

It remains to show that both second derivatives are positive in the admissible range, which we will show below. This establishes the convexity of both functions in ϕ, and thus $c_R \log s(\phi) + \log r(\phi)$. It directly follows that it is sufficient to check the inequality at the endpoints $\phi = \frac{\pi}{4}$ which is trivial, and $\phi = \arctan(\frac{1}{\sqrt{R}})$, which recovers s_∞, r_∞ as defined in Lemma 14.

To show that the second derivatives are positive, note that

$$\forall z > 0: \ \frac{\sqrt{2}}{z} - 2 + z^2 \geq 3 \times 2^{-1/3} - 2 > 0,$$

which establish the non-negativity claims.

□

Lemma 16 (Refinement of (78a))**.** *For $\lambda \geq 0$, the regularization path parameters satisfy*

$$\delta - \gamma \leq (M-1)\log(\frac{r}{s}) \tag{87}$$

Proof. We begin by using (79) to rewrite the difference $\delta - \gamma - M\log(\frac{r}{s})$ in the form of

$$\delta - \gamma - M\log(\frac{r}{s}) = \log\left(\frac{1+\theta_-(\frac{1+r}{r})}{1+\theta_+(\frac{1+s}{s})}\right),$$

where $\theta_+ = \sqrt{R}\,\theta_- + M(\sqrt{R}-1)$. Define the auxiliary variables and function

$$f(\theta) = \frac{1+\theta B}{1+(\sqrt{R}\theta + A)C} \quad B = \frac{1+r}{r} \quad C = \frac{1+s}{s} \quad A = (\sqrt{R}-1)M.$$

Note that r, s depend implicitly on θ as well, so we calculate the partial derivative as

$$\partial_\theta f = \frac{B(1+AC) - \sqrt{R}C}{(1+(\sqrt{R}\theta + A)C)^2}.$$

Using $M \geq 1$, the numerator is lower bounded in the following way:

$$\begin{aligned} B(1+AC) - \sqrt{R}C &\geq B(1+(\sqrt{R}-1)C) - \sqrt{R}C \\ &= \frac{1}{rs}\left((1+r)s - \sqrt{R}(1+s)r + (1+r)(1+s)(\sqrt{R}-1)\right) \\ &= \frac{1}{rs}\left(\sqrt{R}(1+s) - (1+r)\right) \geq 0, \end{aligned}$$

where nonnegativity follows by $(1+r) \leq \sqrt{R}(1+s)$ from (81). Consequently, $\partial_\theta f \geq 0$ and treating s, r as constants, the function is bounded by its limit in $\theta \to \infty$, resulting in

$$f(\theta) \leq \lim_{\theta\to\infty} f(\theta) = \frac{s}{r}\frac{1+r}{\sqrt{R}(1+s)}.$$

Taking the logarithm and using $1 + r \leq \sqrt{R}(1+s)$ once more completes the proof:

$$\delta - \gamma \leq (M-1)\log(\frac{r}{s}) + \log(\frac{1+r}{\sqrt{R}(1+s)}) \leq (M-1)\log(\frac{r}{s}).$$

□

Lemma 17. *For any $R \geq 1$, the inequality*

$$c_R \log s_\infty + \log r_\infty \leq 0 \tag{88}$$

holds for the two following choices of c_R:

$$c_R^{(1)} = \frac{2(R+1)}{\sqrt{R}+1} - 1, \qquad c_R^{(2)} = \sqrt{R}. \tag{89}$$

Proof. Note that since $s_\infty \le 1$ implies $\log s_\infty \le 0$ and

$$c_R^{(1)} - c_R^{(2)} = \left(\frac{2(R+1)}{\sqrt{R}+1} - 1\right) - \sqrt{R} = \frac{(\sqrt{R}-1)^2}{\sqrt{R}+1} \ge 0,$$

it follows that $c_R^{(1)} \log s_\infty + \log r_\infty \le c_R^{(2)} \log s_\infty + \log r_\infty$. As a result, it is sufficient to prove the inequality for the choice $c_R^{(2)}$. Recall the definitions of s_∞, r_∞ and define

$$f(R) \coloneqq \sqrt{R}\log(s_\infty(R)) + \log(r_\infty(R)), \quad s_\infty = \sqrt{\frac{2(R+1)}{R}} - 1 \le 1, \quad r_\infty = \sqrt{2(R+1)} - 1 \ge 1.$$

The derivative with respect to R after simplification is

$$f'(R) = \frac{1}{2\sqrt{R}} \log s_\infty + \frac{(R-1)(\sqrt{2} - \sqrt{R+1})}{\sqrt{2}\, s_\infty r_\infty R^2} \le 0,$$

which results in f being non-increasing over $R \ge 1$. Since $f(1) = 0$, we conclude that $f(R) \le 0$ which is the desired result. □

E.5 Behavior of the threshold $\lambda_{K,R}$

Proposition 11. *Consider the threshold $\lambda_{K,R} \in [1, \sqrt{\frac{R+1}{2}}]$ between phases (ii) and (iii) in Theorem 13, with its value defined by (71). Then, it has the following properties:*

1. *$\lambda_{K,R}$ is decreasing as a function of M for fixed R, and $\lim_{M\to\infty} \lambda_{K,R} = 1$*
2. *For fixed M, $\lambda_{K,R}$ is increasing in R, but $\frac{\lambda_{K,R}}{\sqrt{R}}$ decreases and $\lim_{R\to\infty} \frac{\lambda_{K,R}}{\sqrt{R}} = c_K$, where c_K solely depends on K.*

Proof. The threshold is the value of λ where phases (ii) and (iii) intersect; as a result, $\gamma = \delta = 0$ and at the same time $\beta^2 + b^2 = \lambda^2$. Using these to modify the phase (iii) equations leads us to

$$1 = \frac{2}{(1+s)^2} + \frac{1}{2\lambda_{K,R}^2}, \quad Q(s,\theta_+) = \log\big(s^{M-1}((1+\theta_+)s+\theta_+)\big) = 0, \quad \theta_+ = M[\sqrt{R}\lambda_{K,R}^{-1} - 1] \tag{90}$$

From this point, we drop the subscript from $\lambda_{M,R}$ to reduce clutter. Solving for λ in terms of s yields

$$\lambda(s) = \frac{1+s}{\sqrt{2(s^2+2s-1)}}, \quad \lambda'(s) = -\frac{2}{\sqrt{2}\,(s^2+2s-1)^{3/2}} < 0, \quad \partial_s\theta_+ = -\frac{M\sqrt{R}}{\lambda^2}\lambda'(s) > 0$$

Assuming M is continuous, implicitly differentiating $Q(s,\theta_+) = 0$ with respect to it and simplifying would result in

$$Q'(M) = \partial_M Q + (\partial_s Q)\, s'(M) = 0,$$
$$\partial_M Q = \frac{1 - s^M + \log(s^M)}{M} \le 0, \quad \partial_s Q = \frac{M-1}{s} + \frac{(1+\theta_+) + (1+s)\,\partial_s\theta_+}{(1+\theta_+)s + \theta_+} \ge 0,$$

from which we can conclude $s'(M) \ge 0$ and thus $\lambda'(M) \le 0$.

The same differentiation method can be applied to $\sqrt{R}$ to reach

$$Q'(\sqrt{R}) = \partial_{\sqrt{R}} Q + (\partial_s Q)\, s'(\sqrt{R}) = 0, \quad \partial_s Q = \frac{(M+\theta_+)(1+s)}{\sqrt{R}\,((1+\theta_+)s+\theta_+)} \ge 0,$$

from which follows that s is decreasing in $\sqrt{R}$ (and thus R) and λ is increasing in R. However, to determine the behavior of the ratio to λ_+, we reformulate $Q(s,\theta_+) = 0$ as

$$\frac{\sqrt{R}}{\lambda} = \frac{s^{-(M-1)} - s}{M(1+s)} + 1,$$

where the RHS is decreasing in s. This will result in $\frac{\lambda_{K,R}}{\sqrt{R}}$ to be decreasing in R.

To derive the asymptotics, first note that by using $s \geq \sqrt{2}-1$, the following can be established:

$$s^M(1+\theta_+) \leq s^{M-1}\left((1+\theta_+)s+\theta_+\right) \leq s^M(1+\theta_+)(1+\frac{1}{s}) \leq \frac{\sqrt{2}}{\sqrt{2}-1}s^M(1+\theta_+),$$
$$(\frac{\sqrt{2}}{\sqrt{2}-1})^{1/M}\frac{1}{(1+\theta_+)^{1/M}} \leq s \leq \frac{1}{(1+\theta_+)^{1/M}}.$$

□

Since $\lambda_{K,R} \in [1, \sqrt{\frac{R+1}{2}}]$ stays bounded for all M, $\theta_+ = \mathcal{O}(M)$, and the limit of upper and lower bounds are 1. Consequently, $\lim_{M\to\infty}\lambda_{K,R}=1$. To derive the limit for $R\to\infty$, we substitute for θ_+ is terms of s to reach

$$(s+1)^2-2 = \frac{1}{2R}(\frac{s^{-(M-1)}-s}{M}+1+s)^{-2} \Rightarrow \lim_{R\to\infty}(s+1)^2-2=0 \Rightarrow \lim_{R\to\infty} s=\sqrt{2}-1,$$

where the boundedness of s for all R was used to calculate the LHS limit. Substituting this into $Q(s,\theta_+)=0$ would yield

$$\lim_{R\to\infty}\frac{\lambda}{\sqrt{R}} = \big(1+\frac{(\sqrt{2}-1)^{-M+1}-(\sqrt{2}-1)}{M\sqrt{2}}\big)^{-1} = c_K$$

E.6 Asymptotic behavior of RP solution and Proof of Theorem 3

Proposition 12 (Asymptotic direction as $\lambda\to 0^+$)**.** *For each $\lambda\in(0,1)$, consider the solution of the KKT system in (76) and (77) of the form*

$$\boldsymbol{a}_+ = \sqrt{R}(\gamma - M\log s)\,\mathbf{1}_{M-1},\ \boldsymbol{a}_- = (\delta-\log r)\,\mathbf{1}_{M-1},\ a=\frac{R\gamma+\delta}{\sqrt{2(R+1)}},\ \alpha=\frac{\sqrt{R}(\gamma-\delta)}{\sqrt{2(R+1)}},$$

where all parameters are nonzero. Define $\boldsymbol{a}_\infty \coloneqq (0, \sqrt{\frac{R+1}{2}}, \sqrt{R}\mathbf{1}_{M-1}, \mathbf{1}_{M-1})$. Also define the norm operator $\|\boldsymbol{a}\|_{+1} \coloneqq \sqrt{a^2+\alpha^2}+\|\boldsymbol{a}_+\|_1+\|\boldsymbol{a}_-\|_1$. Then the following limit and rate of convergence hold:

$$\lim_{\lambda\to0^+}\frac{\boldsymbol{a}_\lambda}{\|\boldsymbol{a}_\lambda\|_{+1}} = \frac{\boldsymbol{a}_\infty}{\|\boldsymbol{a}_\infty\|_{+1}}, \qquad \frac{\boldsymbol{a}_\lambda}{\|\boldsymbol{a}_\lambda\|_{+1}} = \frac{\boldsymbol{a}_\infty}{\|\boldsymbol{a}_\infty\|_{+1}} + \mathcal{O}(\frac{1}{\log\lambda^{-1}}) \quad \text{as } \lambda\to0^+$$

Proof. We bound the quantities using (77), along with the bounds established in Lemma 14.

Using the fact that $a_0=\sqrt{R}(\gamma-M\log(s))$ and $a_1=\delta-M\log(r)$, the following bounds hold:

$$s^M(1+\theta_+) \leq s^{M-1}\left((1+\theta_+)s+\theta_+\right) = e^\gamma \leq s^M(1+\theta_+)(1+\frac{1}{s})$$
$$\Rightarrow \log(1+\theta_+) \leq \gamma - M\log(s) \leq \log(1+\theta_+) - \log(1-\sqrt{\frac{R}{2(R+1)}}),$$
$$\log(1+\theta_-) \leq \delta - M\log(r) \leq \log(1+\theta_-)+\log(2),$$

Since s, r are bounded from above and below by Lemma 14,

$$\begin{aligned}
\gamma &= \log(1+\theta_+)+c_\gamma, \quad M\log(\sqrt{\frac{2(R+1)}{R}}-1) \leq c_\gamma \leq -\log(1-\sqrt{\frac{R}{2(R+1)}}) \\
\delta &= \log(1+\theta_-)+c_\delta, \quad 0\leq c_\delta \leq \log(2)+\log(\sqrt{2(R+1)}-1) \\
\frac{a_+}{\sqrt{R}} &= \log(1+\theta_+)+c_{a_+}, \quad 0\leq c_{a_+} \leq -\log(1-\sqrt{\frac{R}{2(R+1)}}) \\
\frac{a_-}{\sqrt{R}} &= \log(1+\theta_-)+c_{a_-}, \quad 0\leq c_{a_-} \leq \log(2)
\end{aligned} \tag{91}$$

Since $\log(1+\theta_\pm)$ scale as $\mathcal{O}(\log(\lambda^{-1}))$, it is evident from the above bounds in (91) that all quantities grow at this rate, which establishes the desired limit. As a result, the norm $\|\boldsymbol{a}_\lambda\|_{+1}$ scales as $\mathcal{O}(\log(\lambda^{-1}))$, which results in the normalized quantity deviating from its limit in the following way:

$$\lim_{\lambda\to 0^+}\frac{\boldsymbol{a}_\lambda}{\|\boldsymbol{a}_\lambda\|_{+1}}=\frac{\boldsymbol{a}_\infty}{\|\boldsymbol{a}_\infty\|_{+1}}+\mathcal{O}(\frac{1}{\log\lambda^{-1}}).$$

It is also worth noting that since $\delta/\gamma\to 1$ as $\lambda\to 0^+$, (77) will lead to $s\to s_\infty,\ r\to r_\infty$. □

Theorem 14 (Direction of convergence of modes (restatement of Theorem 4))**.** *For $\lambda\le\sqrt{R}$, the logit spectral parameters satisfy*

$$\frac{a_{+,\lambda}}{\|\boldsymbol{a}_\lambda\|_{+1}}\ge\frac{a_{+,\infty}}{\|\boldsymbol{a}_\infty\|_{+1}},\quad \frac{a_{-,\lambda}}{\|\boldsymbol{a}_\lambda\|_{+1}}\le\frac{a_{-,\infty}}{\|\boldsymbol{a}_\infty\|_{+1}},\quad \frac{\sqrt{a_\lambda^2+\alpha_\lambda^2}}{\|\boldsymbol{a}_\lambda\|_{+1}}\le\frac{\sqrt{a_\infty^2+\alpha_\infty^2}}{\|\boldsymbol{a}_\infty\|_{+1}}. \tag{92}$$

As a result of (92) and Proposition 12, the normalized majority mode approaches its limit from above, and the normalized minority and majority-minority modes approach their limit from below as $\lambda\to 0^+$

Proof. We suppress the subscript λ to reduce clutter. First, we expand the difference between the normalized a_+ and its limit $a_{+,\infty}$ to reach

$$\frac{a_+}{\|\boldsymbol{a}\|_{+1}}-\frac{a_{+,\infty}}{\|\boldsymbol{a}_\infty\|_{+1}}=\frac{(M-1)(a_+-\sqrt{R}\,a_-)+(\sqrt{\frac{R+1}{2}}a_+-\sqrt{R}\sqrt{a^2+\alpha^2})}{\|\boldsymbol{a}\|_{+1}\|\boldsymbol{a}_\infty\|_{+1}}\ge 0.$$

where both of the grouped terms are non-negative as a result of Proposition 9 and Proposition 10.

The same procedure for a_+ and its limit will lead to

$$\frac{\sqrt{a^2+\alpha^2}}{\|\boldsymbol{a}\|_{+1}}-\frac{\sqrt{a_\infty^2+\alpha_\infty^2}}{\|\boldsymbol{a}_\infty\|_{+1}}=\frac{\sqrt{\frac{R+1}{2}}(M-1)}{\|\boldsymbol{a}\|_{+1}\|\boldsymbol{a}_\infty\|_{+1}}\left[(\sqrt{R}+1)\sqrt{\frac{2}{R+1}}\sqrt{a^2+\alpha^2}-(a_-+a_+)\right]$$

Substituting the phase-wise equations in terms of γ,δ,r,s, alongside $\sqrt{a^2+\alpha^2}=\sqrt{\frac{R\gamma^2+\delta^2}{2}}$ will reduce our claim to

$$(\sqrt{R}+1)\sqrt{\frac{R\gamma^2+\delta^2}{R+1}}-\left[\sqrt{R}(\gamma-M\log s)+\delta-M\log r\right]\le 0.$$

Note that the concavity of $\sqrt{x}$ results in $\sqrt{x+a}\le\sqrt{x}+\frac{a}{2\sqrt{x}}$, which in our case can be applied to reach

$$\sqrt{\frac{R\gamma^2+\delta^2}{R+1}}=\sqrt{\gamma^2+\frac{\delta^2-\gamma^2}{R+1}}\le\gamma+\frac{\delta^2-\gamma^2}{2(R+1)\gamma}.$$

Using the above inequality and rearranging, it is sufficient to prove

$$(\delta-\gamma)\left(\frac{\sqrt{R}+1}{2(R+1)}(\frac{\delta}{\gamma}+1)-1\right)+M\left(\sqrt{R}\log s+\log r\right)\le 0.$$

Both terms are then proved to be negative. The term $\sqrt{R}\log s+\log r$ is negative as a consequence of Lemma 15 and Lemma 17. To prove that the first term is negative, note that $\delta-\gamma\ge 0$ from (78b), and $\delta\le\sqrt{R}\gamma$, which will lead to

$$(\delta-\gamma)\left(\frac{\sqrt{R}+1}{2(R+1)}(\frac{\delta}{\gamma}+1)-1\right)\le(\delta-\gamma)\left(\frac{(\sqrt{R}+1)^2}{2(R+1)}-1\right)\le-(\delta-\gamma)\frac{(\sqrt{R}-1)^2}{2(R+1)}\le 0.$$

The same procedure for a_- and its limit will lead to the claim

$$\frac{a_-}{\|\boldsymbol{a}\|_{+1}}-\frac{a_{-,\infty}}{\|\boldsymbol{a}_\infty\|_{+1}}=\frac{(M-1)(\sqrt{R}\,a_--a_+)+(\sqrt{\frac{R+1}{2}}a_--\sqrt{a^2+\alpha^2})}{\|\boldsymbol{a}\|_{+1}\|\boldsymbol{a}_\infty\|_{+1}}\le 0,$$

where we need to show the numerator to be negative. After substituting for γ, δ, r, s and rearranging, we reach

$$\sqrt{R}(M-1)\left(\delta-\gamma-M\log(\frac{r}{s})\right)+\sqrt{\frac{R+1}{2}}(\delta-M\log r)-\frac{\sqrt{R}\gamma}{2}(1+s)\leq 0.$$

By using the refined bound $\delta-\gamma\leq(M-1)\log(\frac{r}{s})$ from Lemma 16, it is sufficient to prove

$$(M-1)\log(\frac{r}{s})\left(\sqrt{\frac{R+1}{2}}-\sqrt{R}\right)-M\sqrt{\frac{R+1}{2}}\log r+\frac{\sqrt{R}\gamma}{2}\left(\sqrt{\frac{2(R+1)}{R}}-(1+s)\right)\leq 0.$$

All three terms are negative since $\sqrt{\frac{R+1}{2}}\leq\sqrt{R}$, and the bounds $1\leq r/s$, $r\geq 1$ and $s+1\geq\sqrt{\frac{2(R+1)}{R}}$ follow from Lemma 14. This completes the proof. □

E.7 Alternative proof of regularization-path limit

For completeness, we verify here that the max-margin problem which is associated with the regularization path limit of (7) has SEL as its unique solution. Note that Theorem 3 proves that directly by taking an explicit limit of the regularization-path solution. Alternatively, we can show through proof by contradiction, which is a now classical argument due to Rosset et al. (2003; 2004) and used extensively in the recent literature (Ji et al., 2020; Kini et al., 2021; Tarzanagh et al., 2023, e.g.), that the solution of (7) directionally converges in the limit $\lambda\to 0$ to the (normalized) max-margin solution below:

$$\begin{aligned}\min_{\boldsymbol{a}_+,\boldsymbol{a}_-,a,\alpha}\quad & \|\boldsymbol{a}_+\|_1+\|\boldsymbol{a}_-\|_1+\sqrt{a^2+\alpha^2}\\ \text{s.t.}\quad & \forall i\in[M-1]:\frac{1}{\sqrt{R}}(\mathbf{1}-\overline{\boldsymbol{\phi}}_{i+1})^\top\boldsymbol{a}_+\geq 1,\quad \frac{1}{\sqrt{R}}\mathbf{1}^\top\boldsymbol{a}_++\sqrt{\frac{2}{R+1}}\left(a+\frac{\alpha}{\sqrt{R}}\right)\geq 1\\ & \forall i\in[M-1]:(\mathbf{1}-\overline{\boldsymbol{\phi}}_{i+1})^\top\boldsymbol{a}_-\geq 1,\quad \mathbf{1}^\top\boldsymbol{a}_-+\sqrt{\frac{2}{R+1}}\left(a-\sqrt{R}\alpha\right)\geq 1\end{aligned}\tag{93}$$

We omit the details of this standard argument because they bring no new insights. Instead, we verify here that the unique solution of this max-margin problem matches the SEL geometry of Theorem 3. This also yields an alternative proof of (Thrampoulidis et al., 2022, Theorem 1, Proposition 2). Note that compared to our approach in Theorem 3 the overall proof here is not as explicit (cf. proof by contradiction) and does not yield rates of convergence.

Theorem 15. *The optimization problem of (93) admits the unique solution of*

$$\boldsymbol{a}_+=\frac{1}{M}\sqrt{R}\mathbf{1}_{M-1},\quad \boldsymbol{a}_-=\frac{1}{M}\mathbf{1}_{M-1},\quad a=\frac{1}{M}\sqrt{\frac{R+1}{2}},\quad \alpha=0\tag{94}$$

Proof. Define the rescaled variables $\hat{\boldsymbol{a}}_+=\frac{1}{\sqrt{R}}\boldsymbol{a}_+$, $\hat{\boldsymbol{a}}_-=\boldsymbol{a}_-$, $(\hat{a},\hat{\alpha})=\sqrt{\frac{2}{R+1}}(a,\alpha)$. We also omit the dimension subscripts to reduce clutter. The problem can be written as

$$\begin{aligned}\min\quad & \sqrt{R}\|\hat{\boldsymbol{a}}_+\|_1+\|\hat{\boldsymbol{a}}_-\|_1+\sqrt{\frac{R+1}{2}}\sqrt{\hat{a}^2+\hat{\alpha}^2}\\ \text{s.t.}\quad & \boldsymbol{\Psi}\hat{\boldsymbol{a}}_+\geq\mathbf{1},\quad \boldsymbol{\Psi}\hat{\boldsymbol{a}}_-\geq\mathbf{1},\quad \mathbf{1}^\top\hat{\boldsymbol{a}}_++(\hat{a}+\frac{\hat{\alpha}}{\sqrt{R}})\geq 1,\quad \mathbf{1}^\top\hat{\boldsymbol{a}}_-+(\hat{a}-\sqrt{R}\hat{\alpha})\geq 1.\end{aligned}$$

The KKT conditions in terms of the dual variables are

$$\sqrt{R}\,\mathrm{sign}(\hat{\boldsymbol{a}}_+)=\boldsymbol{\Psi}\boldsymbol{\nu}_++\lambda_+\mathbf{1},\quad \mathrm{sign}(\hat{\boldsymbol{a}}_-)=\boldsymbol{\Psi}\boldsymbol{\nu}_-+\lambda_-\mathbf{1},\tag{95a}$$

$$\sqrt{\frac{2}{R+1}}(\lambda_++\lambda_-,\frac{\lambda_+}{\sqrt{R}}-\lambda_-\sqrt{R})\in\partial_{(\hat{a},\hat{\alpha})}(\sqrt{\hat{a}^2+\hat{\alpha}^2})\tag{95b}$$

$$\boldsymbol{\nu}_\pm\geq\mathbf{0},\quad \lambda_\pm\geq 0,\quad \boldsymbol{\nu}_\pm\odot(\boldsymbol{\Psi}\hat{\boldsymbol{a}}_\pm-\mathbf{1})=\mathbf{0},\tag{95c}$$

$$\lambda_+(\mathbf{1}^\top\hat{\boldsymbol{a}}_++(\hat{a}+\frac{\hat{\alpha}}{\sqrt{R}})-1)=0,\quad \lambda_-(\mathbf{1}^\top\hat{\boldsymbol{a}}_-+(\hat{a}-\hat{\alpha}\sqrt{R})-1)=0,\tag{95d}$$

where the sign function is treated in an extended value form. Notice that since the entries of $\boldsymbol{\Psi}$ are nonnegative, (95a) alongside the dual constraints in (95c) result in nonnegative signs for $\boldsymbol{a}_+, \boldsymbol{a}_-$ and thus $\boldsymbol{a}_+ \geq \mathbf{0}, \boldsymbol{a}_- \geq \mathbf{0}$. From this point on, our first step is showing that the only solution satisfying $(\hat{a}, \hat{\alpha}) \neq (0, 0)$ and $\boldsymbol{a}_+ > \mathbf{0}, \boldsymbol{a}_- > \mathbf{0}$ is the solution in the theorem statement. Assuming that these conditions would result in the simplified KKT equations

$$\boldsymbol{\nu}_+ = \boldsymbol{\Psi}^{-1}(\sqrt{R} - \lambda_+)\mathbf{1} = \frac{1}{M}(\sqrt{R} - \lambda_+)\mathbf{1}, \quad \boldsymbol{\nu}_- = \frac{1}{M}(1 - \lambda_-)\mathbf{1} \tag{96a}$$

$$\sqrt{\frac{2}{R+1}}(\lambda_+ + \lambda_-, \frac{\lambda_+}{\sqrt{R}} - \lambda_-\sqrt{R}) = \frac{1}{\sqrt{\hat{a}^2 + \hat{\alpha}^2}}(\hat{a}, \hat{\alpha}). \tag{96b}$$

Notice that (96b) implies $\frac{\lambda_+^2}{R} + \lambda_-^2 = \frac{1}{2}$. This constraint eliminates the possibility for $\lambda_+ = \sqrt{R}$ or $\lambda_- = 1$, from which it follows that $\boldsymbol{\nu}_+ > \mathbf{0}, \boldsymbol{\nu}_- > \mathbf{0}$. As a result of complementary slackness, the constraints $\boldsymbol{\Psi}\hat{\boldsymbol{a}}_+ \geq \mathbf{1}$ and $\boldsymbol{\Psi}\hat{\boldsymbol{a}}_- \geq \mathbf{1}$ are active, from which $\hat{\boldsymbol{a}}_+, \hat{\boldsymbol{a}}_-$ can be recovered as

$$\boldsymbol{\Psi}\hat{\boldsymbol{a}}_+ = \mathbf{1}, \ \boldsymbol{\Psi}\hat{\boldsymbol{a}}_- = \mathbf{1} \Rightarrow \hat{\boldsymbol{a}}_+ = \frac{1}{M}\mathbf{1}, \ \boldsymbol{a}_- = \frac{1}{M}\mathbf{1}.$$

As a result, the remaining KKT conditions on $(\hat{a}, \hat{\alpha})$ can be simplified to

$$\hat{a} + \frac{\hat{\alpha}}{\sqrt{R}} \geq \frac{1}{M}, \quad \hat{a} - \hat{\alpha}\sqrt{R} \geq \frac{1}{M}, \quad \sqrt{\frac{2}{R+1}}(\lambda_+ + \lambda_-, \frac{\lambda_+}{\sqrt{R}} - \lambda_-\sqrt{R}) = \frac{1}{\sqrt{\hat{a}^2 + \hat{\alpha}^2}}(\hat{a}, \hat{\alpha}).$$

Standard case-by-case checking will show that setting either $\lambda_+ = 0$ or $\lambda_- = 0$ will lead to a violation of one of the above simplified constraints; consequently, $\lambda_+, \lambda_- \neq 0$ leads to the scalar constraints being active. Solving for $\hat{a}, \hat{\alpha}$ will give the solution in the theorem statement.

So far, (94) has been shown to be a valid KKT point for the optimization problem. Our final step is proving that it is the unique optimal solution to (93). Note that the strict convexity of the 2-norm over $(\hat{a}, \hat{\alpha})$ guarantees their uniqueness. Using this fact, the primal constraints can be simplified to

$$\boldsymbol{\Psi}(\hat{\boldsymbol{a}}_+ - \frac{1}{M}\mathbf{1}) \geq \mathbf{0}, \quad \boldsymbol{\Psi}(\hat{\boldsymbol{a}}_- - \frac{1}{M}\mathbf{1}) \geq \mathbf{0} \quad \mathbf{1}^\top(\hat{\boldsymbol{a}}_+ - \frac{1}{M}\mathbf{1}) \geq 0, \quad \mathbf{1}^\top(\hat{\boldsymbol{a}}_+ - \frac{1}{M}\mathbf{1}) \geq 0.$$

The scalar constraints above are active for (94). Since the terms $\mathbf{1}^\top\hat{\boldsymbol{a}}_+, \mathbf{1}^\top\hat{\boldsymbol{a}}_-$ appear directly in the objective function as well, the scalar constraints must be active for every other minimizer. Use the fact that $\mathbf{1}^\top\boldsymbol{\Psi} = M\mathbf{1}^\top$ to conclude that while every element of the vector $\boldsymbol{\Psi}(\hat{\boldsymbol{a}}_+ - \frac{1}{M}\mathbf{1})$ is nonnegative, the sum of its elements is zero. Hence, $\boldsymbol{\Psi}(\hat{\boldsymbol{a}}_+ - \frac{1}{M}\mathbf{1}) = \mathbf{0}$ and likewise for $\hat{\boldsymbol{a}}_-$, which concludes the uniqueness argument. □

F Proofs for gradient-flow path

F.1 Proof of Theorem 5 and Corollary 1

To reduce clutter in our proof, we drop the subscript t in our expressions. For the moment, we assume a general decomposition for $\boldsymbol{W}, \boldsymbol{H}$. Suppose $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{W}_\star\boldsymbol{R}^\top$ and $\boldsymbol{H} = \boldsymbol{R}\,[\alpha_H\boldsymbol{e}_0 \quad \boldsymbol{W}_\star]\,\widetilde{\boldsymbol{V}}^\top$. Then we have $\boldsymbol{W}\boldsymbol{H} = \boldsymbol{U}\,[\boldsymbol{W}_\star\alpha_H\boldsymbol{e}_0 \quad \boldsymbol{W}_\star\boldsymbol{W}_\star]\,\widetilde{\boldsymbol{V}}^\top$ and

$$\dot{\boldsymbol{W}} = \boldsymbol{U}\,[\beta\boldsymbol{e}_0 \quad \boldsymbol{B}]\underbrace{\widetilde{\boldsymbol{V}}^\top\widetilde{\boldsymbol{V}}}_{\boldsymbol{I}}[\alpha_H e_0 \quad \boldsymbol{H}_\star]^\top\,\boldsymbol{R}^\top = \boldsymbol{U}(\beta\alpha_H\boldsymbol{e}_0\boldsymbol{e}_0^\top + \boldsymbol{B}\boldsymbol{H}_\star)\boldsymbol{R}^\top$$

$$\dot{\boldsymbol{H}} = \boldsymbol{R}\boldsymbol{W}_\star\,[\beta\boldsymbol{e}_0 \quad \boldsymbol{B}]\,\widetilde{\boldsymbol{V}}^\top = \boldsymbol{R}\,[\beta(\boldsymbol{W}_\star)_0\boldsymbol{e}_0 \quad \boldsymbol{W}_\star\boldsymbol{B}]\,\widetilde{\boldsymbol{V}}^\top.$$

So, the subspaces and structure do not change over time, and $\boldsymbol{W}_\star, \boldsymbol{W}_\star$ can be tracked directly:

$$\frac{\mathrm{d}\boldsymbol{W}_\star}{\mathrm{d}t} = \boldsymbol{B}\boldsymbol{H}_\star + \beta\alpha_H\boldsymbol{e}_0\boldsymbol{e}_0^\top, \quad \frac{\mathrm{d}\boldsymbol{H}_\star}{\mathrm{d}t} = \boldsymbol{W}_\star\boldsymbol{B}, \quad \dot{\alpha_H} = \beta\boldsymbol{W}_\star[0, 0]$$

In addition, the gradient flow equations result in

$$\frac{\mathrm{d}\boldsymbol{W}^\top\boldsymbol{W}}{\mathrm{d}t} = \boldsymbol{W}^\top(\boldsymbol{Y} - \mathbb{S}(\boldsymbol{W}\boldsymbol{H}))\boldsymbol{H} + \boldsymbol{H}(\boldsymbol{Y} - \mathbb{S}(\boldsymbol{W}\boldsymbol{H}))^\top\boldsymbol{W}$$

$$\frac{\mathrm{d}\boldsymbol{H}\boldsymbol{H}^\top}{\mathrm{d}t} = \boldsymbol{W}^\top(\boldsymbol{Y} - \mathbb{S}(\boldsymbol{W}\boldsymbol{H}))\boldsymbol{H} + \boldsymbol{H}(\boldsymbol{Y} - \mathbb{S}(\boldsymbol{W}\boldsymbol{H}))^\top\boldsymbol{W},$$

which directly leads to the conserved quantity equation $\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{W}^\top\boldsymbol{W} - \boldsymbol{H}\boldsymbol{H}^\top) = \boldsymbol{0}$. Combining this with the factorizations of $\boldsymbol{W}, \boldsymbol{H}$, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{R}(\boldsymbol{W}_\star^\top\boldsymbol{W}_\star - \alpha_H^2\boldsymbol{e}_0\boldsymbol{e}_0^\top - \boldsymbol{H}_\star\boldsymbol{H}_\star^\top)\boldsymbol{R}^\top) = \boldsymbol{0} \Rightarrow \frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{W}_\star^\top\boldsymbol{W}_\star - \alpha_H^2\boldsymbol{e}_0\boldsymbol{e}_0^\top - \boldsymbol{H}_\star\boldsymbol{H}_\star^\top) = \boldsymbol{0}.$$

In the diagonal case, we denote the diagonal matrices $\boldsymbol{W}_\star, \boldsymbol{H}_\star$ as

$$\boldsymbol{W}_\star = \operatorname{diag}(w, \boldsymbol{w}_+ \otimes \boldsymbol{e}_+ + \boldsymbol{w}_- \otimes \boldsymbol{e}_-), \quad \boldsymbol{H}_\star = \operatorname{diag}(h, \boldsymbol{h}_+ \otimes \boldsymbol{e}_+ + \boldsymbol{h}_- \otimes \boldsymbol{e}_-),$$

As a result, the dynamics in the previous section can be simplified to

$$\dot{\boldsymbol{w}}_\pm = \boldsymbol{b}_\pm \odot \boldsymbol{h}_\pm, \quad \dot{\boldsymbol{h}}_\pm = \boldsymbol{b}_\pm \odot \boldsymbol{w}_\pm, \quad \dot{w} = \beta\alpha_H + bh, \quad \dot{h} = bw, \quad \dot{\alpha_H} = \beta\widetilde{w} \tag{97}$$

In addition, the logits and the equations governing them would be

$$\boldsymbol{a}_\pm \coloneqq \boldsymbol{w}_\pm \odot \boldsymbol{h}_\pm, \quad \dot{\boldsymbol{a}}_\pm = \boldsymbol{b}_\pm \odot (\boldsymbol{w}_\pm^2 + \boldsymbol{h}_\pm^2), \tag{98}$$

$$a \coloneqq wh, \quad \alpha \coloneqq \alpha_H w, \quad \dot{a} = b(w^2 + h^2) + \beta\alpha_H h, \quad \dot{\alpha} = \beta(w^2 + \alpha_H^2) + bh\,\alpha_H. \tag{99}$$

The matrix conserved quantity equations would also simplify to

$$\frac{\mathrm{d}}{\mathrm{d}t}(w^2 - \alpha_H^2 - h^2) = 0, \quad \frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{w}_\pm^2 - \boldsymbol{h}_\pm^2) = \boldsymbol{0}.$$

Given the conserved quantities, assuming a balanced initialization would simplify the above equations to

$$\boldsymbol{a}_\pm \coloneqq \boldsymbol{w}_\pm^2 \geq \boldsymbol{0}, \quad \dot{\boldsymbol{a}}_\pm = 2\boldsymbol{a}_\pm \odot \boldsymbol{b}_\pm, \quad w^2 = \sqrt{a^2 + \alpha^2}, \tag{100}$$

$$\dot{a} = b(w^2 + \frac{a^2}{w^2}) + \beta\frac{\alpha a}{w^2} = \frac{b(2a^2 + \alpha^2) + \beta a\alpha}{\sqrt{a^2 + \alpha^2}}, \tag{101}$$

$$\dot{\alpha} = \beta(w^2 + \frac{\alpha^2}{w^2}) + b\frac{a\alpha}{w^2} = \frac{\beta(2\alpha^2 + a^2) + ba\alpha}{\sqrt{a^2 + \alpha^2}}. \tag{102}$$

F.2 Proof of Theorem 7

The proof closely follows the method of Garrod et al. (2025) for the balanced-data case. We drop the dimension subscripts from $\boldsymbol{1}_{M-1}, \boldsymbol{1}_M$ and $\overline{\boldsymbol{\Phi}}_M$ to reduce clutter in our proof. Our first step is to expand the KL divergence and its derivative as

$$D_\pm = D_{\mathrm{KL}}\left(\frac{1}{M-1}\boldsymbol{1}_{M-1} || \overline{\boldsymbol{a}_\pm}\right) = -\log(M-1) + \log(\boldsymbol{1}^\top\boldsymbol{a}_\pm) - \frac{1}{M-1}\boldsymbol{1}^\top\log(\boldsymbol{a}_\pm),$$

where we have used $\boldsymbol{a}_\pm > 0$ to simplify the norms with summations. As a result, using the dynamics equations $\dot{\boldsymbol{a}}_\pm = 2\boldsymbol{a}_\pm \odot \boldsymbol{b}_\pm$, the derivative expression is

$$\dot{D_\pm} = 2\frac{\boldsymbol{b}_\pm^\top\boldsymbol{a}_\pm}{\boldsymbol{1}^\top\boldsymbol{a}_\pm} - 2\frac{\boldsymbol{1}^\top\boldsymbol{b}_\pm}{M-1} = \frac{2}{\boldsymbol{1}^\top\boldsymbol{a}_\pm}\left(\boldsymbol{b}_\pm^\top\boldsymbol{a}_\pm - \frac{\boldsymbol{1}^\top\boldsymbol{a}_\pm}{M-1}\boldsymbol{1}^\top\boldsymbol{b}_\pm\right).$$

The expressions simplify to

$$\dot{D_+} = \frac{2MRe^{\gamma/M}}{\boldsymbol{1}^\top\boldsymbol{a}_+}\left(\frac{-\boldsymbol{g}_+^\top\overline{\boldsymbol{\Phi}}^\top\exp(\overline{\boldsymbol{\Phi}}\boldsymbol{g}_+) + \frac{1}{M-1}(\boldsymbol{1}^\top\boldsymbol{g}_+)\boldsymbol{1}^\top\overline{\boldsymbol{\Phi}}^\top\exp(\overline{\boldsymbol{\Phi}}\boldsymbol{g}_+)}{\boldsymbol{1}^\top\exp(\boldsymbol{x}_+) + M}\right)$$

$$\dot{D_-} = \frac{2Me^{\delta/M}}{\boldsymbol{1}^\top\boldsymbol{a}_-}\left(\frac{-\boldsymbol{g}_-^\top\overline{\boldsymbol{\Phi}}^\top\exp(\overline{\boldsymbol{\Phi}}\boldsymbol{g}_-) + \frac{1}{M-1}(\boldsymbol{1}^\top\boldsymbol{g}_-)\boldsymbol{1}^\top\overline{\boldsymbol{\Phi}}^\top\exp(\overline{\boldsymbol{\Phi}}\boldsymbol{g}_-)}{\boldsymbol{1}^\top\exp(\boldsymbol{x}_-) + M}\right),$$

where $\boldsymbol{g}_+ \coloneqq (\sqrt{R}M)^{-1}\boldsymbol{a}_+$, $\boldsymbol{g}_- = M^{-1}\boldsymbol{a}_-$. The above expressions suggest a similar form in the numerator of both derivatives, with strictly positive denominators and coefficients. In the next step, define $\boldsymbol{z} \coloneqq \overline{\boldsymbol{\Phi}}\boldsymbol{g}$, which directly establishes $\boldsymbol{1}^\top\boldsymbol{z} = (\boldsymbol{1}^\top\overline{\boldsymbol{\Phi}})\boldsymbol{g} = 0$. Recall that the first row of $\overline{\boldsymbol{\Phi}}$ is

$\mathbf{1}^\top$, from which $z_0 = \mathbf{1}^\top \boldsymbol{g}$ follows, and $\overline{\boldsymbol{\Phi}}\mathbf{1} = M\boldsymbol{e}_0 - \mathbf{1}$. Using these identities, the numerators can be simplified to

$$\begin{aligned}
-\boldsymbol{z}^\top \exp(\boldsymbol{z}) + \frac{z_0}{M-1}(M\boldsymbol{e}_0 - \mathbf{1})^\top \exp(\boldsymbol{z}) &= -\sum_{i=1}^{M-1} z_i e^{z_i} - \frac{z_0}{M-1}\sum_{i=1}^{M-1} e^{z_i} \\
&= -\left(\sum_{i=1}^{M-1} z_i e^{z_i} - \frac{1}{M-1}(\sum_{i=1}^{M-1} z_i)(\sum_{j=1}^{M-1} e^{z_j})\right) \\
&= -\frac{1}{2(M-1)}\sum_{i,j=1}^{M-1}(z_i - z_j)(e^{z_i} - e^{z_j})
\end{aligned}$$

Notice that each term $(z_i - z_j)(e^{z_i} - e^{z_j})$ is nonnegative regardless of the values of z_i, z_j, and vanishes if and only if $z_i = z_j$. As a result, the summation is nonnegative, with strict inequality unless $\forall i, j \geq 1 :\ z_i = z_j$. This in turn leads to to $\boldsymbol{z} = (M\boldsymbol{e}_0 - \mathbf{1})g$ for some scalar g. Using the definition $\boldsymbol{z} = \overline{\boldsymbol{\Phi}}\boldsymbol{g}$ and solving for $\boldsymbol{g}$ gives the unique answer $\boldsymbol{g} = g\mathbf{1}$ for the derivative to vanish.

Hence the KL-divergence is decreasing and serves as a Lyapunov function pushing $\boldsymbol{a}_\pm$ towards uniformity.

Finally, if $\boldsymbol{a}_{\pm,0} \propto \mathbf{1}$, then $D_{\mathrm{KL}}(\overline{\mathbf{1}} \| \overline{\boldsymbol{a}}_{\pm,0}) = 0$; since the divergence is non-negative and non-increasing, it remains zero for all $t \geq 0$, i.e., $\boldsymbol{a}_{\pm,t} \propto \mathbf{1}$.

It remains to show that $D_{\mathrm{KL}}(\overline{\mathbf{1}} \| \overline{\boldsymbol{a}}_{\pm,0}) \to 0$, rather than approaching some positive lower bound from above. Here the proof is identical for both $\pm$ and so we drop this from the notation for the rest of this section. To allow for this we define variables dependent on the choice of $\pm$

$$(s, h, g) := \begin{cases} (\sqrt{R}, \frac{\gamma}{M}, \frac{\boldsymbol{a}_+}{M\sqrt{R}}), & \text{for } \boldsymbol{a}_+, \\ (1, \frac{\delta}{M}, \frac{\boldsymbol{a}_-}{M}), & \text{for } \boldsymbol{a}_-. \end{cases}$$

Thus, in the notation used above, $x = h\mathbf{1} + \Phi g$, and $z = \Phi g$ as before.

We additionally define $r = \mathbf{1}^T g$, $\hat{g} = \frac{g}{r}$ and $c = \Psi_M \hat{g}$, which by simple Hadamard properties obeys $\sum_j c_j = M$.

Further define

$$S_0 = \sum_j e^{irc_j}, \quad S_1 = \sum_j c_j e^{-rc_j}, \quad E = e^{h+r}, \quad Q = \mathbf{1}^T e^x + M.$$

Then in terms of these variables the KL-divergence derivative is

$$\dot{D} = \frac{2sE}{Q}\left(S_1 - \frac{M}{M-1}S_0\right).$$

Now define the alternative time coordinate τ by

$$\frac{d\tau}{dt} = \frac{2sE}{Q}S_1 > 0 \tag{103}$$

Then in this alternative time coordinate

$$\frac{dD}{d\tau} = 1 - \frac{MS_0}{(M-1)S_1} =: G(r, \hat{g})$$

where we define the function $G(r, \hat{g})$ to be the right hand side. We established earlier that $G(r, \hat{g}) \leq 0$ with equality iff $\hat{g} = \frac{1}{M-1}1$.

We further note that if we keep $\hat{g}$ fixed and take a partial derivative with respects to r, this gives

$$\frac{\partial G}{\partial r} = -\frac{M}{M-1}\frac{S_0 S_2 - S_1^2}{S_1^2} \leq 0, \quad \text{where } S_2 = \sum_j c_j^2 e^{-rc_j}. \tag{104}$$

The inequality follows from a simple application of the Cauchy-Schwarz inequality.

We now establish two additional properties of the trajectory. First, the normalized coordinates remain uniformly bounded away from zero. Since $D(t) \leq D(0)$, using the definition of the KL-divergence, this gives

$$\hat{g}_i(t) \geq \prod_j \hat{g}_j(t) = \frac{1}{(M-1)^{M-1}} e^{-(M-1)D(t)} \geq \eta := \frac{1}{(M-1)^{M-1}} e^{-(M-1)D(0)} > 0$$

This further gives $M\eta \leq c_i(t) \leq 2$ due to the fact Ψ_M has entries in $\{0, 2\}$ and $\Psi_M \mathbf{1} = M\mathbf{1}$.

We next establish $r(t)$ diverges. It's derivative is given by

$$\dot{r} = 2g^T b = 2s\left(r - \frac{e^h z^T e^z}{Q}\right)$$

Next using the following relations

$$z_i = r(1 - c_i), \quad \sum_j c_j = M, \quad Q = M + E(1 + S_0)$$

The derivative simplifies to

$$\frac{\dot{r}}{r} = \frac{2s(M + ES_1)}{Q} = 2s\frac{M/E + S_1}{M/E + 1 + S_0} \tag{105}$$

Since h and r are positive, we have $\frac{M}{E} \leq M$. We also have $S_0 \leq M - 1$, giving

$$\frac{\dot{r}}{r} \geq \frac{s}{M} S_1 = \frac{s}{M} \sum_i c_i e^{-rc_i} \geq M(M-1)\eta e^{-2r}$$

and consequently

$$\frac{d}{dt} e^{2r} \geq 2s(M-1)\eta r(0)$$

and integrating this then implies $r(t) \to \infty$.

We next show that τ also diverges and so is a valid time coordinate. Using the definition of τ in Eq. (103) and the formula for $\dot{r}$ in Eq. (105) gives

$$\frac{d}{dt}(\log(r)) = \frac{2sM}{Q} + \dot{\tau} \tag{106}$$

but the dynamics of (γ, δ) given in Eq. (11) imply (for either choice of $\pm$)

$$\frac{d}{dt}(\log(h)) \geq \sqrt{2}\frac{2sM}{Q} \tag{107}$$

For example, for the positive block

$$\frac{\dot{\gamma}}{\gamma} \geq \frac{(2R\gamma^2 + \delta^2)X_+}{\sqrt{2(R\gamma^2 + \delta^2)\gamma}} \geq \sqrt{2R}X_+ = \sqrt{2}\frac{2\sqrt{R}M}{Q}.$$

Now suppose for contradiction that τ was bounded. Then by Eq. (106) and Eq. (107) we have

$$\frac{d}{dt}(\log(h) - \sqrt{2}\log(r) + \sqrt{2}\tau) \geq 0$$

Then since τ is bounded this gives

$$h(t) \geq Cr(t)^{\sqrt{2}}$$

for some positive constant C. Since $r \to \infty$, it follows that $h(t) - r(t) \to \infty$. But recalling the bound $M\eta \leq c_i \leq 2$ obtained earlier, we have

$$S_1 = \sum_j c_j e^{-rc_j} \geq (M-1)M\eta e^{-2r}$$

and consequently

$$\frac{Q}{2sM}\dot{\tau} = \frac{ES_1}{M} \geq (M-1)\eta e^{h-r} \to \infty.$$

Thus, for all sufficiently large t,

$$\dot{\tau} \geq \frac{2sM}{Q}.$$

but using Eq. (106) then gives

$$\frac{d}{dt}(\log(r)) \leq 2\dot{\tau}.$$

But if τ were bounded this forces $\log(r)$ to be bounded, a contradiction. Hence $\tau(t) \to \infty$.

We can now put this all together. Since $D(t)$ is non-increasing and bounded below by zero, it has a limit. Suppose for contradiction that $D(t) \to \epsilon > 0$. Consider the set

$$\mathcal{S} = \{\bar{a} : \epsilon \leq D_{\mathrm{KL}}(\frac{1}{M-1}\mathbf{1}||\hat{a}) \leq D(0)\}$$

This is a compact subset of the interior of the simplex, and $\frac{1}{M-1}\mathbf{1} \notin \mathcal{S}$.

Let $r_0 = r(0)$. Since $r(t) \geq r(0)$, we have by Eq. (104) that $G(r(t), \hat{g}(t)) \leq G(r_0, \hat{g}(t))$. We also have shown that the function $p \to G(r_0, p)$ is strictly negative on the compact set $\mathcal{S}$. Hence there exists some $\kappa > 0$ such that

$$G(r(t), \hat{g}(t)) \leq -\kappa$$

for all $t \geq 0$ when in $\mathcal{S}$. This gives

$$\frac{dD}{d\tau} \leq -\kappa$$

Since $\tau \to \infty$, this would eventually force $D(t)$ below the threshold ϵ, a contradiction. Hence $D(t) \to 0$. Applying these steps to both signs $\pm$ gives

$$D_{\mathrm{KL}}\left(\frac{1}{M-1}\mathbf{1}||\frac{\boldsymbol{a}_{\pm}(t)}{1^T\boldsymbol{a}_{\pm}(t)}\right) \to 0.$$

In particular, this then enforces

$$\frac{\boldsymbol{a}_{\pm}(t)}{1^T\boldsymbol{a}_{\pm}(t)} \to \frac{1}{M-1}\mathbf{1}.$$

F.3 Proof of Theorem 6

We prove the theorem by separately showing that majority emerges before minority (Proposition 13), and then, we show it emerges before maj-min (Proposition 14).

Throughout, let $R > 1$ and

$$\boldsymbol{a}_t^{\varepsilon} = \left(\gamma_t^{\varepsilon}, \delta_t^{\varepsilon}, \boldsymbol{a}_{+,t}^{\varepsilon}, \boldsymbol{a}_{-,t}^{\varepsilon}\right)$$

be the ODE trajectory in Eq. (11) under positive initialization satisfying, for some constants $C, c > 0$,

$$\|\boldsymbol{a}^{\varepsilon}(0)\|_{\infty} \leq C\varepsilon, \qquad \text{and} \qquad c\varepsilon \leq \|\boldsymbol{a}_{+}^{\varepsilon}(0)\|_{\infty}. \tag{108}$$

For any $\eta > 0$, define

$$T_{+}^{\varepsilon}(\eta) := \inf\{t \geq 0 : \|\boldsymbol{a}_{+,t}^{\varepsilon}\|_{\infty} = \eta\},$$

and

$$T_{0}^{\varepsilon}(\eta) := \inf\{t \geq 0 : a_{+,t}^{\varepsilon} = \eta\},$$

where recall that $a_{+,t}^{\varepsilon} = \frac{1}{\sqrt{2}}\sqrt{R\gamma_t^{\varepsilon}(t)^2 + \delta_t^{\varepsilon}(t)^2}$. We will show that for small enough η, there exists small enough initialization scale ϵ such that $T_{+}^{\varepsilon} < T_{-}^{\varepsilon}$ and $T_{+}^{\varepsilon} < T_{-}^{\varepsilon}$.

Proposition 13 (Majority emerges first)**.** *Assume* (108)*. Then there exists $\eta_0 > 0$ such that for every $\eta \in (0, \eta_0)$, there exists $\varepsilon_0(\eta) > 0$ for which, for all $\varepsilon \in (0, \varepsilon_0(\eta))$,*

$$T_{+}^{\varepsilon}(\eta) < T_{-}^{\varepsilon}(\eta).$$

Proposition 14 (Majority mode emerges before the maj-min mode)**.** *Assume* (108)*. Then there exists $\eta_0 > 0$ such that for every $\eta \in (0, \eta_0)$, there exists $\varepsilon_0(\eta) > 0$ such that, for all $\varepsilon \in (0, \varepsilon_0(\eta))$,*

$$T_{+}^{\varepsilon}(\eta) < T_{0}^{\varepsilon}(\eta).$$

F.3.1 AUXILIARY LEMMAS

Before proving the two main propositions, we need some auxiliary results to locally bound the coefficients $\boldsymbol{b}_\pm$ and $X_\pm$ that enter the ODE. These will allow quantifying the escape times $T_\pm, T_0$ in the next section. To reduce clutter, we drop the dimension subscripts from $\overline{\boldsymbol{\Phi}}_M, \mathbf{1}_M, \mathbf{1}_{M-1}$.

Lemma 18 (Local control of $\boldsymbol{b}_\pm$)**.** *Fix any $\rho > 0$. Then,*

$$\forall t \in [0, T_+(\rho)]\,, \qquad \mathbf{0} \le \sqrt{R}\,\mathbf{1} - \boldsymbol{b}_{+,t}(t) \le \sqrt{R}\,\tanh(\frac{\rho}{\sqrt{R}})\mathbf{1}, \tag{109a}$$

$$\forall t \in [0, T_-(\rho)]\,, \qquad \mathbf{0} \le \mathbf{1} - \boldsymbol{b}_{-,t}(t) \le \tanh(\rho)\mathbf{1}. \tag{109b}$$

Proof. For convenience, recall that

$$\sqrt{R}\mathbf{1} - \boldsymbol{b}_+ = \sqrt{R}\frac{e^{\gamma/M}\overline{\boldsymbol{\Phi}}^\top \exp(\boldsymbol{z}_+)}{e^{\gamma/M}\mathbf{1}^\top \exp(\boldsymbol{z}_+) + M}, \qquad \mathbf{1} - \boldsymbol{b}_- = \frac{e^{\delta/M}\overline{\boldsymbol{\Phi}}^\top \exp(\boldsymbol{z}_-)}{e^{\delta/M}\mathbf{1}^\top \exp(\boldsymbol{z}_-) + M},$$

where similar to the proof method in Section F.2, we have defined

$$\boldsymbol{z}_+ \coloneqq \frac{1}{M\sqrt{R}}\overline{\boldsymbol{\Phi}}\boldsymbol{a}_+, \qquad \boldsymbol{z}_- \coloneqq \frac{1}{M}\overline{\boldsymbol{\Phi}}\boldsymbol{a}_-.$$

The non-negativity bounds in Eq. (109) are proved in Lemma 7. Here, we prove the upper bounds. We will prove the bound in (109b), and the bound for (109a) follows in the exact same way and is omitted for brevity.

For each $j \in [M-1]$, define the index sets $S_j^\pm \coloneqq \{m : \overline{\boldsymbol{\Phi}}_{mj} = \pm 1\}$. By using this definition and the positivity of multiplicative scalar in $\mathbf{1} - \boldsymbol{b}_-$, we can write

$$\begin{aligned}(\mathbf{1} - \boldsymbol{b}_-)_j \le \frac{(\overline{\boldsymbol{\Phi}}\boldsymbol{e}_j)^\top \exp(\boldsymbol{z}_-)}{\mathbf{1}^\top \exp(\boldsymbol{z}_-)} &= \frac{\sum_{m\in S_j^+}\exp(\boldsymbol{z}_{-,m}) - \sum_{m\in S_j^-}\exp(\boldsymbol{z}_{-,m})}{\sum_{m\in S_j^+}\exp(\boldsymbol{z}_{-,m}) + \sum_{m\in S_j^-}\exp(\boldsymbol{z}_{-,m})} \\ &= \tanh\left(\frac{1}{2}\log\left(\frac{\sum_{m\in S_j^+}\exp(\boldsymbol{z}_{-,m})}{\sum_{m\in S_j^-}\exp(\boldsymbol{z}_{-,m})}\right)\right)\end{aligned}$$

Now, on $[0, T_+)$, using the assumptions $\boldsymbol{a}_- \ge \mathbf{0}$ and $\overline{\boldsymbol{\Phi}} \in \{\pm 1\}^{M\times(M-1)}$ having $M/2$ entries of -1 and $M/2 - 1$ entries of 1, we reach

$$\|\boldsymbol{z}_-\|_\infty = \left\|\frac{1}{M}\overline{\boldsymbol{\Phi}}\boldsymbol{a}_-\right\|_\infty \le \frac{M-1}{M}\|\boldsymbol{a}_-\|_\infty \le \rho\,.$$

Substituting the above bound will result in

$$(\mathbf{1} - \boldsymbol{b}_-)_j \le \tanh(\frac{1}{2}\log(\frac{Me^\rho}{Me^{-\rho}})) = \tanh(\rho),$$

which is the desired result. It should be noted that the above bound is tight and can be achieved by setting $\boldsymbol{a} = a\boldsymbol{e}_i$, where the bound holds with equality. □

Lemma 19 (Local control of $X_\pm$)**.** *Fix $\rho > 0$, and define the local stopping time*

$$\tau_\rho \coloneqq \inf\left\{t \ge 0 : \max\left\{\gamma_t(t), \delta_t(t), \|\boldsymbol{a}_{+,t}(t)\|_\infty, \|\boldsymbol{a}_{-,t}(t)\|_\infty\right\} = \rho\right\}.$$

Then, for every $t \in [0, \tau_\rho)$,

$$\frac{2}{1+e^{2\rho}} =: \underline{X}(\rho) \le X_{+,t}, X_{-,t} \le \overline{X}(\rho) \coloneqq \frac{2}{1+e^{-2\rho}},$$

Proof. Recall that

$$X_+ = \frac{2M}{\mathbf{1}^\top e^{\boldsymbol{x}_+} + M}, \qquad X_- = \frac{2M}{\mathbf{1}^\top e^{\boldsymbol{x}_-} + M}.$$

Fix $t \in [0, \tau_\rho)$ such that $\gamma_t, \delta_t, \|\boldsymbol{a}_{+,t}\|_\infty, \|\boldsymbol{a}_{-,t}\|_\infty < \rho$. Using $\overline{\boldsymbol{\Phi}} \in \{\pm 1\}^{M \times (M-1)}$, we have

$$\|\overline{\boldsymbol{\Phi}}\boldsymbol{a}_{\pm,t}\|_\infty \le M\|\boldsymbol{a}_{\pm,t}\|_\infty \implies \|\boldsymbol{x}_{+,t}\|_\infty \le \frac{\gamma_t}{M} + \frac{1}{M\sqrt{R}}\|\overline{\boldsymbol{\Phi}}\boldsymbol{a}_{+,t}\|_\infty \le \frac{\rho}{M} + \frac{\rho}{\sqrt{R}} \le 2\rho\,.$$

Similarly, $\|\boldsymbol{x}_{-,t}\|_\infty \le 2\rho$. Therefore

$$Me^{-2\rho} \le \mathbf{1}^\top e^{\boldsymbol{x}_{\pm,t}} \le Me^{2\rho}.$$

Substituting this into the definitions of $X_\pm$ gives the desired. □

F.3.2 Proof of Proposition 13

Recall that

$$\dot{\boldsymbol{a}}_\pm = 2\boldsymbol{a}_\pm \odot \boldsymbol{b}_\pm.$$

Fix $\rho > 0$ and assume first that $\eta \in (0, \rho)$. By Lemma 18, for every $t \in [0, T^\varepsilon_+(\rho)]$,

$$\|\boldsymbol{b}_{+,t} - \sqrt{R}\,\mathbf{1}\|_\infty \le \sqrt{R}\,\tanh\left(\frac{\rho}{\sqrt{R}}\right) \le \sqrt{R}\,\tanh(\rho).$$

Hence, for every coordinate i and every $t \in [0, T^\varepsilon_+(\rho)]$,

$$\sqrt{R}\big(1 - \tanh(\rho)\big) \le b_{+,i}(t) \le \sqrt{R},$$

and therefore

$$2\sqrt{R}\big(1 - \tanh(\rho)\big)a_{+,i}(t) \le \dot{a}_{+,i}(t) \le 2\sqrt{R}\,a_{+,i}(t).$$

Applying Gronwall's inequality coordinatewise and taking maximum over i gives, for every $t \in [0, T^\varepsilon_+(\rho)]$,

$$\|\boldsymbol{a}^\varepsilon_{+,0}\|_\infty e^{2\sqrt{R}(1-\tanh(\rho))t} \le \|\boldsymbol{a}^\varepsilon_{+,t}\|_\infty \le \|\boldsymbol{a}^\varepsilon_{+,0}\|_\infty e^{2\sqrt{R}t}. \tag{110}$$

In the exact same manner, for every $t \in [0, T^\varepsilon_-(\rho)]$,

$$\|\boldsymbol{a}^\varepsilon_{-,0}\|_\infty e^{2(1-\tanh(\rho))t} \le \|\boldsymbol{a}^\varepsilon_{-,t}\|_\infty \le \|\boldsymbol{a}^\varepsilon_{-,0}\|_\infty e^{2t}. \tag{111}$$

Because $\eta < \rho$, we have

$$T^\varepsilon_+(\eta) \le T^\varepsilon_+(\rho), \qquad T^\varepsilon_-(\eta) \le T^\varepsilon_-(\rho),$$

and therefore (110) may be evaluated at $t = T^\varepsilon_+(\eta)$, while (111) may be evaluated at $t = T^\varepsilon_-(\eta)$. Doing this, using

$$\|\boldsymbol{a}^\varepsilon_+(T^\varepsilon_+(\eta))\|_\infty = \eta, \qquad \|\boldsymbol{a}^\varepsilon_-(T^\varepsilon_-(\eta))\|_\infty = \eta,$$

and rearranging gives

$$\frac{1}{2\sqrt{R}}\log\frac{\eta}{\|\boldsymbol{a}^\varepsilon_+(0)\|_\infty} \le T^\varepsilon_+(\eta) \le \frac{1}{2\sqrt{R}\big(1-\tanh(\rho)\big)}\log\frac{\eta}{\|\boldsymbol{a}^\varepsilon_+(0)\|_\infty}, \tag{112}$$

$$\frac{1}{2}\log\frac{\eta}{\|\boldsymbol{a}^\varepsilon_-(0)\|_\infty} \le T^\varepsilon_-(\eta) \le \frac{1}{2\big(1-\tanh(\rho)\big)}\log\frac{\eta}{\|\boldsymbol{a}^\varepsilon_-(0)\|_\infty}, \tag{113}$$

Because $R > 1$ and $\tanh(\rho) \to 0$ as $\rho \to 0$, we may choose $\eta_0 > 0$ such that for every $\rho \in (0, \eta_0)$, $\sqrt{R}\big(1 - \tanh(\rho/M)\big) > 1$. Fix any $\eta \in (0, \eta_0)$ and choose any $\rho \in (\eta, \eta_0)$. Then (112), (113), together with

$$c\varepsilon \le \|\boldsymbol{a}^\varepsilon_+(0)\|_\infty, \qquad \|\boldsymbol{a}^\varepsilon_-(0)\|_\infty \le \|\boldsymbol{a}^\varepsilon(0)\|_\infty \le C\varepsilon,$$

would imply

$$T^\varepsilon_+(\eta) \le \frac{1}{2\sqrt{R}(1-\tanh(\rho))}\log\frac{\eta}{c\varepsilon}, \qquad T^\varepsilon_-(\eta) \ge \frac{1}{2}\log\frac{\eta}{C\varepsilon}.$$

By the condition $\sqrt{R}\big(1 - \tanh(\rho)\big) > 1$, the upper bound for $T^\varepsilon_+(\eta)$ has a strictly smaller coefficient in front of $\log(1/\varepsilon)$ than the lower bound for $T^\varepsilon_-(\eta)$. Therefore, for all sufficiently small ε, $T^\varepsilon_+(\eta) < T^\varepsilon_-(\eta)$, as desired.

F.3.3 Proof of Proposition 14

We suppress the superscript ε when clear from context. Recall that

$$\dot{\gamma} = \frac{(2R\gamma^2 + \delta^2)X_+ + \gamma\delta X_-}{\sqrt{2(R\gamma^2 + \delta^2)}}, \qquad \dot{\delta} = \frac{(R\gamma^2 + 2\delta^2)X_- + R\gamma\delta X_+}{\sqrt{2(R\gamma^2 + \delta^2)}}.$$

By direct differentiation,

$$\dot{a}_0(t) = \frac{R\gamma(t)\dot{\gamma}(t) + \delta(t)\dot{\delta}(t)}{2a_0(t)}.$$

Substituting the formulas for $\dot{\gamma}, \dot{\delta}$, and simplifying using $2a_0(t)^2 = R\gamma(t)^2 + \delta(t)^2$, yields

$$\dot{a}_0(t) = R\gamma(t)X_{+,t} + \delta(t)X_{-,t}. \tag{114}$$

By Lemma 19, for every $t \in [0, \tau_\rho)$,

$$X_{+,t}, X_{-,t} \le \overline{X}(\rho), \qquad \overline{X}(\rho) = \frac{2}{1 + e^{-2\rho}}.$$

Hence, on $[0, \tau_\rho)$,

$$\dot{a}_0(t) \le \overline{X}(\rho)\big(R\gamma(t) + \delta(t)\big) \le \sqrt{2(R+1)}\,\overline{X}(\rho)\,a_0(t), \tag{115}$$

where the second inequality follows from $R\gamma + \delta \le \sqrt{R+1}\sqrt{R\gamma^2 + \delta^2} = \sqrt{2(R+1)}\,a_0$.

Next, by Lemma 18, exactly as in the proof of Proposition 13,

$$\|\boldsymbol{a}_{+,t}(t)\|_\infty \ge \|\boldsymbol{a}_+(0)\|_\infty e^{2\sqrt{R}(1-E(\rho))t}, \qquad t \in [0, T_+(\rho)]. \tag{116}$$

Let

$$\mu(\rho) := 2\sqrt{R}(1 - \tanh(\rho)), \qquad \nu(\rho) := \sqrt{2(R+1)}\,\overline{X}(\rho).$$

Since $\tanh(\rho) \to 0$ and $\overline{X}(\rho) \to 1$ as $\rho \to 0$, we have

$$\mu(\rho) \to 2\sqrt{R}, \qquad \nu(\rho) \to \sqrt{2(R+1)}.$$

Because $R > 1$,

$$2\sqrt{R} > \sqrt{2(R+1)}.$$

Hence there exists $\rho_{\mathrm{rate}} > 0$ such that

$$\mu(\rho) > \nu(\rho) \qquad \text{for all } \rho \in (0, \rho_{\mathrm{rate}}). \tag{117}$$

By Proposition 13, there exists $\rho_{\mathrm{maj}} > 0$ such that its conclusion holds for every threshold $\rho \in (0, \rho_{\mathrm{maj}})$. Define

$$\rho_0 := \min\{\rho_{\mathrm{rate}}, \rho_{\mathrm{maj}}\}, \qquad \eta_0 := \rho_0/2.$$

Fix any $\eta \in (0, \eta_0)$, and set

$$\rho := 2\eta.$$

Then $\rho \in (0, \rho_0)$, so both (117) and Proposition 13 apply at threshold ρ.

We claim that $T_+^\varepsilon(\eta) < T_0^\varepsilon(\eta)$ for all sufficiently small ε. Suppose, toward a contradiction, that

$$T_0^\varepsilon(\eta) \le T_+^\varepsilon(\eta),$$

and set

$$T := T_0^\varepsilon(\eta).$$

For sufficiently small ε, the initialization satisfies

$$\|\boldsymbol{a}_\pm^\varepsilon(0)\|_\infty < \eta, \qquad a_0^\varepsilon(0) < \eta.$$

Since $0 < \eta < \rho$, continuity gives

$$T_+^\varepsilon(\eta) < T_+^\varepsilon(\rho).$$

Hence $T < T_+^\varepsilon(\rho)$. Moreover, by Proposition 13, after possibly decreasing $\varepsilon_0(\eta)$,

$$T_+^\varepsilon(\rho) < T_-^\varepsilon(\rho).$$

Therefore

$$T < T_+^\varepsilon(\rho), \qquad T < T_-^\varepsilon(\rho).$$

Also, for every $t \le T$, we have $a_0(t) \le \eta$, and hence, since $\rho = 2\eta$:

$$\gamma(t) \le \sqrt{\frac{2}{R}}\eta < \rho, \qquad \delta(t) \le \sqrt{2}\,\eta < \rho,$$

Thus

$$T < \tau_\rho\,, \tag{118}$$

and we can plug in T to (115) and (116).

Specifically, By (115), (118), and Gronwall's inequality,

$$a_0(T) \le a_0(0)e^{\nu(\rho)T}.$$

Since $a_0(T) = \eta$, this implies

$$e^T \ge \left(\frac{\eta}{a_0(0)}\right)^{1/\nu(\rho)}. \tag{119}$$

Using (116), (119), and $\|\boldsymbol{a}_+(0)\|_\infty \ge c\varepsilon$, we get $\|\boldsymbol{a}_{+,t}(T)\|_\infty \ge c\varepsilon \left(\frac{\eta}{a_0(0)}\right)^{\mu(\rho)/\nu(\rho)}$. Since

$$a_0(0) \le C\sqrt{\frac{R+1}{2}}\,\varepsilon,$$

we obtain

$$\|\boldsymbol{a}_{+,T}\|_\infty \ge c\left(C\sqrt{\frac{R+1}{2}}\right)^{-\mu(\rho)/\nu(\rho)} \eta^{\mu(\rho)/\nu(\rho)}\varepsilon^{1-\mu(\rho)/\nu(\rho)}. \tag{120}$$

By (117), $\mu(\rho) > \nu(\rho)$. Hence the exponent $1 - \mu(\rho)/\nu(\rho)$ is negative, so the right-hand side of (120) diverges as $\varepsilon \to 0$. In particular, for all sufficiently small ε,

$$\|\boldsymbol{a}_{+,T}\|_\infty > \eta.$$

But $T = T_0^\varepsilon(\eta) \le T_+^\varepsilon(\eta)$, so by definition of $T_+^\varepsilon(\eta)$, $\|\boldsymbol{a}_{+,T}\|_\infty \le \eta$, a contradiction. Therefore $T_+^\varepsilon(\eta) < T_0^\varepsilon(\eta)$ completing the proof.

F.4 Lyapunov function counterexample

Here, we provide a counterexample showing that the KL distance cannot be a Lyapunov function under our diagonal-plus-rank-one parameterization even for $R = 1$. This is in contrast to the diagonal parameterization for $R = 1$ by (Garrod et al., 2025). Of course, as made clear by our work, the diagonal-plus-rank-one parameterization is necessary to capture the RP and the GF evolution in the general imbalanced case.

Proposition 15 (Counterexample to KL monotonicity)**.** *Let $R = 1$ and $K = 4$. Define*

$$\boldsymbol{z} = \left[\sqrt{\frac{\gamma^2+\delta^2}{2}}, a_+, a_-\right], \qquad \bar{\boldsymbol{z}} := \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_1}, \qquad \bar{\mathbf{1}} := \frac{1}{3}\mathbf{1}_3.$$

Then $D_{\mathrm{KL}}(\bar{\mathbf{1}}\|\bar{\boldsymbol{z}})$ is not necessarily decreasing along the ODE. More precisely, take

$$\gamma = \kappa\varepsilon, \qquad \delta = \varepsilon, \qquad a_+ = a_- = \varepsilon,$$

with $0 < \kappa < 1$ and $\varepsilon > 0$. Then, for all sufficiently small $\varepsilon > 0$, $\dot{D}_{\mathrm{KL}}(\bar{\mathbf{1}}\|\bar{\boldsymbol{z}}) > 0$.

Proof. For $K = 4$, $R = 1$, we have

$$\boldsymbol{x}_+ = \frac{1}{2}\begin{bmatrix}\gamma + a_+\\ \gamma - a_+\end{bmatrix}, \qquad \boldsymbol{x}_- = \frac{1}{2}\begin{bmatrix}\delta + a_-\\ \delta - a_-\end{bmatrix}, \quad X_+ = \frac{4}{\mathbf{1}^\top e^{x_+} + 2}, \qquad X_- = \frac{4}{\mathbf{1}^\top e^{x_-} + 2},$$

and

$$b_+ = 1 - \frac{X_+}{4}\begin{bmatrix}1 & -1\end{bmatrix}e^{\boldsymbol{x}_+}, \qquad b_- = 1 - \frac{X_-}{4}\begin{bmatrix}1 & -1\end{bmatrix}e^{\boldsymbol{x}_-}.$$

As in the proposition's statement, let $\gamma = \kappa\varepsilon, \delta = \varepsilon, a_+ = a_- = \varepsilon$, and

$$c := \sqrt{\frac{\kappa^2+1}{2}}. \tag{121}$$

Then $\boldsymbol{z} = \varepsilon \begin{bmatrix} c & 1 & 1 \end{bmatrix}^\top$. Hence

$$\bar{\boldsymbol{z}} = \frac{\boldsymbol{z}}{\|\boldsymbol{z}\|_1} = \frac{1}{c+2}\begin{bmatrix} c \\ 1 \\ 1 \end{bmatrix}.$$

Let $V(z) := D_{\mathrm{KL}}(\bar{\mathbf{1}} \| \bar{\boldsymbol{z}})$. With $r_i := \dot{z}_i / z_i$, we have

$$\dot{V} = \sum_{i=0}^{2} \bar{z}_i r_i - \frac{1}{3}\sum_{i=0}^{2} r_i \tag{122}$$

As $\varepsilon \to 0^+$, we have

$$\boldsymbol{x}_+ \to 0, \qquad \boldsymbol{x}_- \to 0,$$

and therefore

$$X_+ \to 1, \qquad X_- \to 1, \qquad b_+ \to 1, \qquad b_- \to 1.$$

Consequently,

$$r_1 := \frac{\dot{a}_+}{a_+} = 2b_+ \to 2, \qquad r_2 := \frac{\dot{a}_-}{a_-} = 2b_- \to 2.$$

Moreover, calling $z_0 = \sqrt{\frac{\gamma^2+\delta^2}{2}}$, we have

$$\frac{\dot{z}_0}{z_0} = \frac{\gamma\dot{\gamma} + \delta\dot{\delta}}{\gamma^2+\delta^2},$$

and substituting $\gamma = \kappa\varepsilon, \delta = \varepsilon$, and using $X_+, X_- \to 1$, we obtain

$$r_0 := \frac{\dot{z}_0}{z_0} \to \frac{\kappa+1}{c}.$$

Thus, from Eq. (122)

$$\lim_{\varepsilon\to 0^+} \dot{V} = \frac{c(\kappa+1)/c + 2 + 2}{c+2} - \frac{1}{3}((\kappa+1)/c + 2 + 2) = \frac{\kappa+5}{c+2} - \frac{1}{3}\left(\frac{\kappa+1}{c} + 4\right).$$

To see that this limit is strictly positive for $\kappa \in (0,1)$, we can rearrange the inequality $\lim_{\varepsilon\to 0^+} \dot{V} > 0$ to $c(\kappa+3) > \kappa^2+\kappa+2$. Squaring both sides and substituting $c^2 = (\kappa^2+1)/2$ reduces the condition to:

$$\kappa^4 - 2\kappa^3 + 2\kappa - 1 < 0 \implies (\kappa+1)(\kappa-1)^3 < 0\,.$$

Since $\kappa \in (0,1)$, this inequality strictly holds, completing the proof. □

G Proof of Theorem 8

G.1 Auxiliary results

We begin by upper and lower bounding a scalar variable $x(t) \geq 0$ satisfying differential inequalities of the form $c_1 x e^{-x} \leq \dot{x} \leq c_2 x e^{-x}$. We will later use the following lemmas in our analysis of the dynamics.

Lemma 20. *Let $x(t) \in \mathbb{R}$ satisfy the differential inequality*

$$\dot{x}(t) \geq c\, x(t) e^{-x(t)},\ c > 0, \quad x(0) = x_0 > 0.$$

The following bounds can be established for all $t \geq 0$:

$$x(t) \geq \log(e^{x_0} + c x_0 t) \tag{123}$$

$$\begin{aligned} x(t) &\geq \log\left(e^{x_0} + \frac{c x_0 t}{1+x_0}\log(e^{x_0} + c x_0 t)\right) \\ &\geq \log\left(1 + c x_0 t \log(1 + c x_0 t)\right) - \log(1+x_0) \end{aligned} \tag{124}$$

The same bounds hold for x in case we replace $\dot{x}$ with D_+x when x is non-differentiable but its right derivative $D_+x(t)$ exists; that is,

$$D_+x(t) \geq c\, x(t) e^{-x(t)},\ c > 0, \quad x(0) = x_0 > 0.$$

Proof. We drop the explicit dependence of time from $x(t)$ for notational simplicity. Integrating the differential gives

$$ct \leq I(x, x_0), \quad I(x, x_0) \coloneqq \int_{x_0}^{x} \frac{du}{u} e^u,$$

and we subsequently upper bound $I(x, x_0)$. The simpler lower bound (123) can be reached from

$$ct \leq I(x, x_0) = \int_{x_0}^{x} \frac{du}{u} e^u \leq \frac{1}{x_0} \int_{x_0}^{x} e^u du = \frac{e^x - e^{x_0}}{x_0} \Rightarrow x(t) \geq \log(e^{x_0} + cx_0 t).$$

However, we reach a sharper lower bound for $x(t)$, by noting that

$$I(x, x_0) = e^x \int_0^{x - x_0} \frac{e^{-y}}{x - y} dy, \quad \frac{1}{x - y} \leq \frac{\frac{1}{x_0} - \frac{1}{x}}{x - x_0} y + \frac{1}{x} = \frac{y}{x x_0} + \frac{1}{x},\ \forall y \in [0, x - x_0],$$

where the last inequality follows from the convexity of $(x - y)^{-1}$ over y in the interval $[0, x - x_0]$. Substituting the convexity inequality in the integration leads to

$$ct \leq \frac{e^x}{x} \left(\int_0^{x - x_0} e^{-y} (1 + \frac{y}{x_0}) dy \right) = \frac{1}{x} \left((e^x - e^{x_0})(1 + \frac{1}{x_0}) - (x - x_0) e^{x_0} \right).$$

We can subsequently use $x - x_0 \geq 0$ and the logarithmic lower bound (123) to reach the desired result of (124):

$$(e^x - e^{x_0})(1 + x_0) \geq ctx_0 x \geq ctx_0 \log(e^{x_0} + ctx_0),$$

For the case of nondifferentiable x with an existing right derivative, we can lower bound the trajectory as

$$x(t) \geq y(t), \qquad \dot{y}(t) = c\, y(t) e^{-y(t)},\ c > 0, \quad y(0) = x_0 > 0,$$

after which the lower bounds of (123) and (124) apply to $y(t)$ and by extension, $x(t)$. □

Lemma 21. *Let $x \in \mathbb{R}$ satisfy*

$$0 \leq \dot{x} \leq cxe^{-x},\ c > 0, \quad x(0) = x_0 > 0.$$

Then $x(t)$ satisfies the upper bound

$$x(t) \leq \log(\frac{e^{x_0}}{x_0} + ct) + \log\log(\frac{e^{x_0}}{x_0} + ct) + \log\frac{e}{e - 1}. \tag{125}$$

Proof. Similar to the previous lemma, we integrate the differential inequality. However, this time we lower bound the integral $I(x, x_0)$ as

$$ct \geq I(x, x_0) = \int_{x_0}^{x} \frac{du}{u} e^u \geq \frac{1}{x} \int_{x_0}^{x} du e^u = \frac{e^x - e^{x_0}}{x} \geq \frac{e^x}{x} - \frac{e^{x_0}}{x_0},$$
$$\Rightarrow x - \log x \leq \log(\frac{e^{x_0}}{x_0} + ct),$$

where we have used $u^{-1} \geq x^{-1}$ for $u \in [x_0, x]$ and $x \geq x_0$. By using the standard inequality $\log x \leq x/e$, we reach

$$x \leq (1 - \frac{1}{e})^{-1} \log(\frac{e^{x_0}}{x_0} + ct).$$

Substituting the above bound into the logarithmic inequality leads to the desired result:

$$x(t) \leq \log(\frac{e^{x_0}}{x_0} + ct) + \log\left((1 - \frac{1}{e})^{-1} \log(\frac{e^{x_0}}{x_0} + ct) \right)$$
$$= \log(\frac{e^{x_0}}{x_0} + ct) + \log\log(\frac{e^{x_0}}{x_0} + ct) + \log\frac{e}{e - 1}.$$

□

G.2 Logarithmic growth

We are now able to analyze the gradient flow dynamics in the limit of $t \to \infty$. The following lemma establishes a lower bound growth rate on the norms of $\|\boldsymbol{a}_\pm\|_\infty$ by only assuming $\|\boldsymbol{a}_\pm(0)\|_\infty > 0$:

Proposition 16. *Consider the gradient flow dynamics of (11), initialized such that $\gamma, \delta > 0$ and $\boldsymbol{a}_\pm > \mathbf{0}$. Define $A_\pm(t) \coloneqq \|\boldsymbol{a}_\pm(t)\|_\infty$ and $A_{\pm,0} = A_\pm(0) > 0$. Then, the norms grow at least as $\log t + \log\log t$. More specifically,*

$$A_+ \geq \frac{\sqrt{R}}{2}\left(\log\left(1 + 4A_{+,0}t\log(1+4A_{+,0}t)\right) - \log(1+\frac{2A_{+,0}}{\sqrt{R}})\right) \tag{126a}$$

$$A_- \geq \frac{1}{2}\left(\log\left(1 + 4A_{-,0}t\log(1+4A_{-,0}t)\right) - \log(1+2A_{-,0})\right) \tag{126b}$$

Proof. The local control lemma of Lemma 18 can be written alternatively as

$$\sqrt{R}\left(1 - \tanh(\frac{\|\boldsymbol{a}_+\|_\infty}{\sqrt{R}})\right)\mathbf{1} \leq \boldsymbol{b}_{+,t}(t) \leq \sqrt{R}\,\mathbf{1}, \tag{127a}$$

$$\left(1 - \tanh(\|\boldsymbol{a}_-\|_\infty)\right)\mathbf{1} \leq \boldsymbol{b}_{-,t}(t) \leq \mathbf{1}. \tag{127b}$$

Note that by using $1 - \tanh(x) = \frac{2}{e^{2x}+1} \geq e^{-2x}$ for $x \geq 0$, the control bounds can be simplified to

$$\sqrt{R}e^{-\frac{2\|\boldsymbol{a}_+\|_\infty}{\sqrt{R}}}\mathbf{1} \leq \boldsymbol{b}_{+,t}(t) \leq \sqrt{R}\,\mathbf{1}, \tag{128a}$$

$$e^{-2\|\boldsymbol{a}_-\|_\infty}\mathbf{1} \leq \boldsymbol{b}_{-,t}(t) \leq \mathbf{1}. \tag{128b}$$

The control bounds for $\boldsymbol{b}_\pm$ only depend on $\|\boldsymbol{a}_\pm\|_\infty$; as a result, we can use the alongside $\dot{\boldsymbol{a}}_\pm = 2\boldsymbol{a}_\pm \odot \boldsymbol{b}_\pm$ to control the growth of $\|\boldsymbol{a}_\pm\|_\infty$. We focus on the lower bounds, as we establish later that the upper control bounds for $\boldsymbol{b}_\pm$ are too loose to characterize the asymptotic dynamics.

Using $\dot{\boldsymbol{a}}_\pm = 2\boldsymbol{a}_\pm \odot \boldsymbol{b}_\pm \geq 0$ and $\boldsymbol{a}_\pm \geq 0$, we reach

$$D_+ \log\|\boldsymbol{a}_\pm(t)\|_\infty = \max_{i\in I(t)} \frac{\mathrm{dlog}(\boldsymbol{a}_\pm)_i}{\mathrm{d}t} = 2\max_{i\in I(t)}(\boldsymbol{b}_\pm)_i,$$

where D_+ is the right derivative operator, and $I(t)$ is the set maximizer indices for $\|\boldsymbol{a}_\pm\| = \max_i a_i$. Substituting the uniform lower bounds for $\boldsymbol{b}_\pm$ leads to

$$D_+A_+ \geq 2\sqrt{R}A_+e^{-\frac{2A_+}{\sqrt{R}}}, \quad D_+A_- \geq 2A_-e^{-2A_-}.$$

Consequently, after rescaling $A_\pm$, we can lower bound the growth of $A_\pm$ with Lemma 20, which directly leads to (126), as desired. □

G.3 Asymptotics

Notice that Theorem 7 shows attraction in direction towards $\boldsymbol{a}_\pm \propto \mathbf{1}$. Therefore, we naturally expect the dynamics for $\boldsymbol{a}_\pm = a_\pm\mathbf{1}$ to determine the asymptotic behavior. Motivated by this, we focus on the initialization $\boldsymbol{a}_\pm(0) = a_\pm(0)\mathbf{1}_{M-1}$, for which the dynamics reduce to the four scalar variables $(\gamma, \delta, a_+, a_-)$. For this case, we prove that $a_{\pm,t} = \Theta(\log t)$ and that the majority-to-minority ratio $a_{+,t}/(\sqrt{R}\,a_{-,t})$ converges to its RP-limit value 1. Along the way, we also establish the asymptotic lower bounds $\liminf_{t\to\infty}\sqrt{R}\,\gamma_t/a_{+,t} \geq 1$ and $\liminf_{t\to\infty}\delta_t/a_{-,t} \geq 1$ (Proposition 17); matching upper bounds, which would establish convergence of the full logit direction to the SEL geometry, remain open. For convenience, we define the rescaled variables $\hat{\boldsymbol{a}}_+ = \frac{1}{\sqrt{R}}\boldsymbol{a}_+$, $\hat{\boldsymbol{a}}_- = \boldsymbol{a}_-$ throughout this section.

G.3.1 Simplified equations

We begin with a simplification of the parameters that enter the logit dynamics in Corollary 1, which we recall here for convenience:

$$\dot{\gamma} = \frac{(2R\gamma^2 + \delta^2)\,X_+ + \gamma\delta\,X_-}{\sqrt{2(R\gamma^2+\delta^2)}}, \quad \dot{\delta} = \frac{(R\gamma^2 + 2\delta^2)\,X_- + R\gamma\delta\,X_+}{\sqrt{2(R\gamma^2+\delta^2)}}, \quad \dot{\boldsymbol{a}}_\pm = 2\,\boldsymbol{a}_\pm \odot \boldsymbol{b}_\pm. \tag{129}$$

Lemma 22. *Assume $\gamma_0, \delta_0 > 0$ and $\hat{\boldsymbol{a}}_{\pm,0} > \mathbf{0}$. If $\boldsymbol{a}_\pm \propto \mathbf{1}_{M-1}$ at any time T, it also holds true at all times $t \geq T$. Therefore, if $\hat{\boldsymbol{a}}_{\pm,0} = \hat{a}_{\pm,0}\mathbf{1}$, only the scalars $\gamma, \delta, \hat{a}_\pm$ evolve with time and are increasing. In addition, $\boldsymbol{b}_\pm = b_\pm \mathbf{1}_{M-1}$ and the parameters $b_\pm, X_\pm$ simplify to*

$$
\begin{aligned}
X_+ &= \frac{2M}{M + e^{\frac{\gamma - \hat{a}_+}{M}}(e^{\hat{a}_+} + M - 1)}, \quad X_- = \frac{2M}{M + e^{\frac{\delta - \hat{a}_-}{M}}(e^{\hat{a}_-} + M - 1)}, \\
b_+ &= \sqrt{R}\frac{X_+}{2}(1 + e^{\frac{\gamma - \hat{a}_+}{M}}) = \sqrt{R}\frac{M(1 + e^{\frac{\gamma - \hat{a}_+}{M}})}{M + e^{\frac{\gamma - \hat{a}_+}{M}}(e^{\hat{a}_+} + M - 1)} \\
b_- &= \frac{X_-}{2}(1 + e^{\frac{\delta - a_-}{M}}) = \frac{M(1 + e^{\frac{\delta - a_-}{M}})}{M + e^{\frac{\delta - \hat{a}_-}{M}}(e^{\hat{a}_-} + M - 1)}
\end{aligned}
\tag{130}
$$

Proof. Recall the expressions for $\boldsymbol{x}_\pm$, $\boldsymbol{b}_\pm$ and $X_\pm$ from (5):

$$
(\boldsymbol{x}_+, \boldsymbol{x}_-) = \frac{1}{M}\big((\gamma, \delta)\mathbf{1}_M + \overline{\boldsymbol{\Phi}}_M(\hat{\boldsymbol{a}}_+, \hat{\boldsymbol{a}}_-)\big) \quad X_\pm = \frac{2M}{M + \mathbf{1}_M^\top \exp(\boldsymbol{x}_\pm)}
$$
$$
\boldsymbol{b}_+ = \sqrt{R}\left(\mathbf{1}_{M-1} - \frac{X_+}{2M}\overline{\boldsymbol{\Phi}}_M^\top \exp(\boldsymbol{x}_+)\right), \quad \boldsymbol{b}_- = \mathbf{1}_{M-1} - \frac{X_-}{2M}\overline{\boldsymbol{\Phi}}_M^\top \exp(\boldsymbol{x}_-).
$$

By assuming $\hat{\boldsymbol{a}}_\pm = \hat{a}_\pm \mathbf{1}_{M-1}$ for time t and recalling $\overline{\boldsymbol{\Phi}}_M \mathbf{1}_{M-1} = M\boldsymbol{e}_0 - \mathbf{1}_M$, we can simplify the exponential expressions as

$$
\begin{aligned}
\exp(\boldsymbol{x}_+) &= e^{\frac{\gamma - \hat{a}_+}{M}}((e^{\hat{a}_+} - 1)\boldsymbol{e}_0 + \mathbf{1}_M), \quad \exp(\boldsymbol{x}_-) = e^{\frac{\delta - \hat{a}_-}{M}}((e^{\hat{a}_-} - 1)\boldsymbol{e}_0 + \mathbf{1}_M), \\
\overline{\boldsymbol{\Phi}}_M^\top \exp(\boldsymbol{x}_+) &= e^{\frac{\gamma - \hat{a}_+}{M}}(e^{\hat{a}_+} - 1)\mathbf{1}, \quad \overline{\boldsymbol{\Phi}}_M^\top \exp(\boldsymbol{x}_-) = e^{\frac{\delta - \hat{a}_-}{M}}(e^{\hat{a}_-} - 1)\mathbf{1}, \\
\mathbf{1}_M^\top \exp(\boldsymbol{x}_+) &= e^{\frac{\gamma - \hat{a}_+}{M}}(e^{\hat{a}_+} + M - 1), \quad \mathbf{1}_M^\top \exp(\boldsymbol{x}_-) = e^{\frac{\delta - \hat{a}_-}{M}}(e^{\hat{a}_-} + M - 1),
\end{aligned}
$$

after which we reach $\boldsymbol{b}_\pm = b_\pm \mathbf{1}_M$ and the expressions in (130). Since $\dot{\boldsymbol{a}}_\pm = 2\boldsymbol{a}_\pm \odot \boldsymbol{b}_\pm$, $\boldsymbol{a}_\pm = a_\pm \mathbf{1}_{M-1}$ and $\boldsymbol{b}_\pm = b_\pm \mathbf{1}_{M-1}$, we conclude that the direction of $\boldsymbol{a}_\pm$ remains proportional to $\mathbf{1}_M$. Since $X_\pm \geq 0, \boldsymbol{b}_\pm \geq \mathbf{0}$, $\gamma, \delta, a_\pm$ are increasing in time, completing our proof. □

G.3.2 Growth bounds for majority and minority logits

Our next step is to bound the growth of $a_\pm$ in the spirit of Proposition 16. Here, the simplified equations in Lemma 22 enable us to provide a tighter lower bound in addition to an upper bound, relative to relying only on the infinity norm as in Proposition 16. The following lemma makes this claim precise.

Lemma 23 (Crude growth bounds on $a_\pm$)**.** *Consider the setting of Lemma 22. Define*

$$
C_{+,t} = M\sqrt{R}\left(4t + \frac{e^{\frac{a_{+,0}}{\sqrt{R}}(1 - \frac{1}{M})}}{(M-1)a_{+,0}}\right), \quad \text{and} \quad C_{-,t} = M\left(4t + \frac{e^{a_{-,0}(1 - \frac{1}{M})}}{(M-1)a_{-,0}}\right).
$$

The following upper and lower bounds hold for $a_\pm$:

$$
a_+ \geq \sqrt{R}\left(\log\left(1 + 2a_{+,0}\, t \log(1 + 2a_{+,0}\, t)\right) - \log(1 + \frac{a_{+,0}}{\sqrt{R}})\right) \tag{131a}
$$
$$
a_- \geq \log\left(1 + 2a_{-,0}\, t \log(1 + 2a_{-,0}\, t)\right) - \log(1 + a_{-,0}) \tag{131b}
$$
$$
a_+ \leq \frac{\sqrt{R}\,1}{1 - 1/M}\left(\log(C_{+,t}) + \log\log(C_{+,t}) + \log(\frac{e}{e-1})\right) \tag{131c}
$$
$$
a_- \leq \frac{1}{1 - 1/M}\left(\log(C_{-,t}) + \log\log(C_{-,t}) + \log(\frac{e}{e-1})\right). \tag{131d}
$$

Proof. We prove the bounds (131b) and (131d) here. The proof for $\hat{a}_+$ mirrors $\hat{a}_-$, with γ replacing δ throughout the expressions, and $\sqrt{R}$ factors appearing through rescaling to a_+.

Following the spirit of Proposition 16, we upper bound $1-b_-$, and we additionally upper bound b_-. Recall the expressions

$$b_- = \frac{M(1+e^{\frac{\delta-\hat{a}_-}{M}})}{M+e^{\frac{\delta-\hat{a}_-}{M}}(e^{\hat{a}_-}+M-1)}, \quad 1-b_- = \frac{e^{\frac{\delta-\hat{a}_-}{M}}(e^{\hat{a}_-}-1)}{M+e^{\frac{\delta-\hat{a}_-}{M}}(e^{\hat{a}_-}+M-1)}.$$

Upper bounding $1-b_-$ by using $M \geq 2$ results in

$$1-b_- \leq \frac{e^{\hat{a}_-}-1}{e^{\hat{a}_-}+M-1} \Rightarrow b_- \geq \frac{M}{e^{\hat{a}_-}+M-1} \geq \frac{2}{e^{\hat{a}_-}+1} \geq e^{-\hat{a}_-}. \tag{132}$$

On the other side, upper bounding b_- directly gives

$$b_- \leq \frac{M(e^{\frac{a_- -\delta}{M}}+1)}{e^{a_-}+M-1} \leq Me^{-\hat{a}_-}(e^{\frac{a_- -\delta}{M}}+1) \tag{133a}$$

$$\leq Me^{-\hat{a}_-}(e^{\frac{\hat{a}_-}{M}}+1) \leq 2Me^{-a_-(1-\frac{1}{M})}, \tag{133b}$$

where we used $\delta \geq 0$ (from the initialization assumptions of $\gamma_0, \delta_0 > 0$ in Lemma 22). Substituting (132) and (133b), and the corresponding bounds for b_+ in $\dot{\hat{a}}_\pm = 2\hat{a}_\pm b_\pm$ gives

$$2\sqrt{R}\,\hat{a}_+\,e^{-\hat{a}_+} \leq \dot{\hat{a}}_+ \leq 4\sqrt{R}\,M\hat{a}_+e^{-\hat{a}_+(1-\frac{1}{M})}, \qquad 2\hat{a}_-e^{-\hat{a}_-} \leq \dot{\hat{a}}_- \leq 4M\hat{a}_-e^{-\hat{a}_-(1-\frac{1}{M})}.$$

Substituting (125) and (124) from Lemma 21 and Lemma 20, and rescaling recovers the desired equations. □

G.3.3 Majority and minority logits converge in direction to their SEL limit

So far, we are able to prove that $a_+ = \Theta(a_-) = \Theta(\log t)$ and

$$1-\frac{1}{M} \leq \liminf_{t\to\infty} \frac{a_+}{\sqrt{R}a_-}, \quad \limsup_{t\to\infty} \frac{a_+}{\sqrt{R}a_-} \leq \frac{1}{1-1/M}.$$

First, note that lower bounding by the infinity norm in Proposition 16 is less tight by a factor of $\frac{1}{2}$ in the asymptotic limit relative to what we achieved above. As a second note, we intuitively expect δ, γ to be of the same order as $\hat{a}_\pm$ based on the RP characterization and the conjectured Eq. (12). However, using the crude bound $\delta \geq 0$ to reach (133b) from (133a) created an $1-\frac{1}{M}$ instead, which in turn resulted in a $1-\frac{1}{M}$ mismatch between our upper and lower asymptotic bounds. Therefore, our next step is to characterize the relation between $\gamma, \hat{a}_+$ and $\delta, \hat{a}_-$, respectively, with the aim of refining this bounding step.

Lemma 24 (Relative growth lower bounds for γ, δ vs $a_\pm$)**.** *Consider the gradient flow settings of Lemma 22 with $\gamma_0, \delta_0, a_{\pm,0} > 0$. The quantities $\log a_\pm, \log\gamma, \log\delta$ are well-defined and satisfy*

$$\frac{\mathrm{d}}{\mathrm{d}\log\hat{a}_+}\log(\frac{\gamma}{\hat{a}_+}) \geq \frac{\sqrt{2}}{1+e^{\frac{\gamma-\hat{a}_+}{M}}} - 1 \quad \textit{and} \quad \frac{\mathrm{d}}{\mathrm{d}\log\hat{a}_-}\log(\frac{\delta}{\hat{a}_-}) \geq \frac{\sqrt{2}}{1+e^{\frac{\delta-\hat{a}_-}{M}}} - 1. \tag{134}$$

Proof. Recall the differential equations for γ, δ:

$$\dot{\gamma} = \frac{(2R\gamma^2+\delta^2)X_+ + \gamma\delta X_-}{\sqrt{2(R\gamma^2+\delta^2)}}, \qquad \dot{\delta} = \frac{(R\gamma^2+2\delta^2)X_- + R\gamma\delta X_+}{\sqrt{2(R\gamma^2+\delta^2)}}.$$

As a result, γ, δ are increasing with time and strictly positive, and have well-defined logarithms. Using $X_\pm \geq 0$ allows to reach

$$\frac{\mathrm{d}\log\gamma}{\mathrm{d}t} \geq \frac{A_0^2+R\gamma^2}{\sqrt{2}A_0\gamma}X_+ \geq \frac{2\sqrt{R}\gamma A_0}{\sqrt{2}A_0\gamma}X_+ = \frac{\sqrt{2}}{1+e^{\frac{\gamma-\hat{a}_+}{M}}}(2b_+),$$

$$\frac{\mathrm{d}\log\delta}{\mathrm{d}t} \geq \frac{A_0^2+\delta^2}{\sqrt{2}A_0\delta}X_- \geq \frac{2\delta A_0}{\sqrt{2}A_0\delta}X_- = \frac{\sqrt{2}}{1+e^{\frac{\delta-\hat{a}_-}{M}}}(2b_-),$$

where $A_0 \coloneqq \sqrt{R\gamma^2+\delta^2}$. We know that $a_\pm$ are positive and increasing with time; as a result, $\log a_\pm$ is also well-defined. The desired inequalities follow from $\frac{\mathrm{d}\log a_\pm}{\mathrm{d}t} = 2b_\pm$ and rearranging. □

Proposition 17. *For every $\epsilon \in (0,1)$, there exists a finite time $T_\pm(\epsilon) < \infty$ such that*

$$\delta \geq (1-\epsilon)\hat{a}_- \ \forall t \geq T_-(\epsilon), \qquad \gamma \geq (1-\epsilon)\hat{a}_+ \ \forall t \geq T_+(\epsilon). \tag{135}$$

Consequently, $\liminf \frac{\delta}{\hat{a}_-} \geq 1$ *and* $\liminf \frac{\gamma}{\hat{a}_+} \geq 1$

Proof. Fix any finite $0 < c < \sqrt{2} - 1$. For every desired $\epsilon \in (0,1)$, denote $\tau_\pm$ as the smallest time that $-M/\epsilon \log(\frac{\sqrt{2}}{1+c} - 1) < \hat{a}_\pm$. Such a time exists as we know that $a_\pm$ is positive, is increasing with time, and diverges by the lower bounds (131b) and (131a). The rest of the proof considers only $\delta, \hat{a}_-$. The argument for $\gamma, \hat{a}_+$ mirrors the case of $\delta, \hat{a}_-$ and is omitted for brevity. At time τ_-, we consider the two cases $\delta \geq (1-\epsilon)\hat{a}_-$ and $\delta < (1-\epsilon)\hat{a}_-$ separately:

1. The relation $\delta \geq (1-\epsilon)\hat{a}_-$ already holds at τ_-. In this case, we show that it holds for all $t \geq \tau_-$. Consider any time $t' \geq \tau_-$ when $\delta/\hat{a}_-)$ has decreased to $1-\epsilon$. By the definition of τ_-,
$$\frac{\mathrm{d}}{\mathrm{d}\log \hat{a}_-} \log(\frac{\delta}{\hat{a}_-})\Big|_{t'} \geq \frac{\sqrt{2}}{1+\exp(\frac{-\epsilon \hat{a}_-}{M})} - 1 \geq c > 0,$$
hence, the ratio increases again, and moves away from $1-\epsilon$. As a result, $\delta \geq (1-\epsilon)\hat{a}_-$ for all $t \geq \tau_-$.

2. In the opposite case, $\delta < (1-\epsilon)\hat{a}_-$ at τ_-. This time, instead of lower bounding at a specific time, we conclude
$$\frac{\mathrm{d}}{\mathrm{d}\log \hat{a}_-} \log(\frac{\delta}{\hat{a}_-}) \geq \frac{\sqrt{2}}{1+\exp(\frac{-\epsilon \hat{a}_-}{M})} - 1 \geq c > 0, \qquad \forall t \geq \tau_- : \delta < (1-\epsilon)\hat{a}_-.$$
Integrating the above up to a time T_- when $\delta = (1-\epsilon)\hat{a}_-$ yields
$$\log(1-\epsilon) - \log(\frac{\delta}{\hat{a}_-})\Big|_{\tau_-} \geq c \log \frac{\hat{a}_{-,T_-}}{\hat{a}_{-,\tau_-}},$$
which can be used alongside the bounds for $\hat{a}_-$ to provide an estimate for T_- in terms of τ_-.

□

We are now ready to prove Theorem 8.

Proof of Theorem 8. By using the result of Proposition 17, we refine the bounding step from (133a) to (133b) in Lemma 23. As in the statement of Proposition 17, there exists a finite time $T_-(\epsilon)$ for which $\delta \geq (1-\epsilon)\hat{a}_- \ \forall t \geq T_-(\epsilon)$. Substituting that in (133a) yields

$$b_- \leq Me^{-\hat{a}_-}(e^{\frac{\epsilon}{M}\hat{a}_-} + 1) \leq 2Me^{-(1-\frac{\epsilon}{M})\hat{a}_-} \qquad \forall t \geq T_-(\epsilon)$$
$$b_+ \leq \sqrt{R}Me^{-\hat{a}_+}(e^{\frac{\epsilon}{M}\hat{a}_+} + 1) \leq 2\sqrt{R}Me^{-(1-\frac{\epsilon}{M})\hat{a}_+} \qquad \forall t \geq T_+(\epsilon).$$

Repeating all subsequent steps of Lemma 23 to obtain an upper bound gives

$$a_+ \leq \frac{\sqrt{R}}{1-\epsilon/M}\left(\log(C'_{+,t}) + \log\log(C'_{+,t}) + \log(\frac{e}{e-1})\right) \tag{136a}$$
$$a_- \leq \frac{1}{1-\epsilon/M}\left(\log(C'_{-,t}) + \log\log(C'_{-,t}) + \log(\frac{e}{e-1})\right) \tag{136b}$$

where $C'_{+,t} = M\sqrt{R}\left(4t + \frac{e^{\frac{1}{\sqrt{R}}(1-\frac{\epsilon}{M})a_{+,T_+}}}{(M-1)a_{+,T_+}}\right), \quad C'_{-,t} = M\left(4t + \frac{e^{a_{-,T_-}(1-\frac{\epsilon}{M})}}{(M-1)a_{-,T_-}}\right).$

Note that $C'_{\pm,t}$ use the values of $a_\pm$ at $T_\pm(\epsilon)$, as opposed to $C_{\pm,t}$ in (131c) and (131d). Nevertheless, for $t \gg T_\pm(\epsilon)$, the above bounds simplify to

$$a_+ \leq \frac{\sqrt{R}}{1-\epsilon/M}\left(\log t + \log\log t + O(1)\right) \qquad a_- \leq \frac{1}{1-\epsilon/M}\left(\log t + \log\log t + O(1)\right).$$

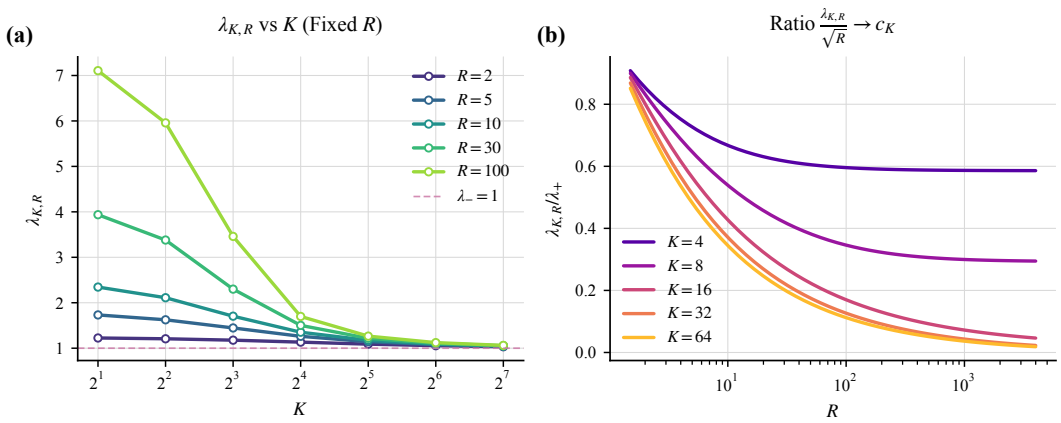


Figure 2: Behavior of regularization-path emergence thresholds for varying number of classes (K) and imbalance ratios (R).

Combining the resulting upper bounds with the lower bounds (131a) (131b) yields

$$1 - \frac{\epsilon}{M} \leq \liminf_{t\to\infty} \frac{a_+}{\sqrt{R}a_-}, \quad \limsup_{t\to\infty} \frac{a_+}{\sqrt{R}a_-} \leq \frac{1}{1-\epsilon/M}, \quad \forall \epsilon \in (0,1).$$

As we take the limit of $\epsilon \to 0^+$ in the above expression, the $\limsup$ and $\liminf$ bounds become equal to 1. Consequently, the limit of $a_+/\sqrt{R}a_-$ as $t \to \infty$ exists and is equal to 1. □

H Numerical results

H.1 Regularization-path experiments

In this section, we provide additional numerical results supporting our analysis. We write σ_+, σ_-, and $\sigma_{+/-}$ for the singular values corresponding, respectively, to the majority, minority, and majority–minority modes. Since these singular values diverge along the CE regularization path as $\lambda \to 0$, we compare their relative evolution after factoring out both the overall scale and the limiting SEL profile.

Specifically, denote in this section the singular values of the SEL matrix as

$$s_+^\star = \sqrt{R}, \qquad s_-^\star = 1, \qquad s_{+/-}^\star = \sqrt{\frac{R+1}{2}},$$

and define the path-dependent scale

$$C_\lambda := \frac{|\sigma_{+,\lambda}| + |\sigma_{-,\lambda}| + |\sigma_{+/-,\lambda}|}{s_+^\star + s_-^\star + s_{+/-}^\star}.$$

We then plot the rescaled coordinates

$$\widehat{\sigma}_{+,\lambda} := \frac{\sigma_{+,\lambda}}{C_\lambda \sqrt{R}}, \qquad \widehat{\sigma}_-(\lambda) := \frac{\sigma_{-,\lambda}}{C_\lambda}, \qquad \widehat{\sigma}_{+/-,\lambda} := \frac{\sigma_{+/-,\lambda}}{C_\lambda \sqrt{(R+1)/2}}.$$

With this normalization, all three coordinates converge to 1 whenever the singular-value vector converges in direction to the SEL direction. Thus, the plots emphasize deviations from the limiting geometry rather than the overall divergence of the singular values.

H.1.1 Regularization-path comparison for different K and R

We complement the theoretical analysis with numerical comparisons of three ways of tracing the regularization path. For each value of λ, we overlay the singular values predicted by our analytic solution, the singular values obtained by solving the convex logit-space problem (Z-UFM$_\lambda$) in CVXPY, and the singular values obtained by projected gradient descent on the sphere of radius B

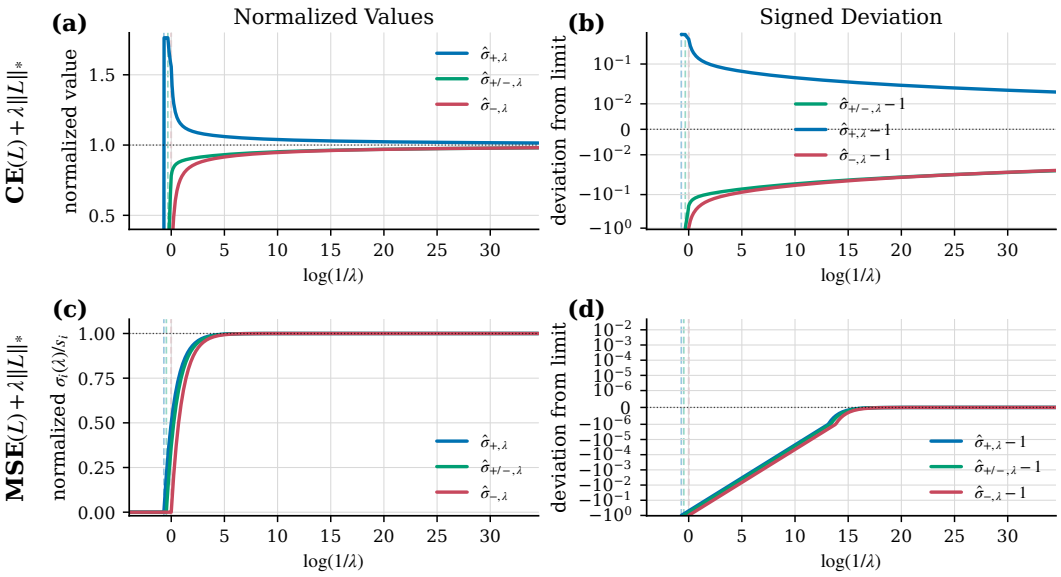
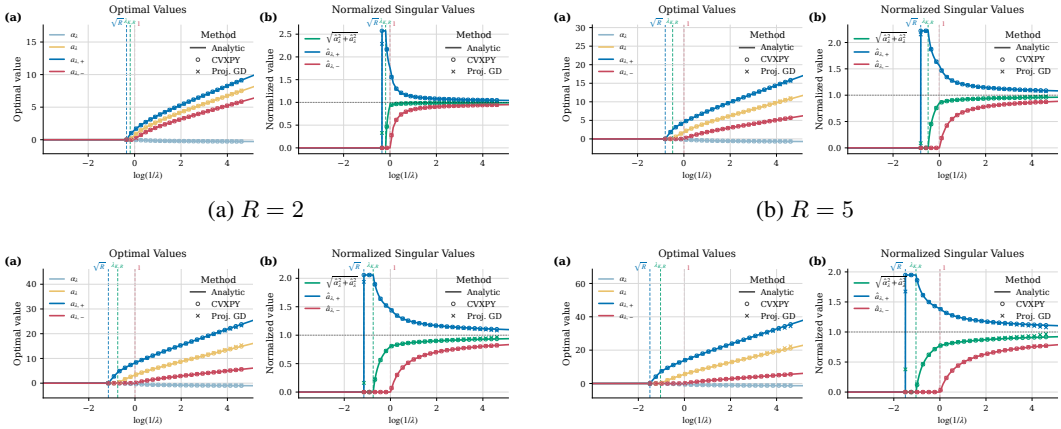


Figure 3: Comparison of the regularization-path trajectories for the three normalized singular coordinates under CE and MSE, shown for $R = 4$ and $K = 8$. The panels track the majority, majority–minority, and minority coordinates as functions of λ; horizontal dashed lines indicate their limiting normalized values



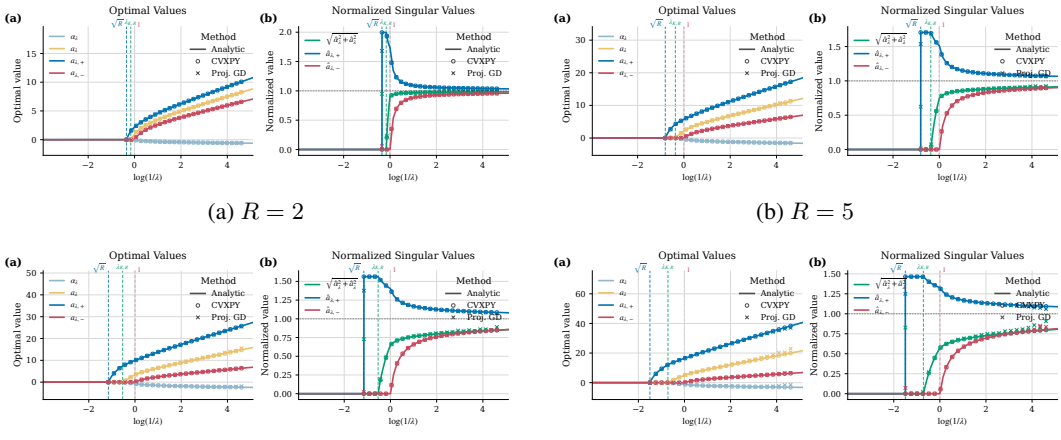
(a) $R = 2$ (b) $R = 5$

(c) $R = 10$ (d) $R = 20$

Figure 4: Numerical comparison of the regularization-path singular values for $K = 4$ across different imbalance ratios R. In each panel, the curves obtained from the analytic solution are overlaid with those produced by a direct CVXPY solve of (Z-UFM$_\lambda$) and by projected gradient descent on the sphere with radius B matched to the same value of λ.

corresponding to that value of λ. This comparison serves two purposes: it verifies the correctness of the analytic formulas, and it shows that the projected optimization procedure tracks the same path across a range of imbalance ratios and class counts. The resulting comparisons are shown in Figures 4 and 5.

Beyond powers of two. Our presentation in the main text assumes $K = 2^m$, which allows us to use Sylvester-Hadamard matrices and keeps the spectral coordinates especially explicit. This assumption is convenient for the proof, but the resulting regularization-path formulas depend only on the three-mode structure induced by the step imbalance: majority, majority–minority, and minority. This raises the possibility that the same path description remains valid more generally, even when K



(a) $R = 2$ (b) $R = 5$

(c) $R = 10$ (d) $R = 20$

Figure 5: Numerical comparison of the regularization-path singular values for $K = 8$ across different imbalance ratios R. In each panel, the curves obtained from the analytic solution are overlaid with those produced by a direct CVXPY solve of (Z-UFM$_\lambda$) and by projected gradient descent on the sphere with radius B matched to the same value of λ.

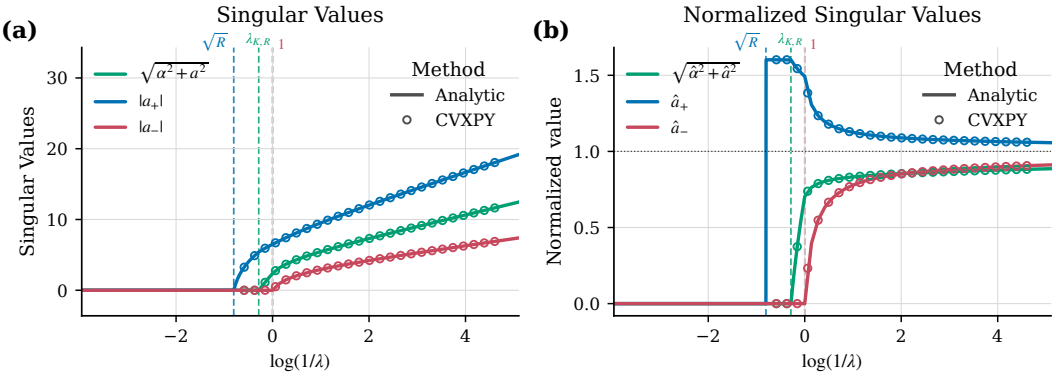


Figure 6: Regularization-path validation for $K = 12$ and $R = 5$. The figure compares the singular-coordinate trajectories predicted by the analytic formulas with those obtained from a direct CVXPY solve of the convex nuclear-norm problem (Z-UFM$_\lambda$). The numerical agreement supports the claim that the regularization-path description is not an artifact of the Sylvester-Hadamard coordinate construction.

is not a power of two and the particular Hadamard construction is unavailable. As a numerical check, we compare in Figure 6 the analytic singular-coordinate trajectories with a direct CVXPY solution of the convex nuclear-norm problem for $K = 12$ and $R = 5$. The close agreement provides evidence that the regularization-path characterization is not merely an artifact of the Sylvester-Hadamard coordinates.

Role of the nuclear norm. The phase transitions identified in the regularization path are not a generic consequence of imbalance alone. To illustrate this, we repeat the same path computation after replacing the nuclear-norm penalty in (Z-UFM$_\lambda$) by a Frobenius-norm penalty in logit space. The resulting trajectories are shown in Figure 7 for both CE and MSE losses.

The contrast with the nuclear-norm path is clear. Under Frobenius regularization, the penalty controls the overall logit scale but does not separately threshold individual singular modes. As a result, once the solution leaves zero, the singular coordinates grow together in their limiting proportions. After normalization by the limiting singular-value direction, all coordinates remain essentially constant at one along the path. Thus, the staged activation observed for the nuclear-norm regularization path disappears.

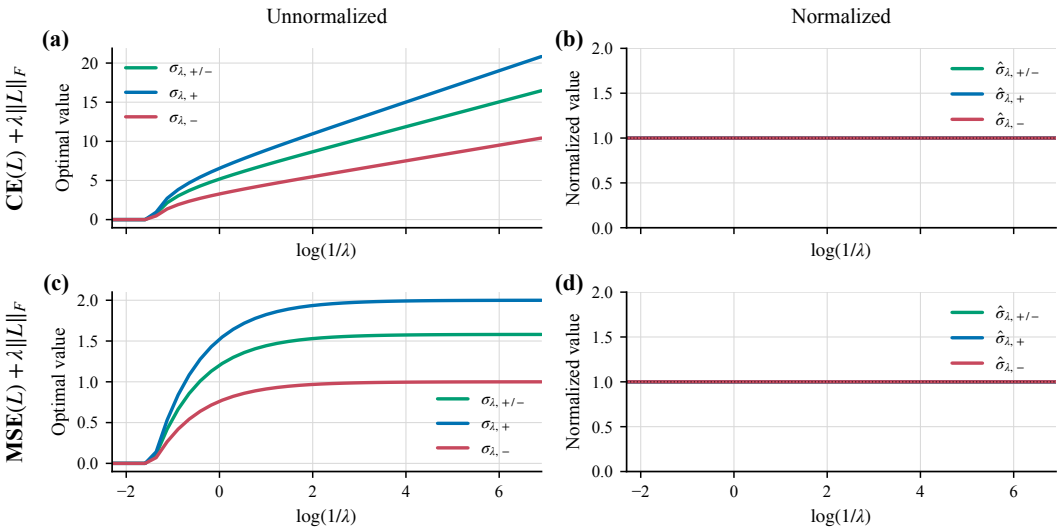


Figure 7: Regularization paths with a Frobenius-norm penalty in logit space. The top row shows the CE objective with Frobenius regularization, and the bottom row shows the corresponding MSE objective. Left panels plot the unnormalized singular coordinates, while right panels plot the same coordinates after normalization by their limiting direction. In contrast to the nuclear-norm path, the normalized coordinates remain aligned throughout the trajectory, indicating that the modes do not emerge sequentially.

This comparison highlights the mechanism behind mode emergence in our analysis. The nuclear norm acts as an ℓ_1-type penalty on the singular values, allowing different modes to become active at different regularization strengths. By contrast, the Frobenius norm behaves as an ℓ_2-type penalty on the singular-value vector, coupling the modes through a common scale and thereby suppressing sequential emergence.

Relative convergence of the minority and majority–minority modes. The order in which modes emerge does not, by itself, determine how quickly active modes approach their limiting normalized values. To compare these two notions, let $e_-(L) := |\widehat{\sigma}_{-,\lambda} - 1|$ and $e_{+/-}(L) := |\widehat{\sigma}_{+/-,\lambda} - 1|$, where $L = \log(1/\lambda)$, and record when each error enters within a fixed tolerance of zero. As shown in Figure 8, the ordering depends on K and R. For $K = 4$ and $R = 10$, the majority–minority mode enters the 5% tolerance region at approximately $L = 7.0$, before the minority mode at $L = 22.7$. For $K = 16$ and the same R, the ordering reverses: the minority mode enters at approximately $L = 19.3$, whereas the majority–minority mode does so only at $L = 64.7$. The same ordering holds for these two examples at the 10% and 2% tolerances as well. Thus, although the minority mode is the last to become active along the regularization path, it can subsequently approach the limiting SEL direction faster than the majority–minority mode.

H.2 Gradient-flow experiments

H.2.1 Setup

We numerically simulate the gradient flow in (2) using gradient descent with stepsize 10^{-2}. Throughout this experiment we set $K = 4$. We compare two initialization schemes.

Hadamard initialization. We initialize the model according to the spectrally aligned construction in Theorem 5, using randomly chosen initial singular coordinates. For this initialization, the matrix-valued gradient flow is predicted to remain in the invariant subspace induced by the $(\boldsymbol{U}, \widetilde{\boldsymbol{V}})$ coordinates, and its evolution should therefore be exactly captured by the reduced vector ODE derived in Theorem 5. We simulate both the original gradient descent dynamics and the reduced ODE from the same initial coordinates.

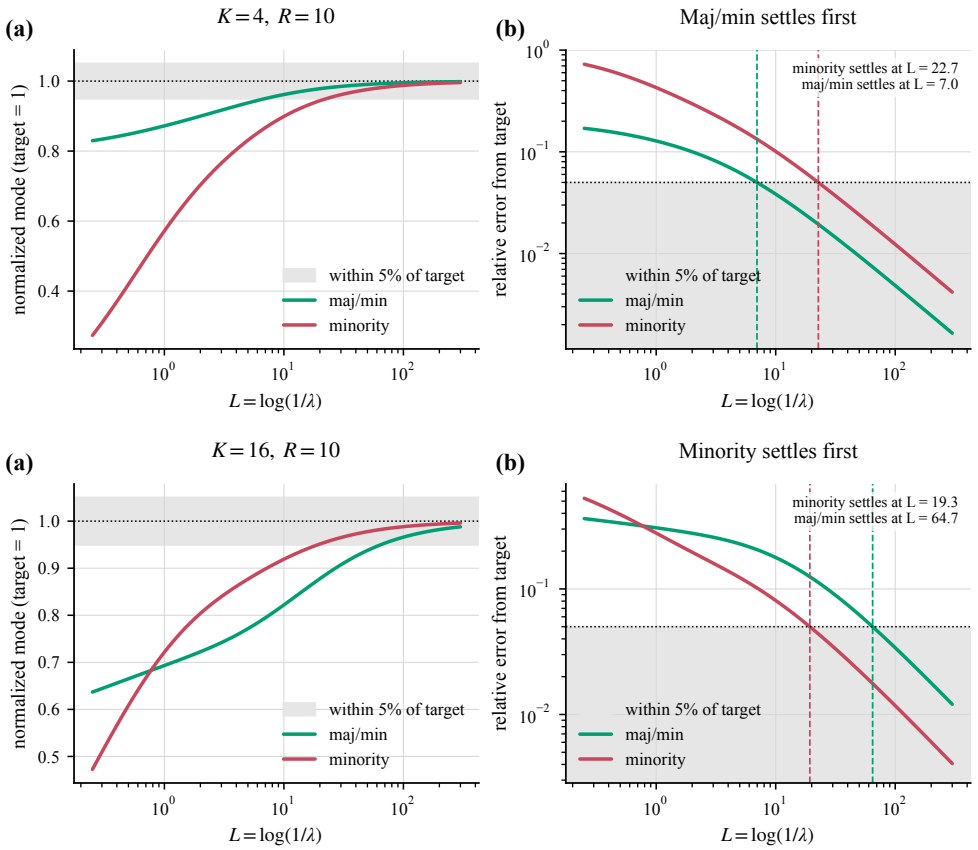


Figure 8: Dependence of the relative convergence ordering on K at fixed imbalance ratio $R = 10$. *Top:* for $K = 4$, the majority–minority mode settles first. *Bottom:* for $K = 16$, the minority mode settles first. In each plot, the left panel shows the normalized coordinates approaching their common target value one. The right panel shows their absolute relative errors; the shaded region denotes error at most 5%, and each colored dashed line marks the first value of $L = \log(1/\lambda)$ after which the corresponding mode remains in that region.

Random spectral initialization. As a comparison, we also initialize $\boldsymbol{W}$ and $\boldsymbol{H}$ with the same initial singular values as in the Hadamard initialization, but with random left and right singular directions. This keeps the initial scale and spectrum fixed while removing the spectral alignment required by the theory.

All results are averaged over 10 independent runs. Shaded regions indicate the standard error of the mean. Metrics are recorded at 92 geometrically spaced iterations, which allows the full trajectory to be visualized on a logarithmic time axis.

Metrics. We track six quantities. First, we record the cross-entropy loss. Second, we measure neural-collapse-style alignment through the Frobenius distances between the normalized Gram matrices of $\boldsymbol{W}$, $\boldsymbol{H}$, and $\boldsymbol{W}\boldsymbol{H}$ and their corresponding limiting SELI geometries. Third, we measure the relative residual energy outside the signal subspace associated with the Hadamard coordinates,

$$\frac{\|\boldsymbol{L} - \boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{L}\,\widetilde{\boldsymbol{V}}\widetilde{\boldsymbol{V}}^\top\|_F}{\|\boldsymbol{L}\|_F},$$

where $\boldsymbol{L} = \boldsymbol{W}\boldsymbol{H}$ denotes the logit matrix. This quantity is zero exactly when the logits lie in the reduced subspace. Fourth, we track the singular coordinates of the logits. Fifth, we plot their ℓ_1-normalized versions, which removes the diverging scale and isolates convergence in direction. Finally, we report the KL divergence between the normalized singular-coordinate profile and the limiting SELI singular-value profile. In the present imbalanced setting, this KL quantity is used only

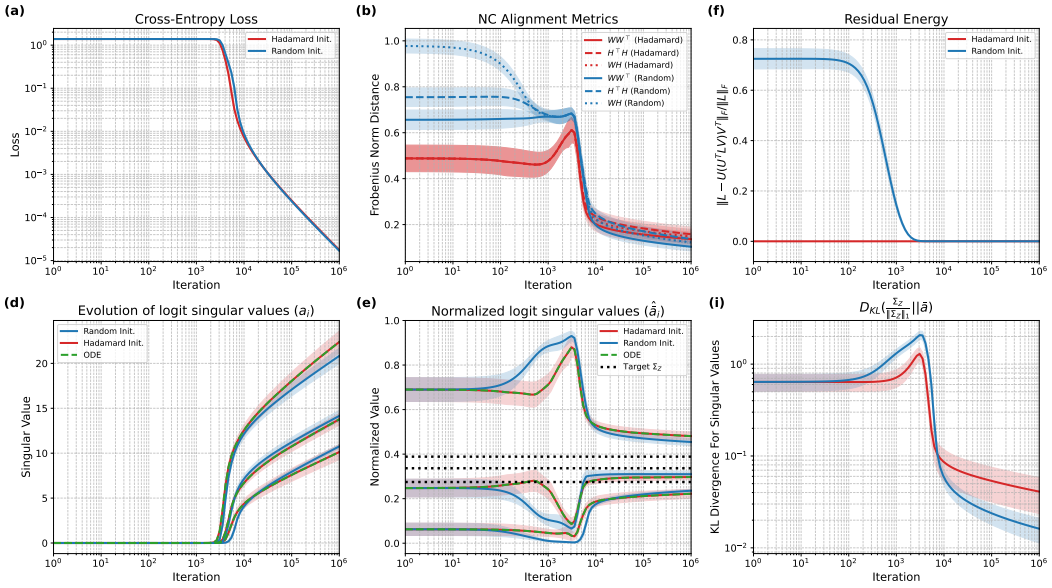


Figure 9: Gradient-flow experiments for $K = 4$ and $R = 2$. We compare Hadamard initialization, random spectral initialization with the same initial singular values, and the reduced vector ODE initialized from the Hadamard coordinates. The panels show the cross-entropy loss, neural-collapse-style alignment metrics, residual energy outside the Hadamard signal subspace, raw singular-coordinate trajectories, ℓ_1-normalized singular-coordinate trajectories, and the KL discrepancy between the normalized singular-coordinate profile and the limiting SELI profile. Shaded regions denote standard errors over 10 runs.

as a directional discrepancy measure; unlike in the balanced setting, it is not a Lyapunov function for the dynamics.

H.2.2 Results

The results are summarized in Figure 9. They show three main trends.

Reduction to the vector ODE. Under Hadamard initialization, the residual energy outside the $(\boldsymbol{U}, \widetilde{\boldsymbol{V}})$ signal subspace remains at its initial value of zero throughout training. This verifies the invariance predicted by the softmax diagonalization result in Theorem 1: the Hadamard coordinates close under the gradient-flow dynamics, so the matrix-valued trajectory does not leave the reduced subspace.

Validation of the reduced dynamics. The singular-coordinate trajectories obtained from the full gradient descent simulation under Hadamard initialization agree with the solution of the reduced vector ODE. This agreement is visible both for the raw singular coordinates and for their normalized versions. Thus, the reduced ODE in Theorem 5 accurately captures the evolution of the original matrix-valued system when the initialization is spectrally aligned.

Behavior beyond the invariant initialization. Random spectral initialization does not begin in the Hadamard signal subspace, and therefore initially has substantial residual energy outside that subspace. Nevertheless, this residual energy decreases during training, and the normalized singular-coordinate profile moves toward the same limiting SELI profile. This suggests that the Hadamard-reduced dynamics capture not only an exactly invariant initialization, but also a useful organizing description of the late-time behavior of more generic initializations. The KL diagnostic exhibits a transient increase before decreasing, which is consistent with the fact that it is not a Lyapunov function in this imbalanced setting.

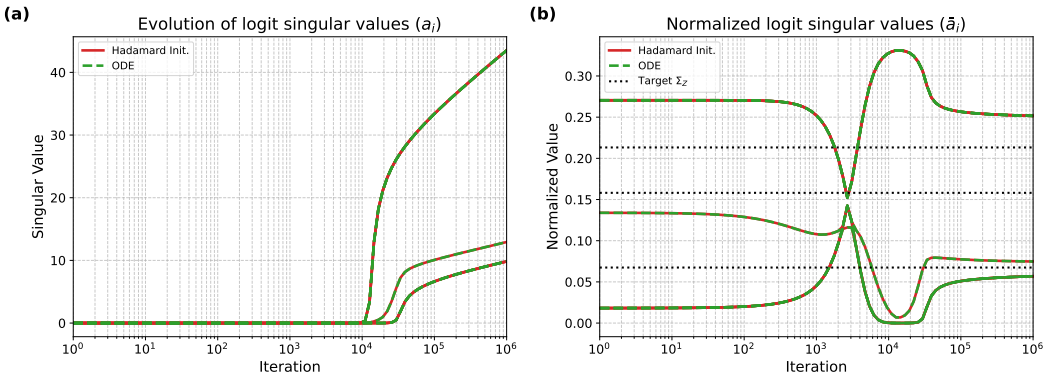


Figure 10: Preservation of block-constant spectral coordinates under gradient flow. The model uses $K = 8$ and $R = 10$, and is initialized with $\boldsymbol{a}_{+,0} = a_{+,0}\mathbf{1}_{M-1}$ and $\boldsymbol{a}_{-,0} = a_{-,0}\mathbf{1}_{M-1}$. The trajectories show that the coordinates within each majority and minority block remain identical throughout training, confirming preservation of the block-constant subspace. The same run also exhibits the qualitative phase structure observed along the regularization path: majority modes emerge first, followed by the majority–minority mode, and then the minority modes.

H.2.3 Preservation of block-constant spectral coordinates

The reduced dynamics in Theorem 5 still contain vector-valued coordinates

$$\boldsymbol{a}_{+,t}, \boldsymbol{a}_{-,t} \in \mathbb{R}^{M-1},$$

corresponding to the majority and minority singular blocks. A particularly important special case is the block-constant initialization

$$\boldsymbol{a}_{+,0} = a_{+,0}\mathbf{1}_{M-1}, \qquad \boldsymbol{a}_{-,0} = a_{-,0}\mathbf{1}_{M-1}.$$

This condition is preserved by the reduced gradient-flow dynamics, using the fact that under this initialization, we get

$$\boldsymbol{b}_{+,0} = b_{+,0}\mathbf{1}_{M-1}, \qquad \boldsymbol{b}_{-,0} = b_{-,0}\mathbf{1}_{M-1},$$

which implies

$$\boldsymbol{a}_{+,t} = a_{+,t}\mathbf{1}_{M-1}, \qquad \boldsymbol{a}_{-,t} = a_{-,t}\mathbf{1}_{M-1} \qquad \text{for all } t \geq 0.$$

Consequently, the dynamics further reduce from vector-valued coordinates to four scalar variables. This is the gradient-flow analogue of the symmetry reduction used in the regularization-path analysis, where the optimizer is forced by the same permutation symmetry to have block-constant majority and minority coordinates. See Thm. 7.

This block-constant initialization also makes the phase structure visible at the level of gradient-flow time. Although gradient flow is not parametrized by the regularization strength λ, the same mode-wise ordering observed along the regularization path appears empirically in the trajectory: the majority coordinates grow first, followed by the majority–minority coordinate, and finally the minority coordinates. Thus, under the all-ones spectral initialization, the regularization-path phase transition has a clear dynamical analogue.

We verify this invariance numerically in Figure 10. Starting from block-constant Hadamard coordinates, the entries of each vector block remain indistinguishable throughout training, up to numerical precision. Equivalently, the within-block variance of both $\boldsymbol{a}_{+,t}$ and $\boldsymbol{a}_{-,t}$ stays at machine precision along the trajectory.