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

Arman Lotfalikhani ARMANL@ECE.UBC.CA
Department of Electrical and Computer Engineering, University of British Columbia, Canada

Mohammad Moshtaghifar MMOSHTAQ@CS.UBC.CA
Department of Computer Science, University of British Columbia, Canada

Connall Garrod CONNALL.GARROD@MATHS.OX.AC.UK
Mathematical Institute, University of Oxford, United Kingdom

Jonathan P. Keating JON.KEATING@MATHS.OX.AC.UK
Mathematical Institute, University of Oxford, United Kingdom

Christos Thrampoulidis CTHRAMPO@ECE.UBC.CA
Department of Electrical and Computer Engineering, University of British Columbia, Canada

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. We derive a closed-form expression for the full regularization path, which exhibits sequential mode emergence, but with novel behavior: active singular values diverge, only normalized logits converge and can overshoot the limiting geometry. We then show that a related sequential picture holds for gradient flow under appropriate spectrally aligned initialization. Our analysis relies on a novel imbalance-adapted Hadamard basis in which 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 [3–5, 21, 35, 53]. A particularly influential example is the spectral-dynamics framework of Saxe et al. [44, 45], 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 [7, 26, 39, 55, 59, 60].

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 convergence and 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: Step-imbalanced data is the minimal label-symmetry distortion that produces nontrivial mode structure: majority, majority-minority, and minority modes [60]. A

© A. Lotfalikhani, M. Moshtaghifar, C. Garrod, J.P. Keating & C. Thrampoulidis.

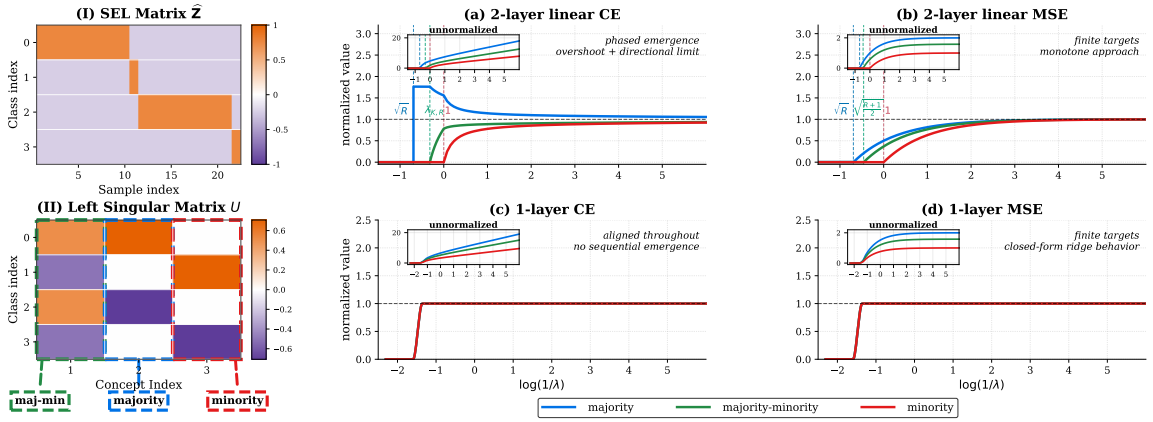


Figure 1: We study sequential learning in a deliberately minimal setting: step-imbalanced classification with $K/2$ majority/minority classes and imbalance ratio $R > 1$. **(I)** The centered one-hot encoding matrix $\widehat{\boldsymbol{Z}} = (\boldsymbol{I} - \frac{1}{K}\boldsymbol{1}\boldsymbol{1}^\top)\boldsymbol{Y}$ has three distinct singular modes: majority–minority, majority, and minority. **(II)** These are reflected in the structure of the left singular matrix $\boldsymbol{U}$ [60]. **(a)–(d)** shows 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. Coordinates are normalized so that the limiting profile equals one; unnormalized profiles shown in insets. **(a)** We analytically solve the CE regularization path: modes activate in phases, but activation thresholds are more intricate than in MSE, unnormalized singular values diverge, and normalized coordinates converge only directionally and nonmonotonically, with possible overshoot. **(b)** The well-studied MSE picture [44, 45]: finite targets approached monotonically along sigmoidal curves.[1] **(c, d)** One-layer CE and MSE paths show no sequential emergence.

two-layer linear architecture, unlike a single-layer linear model, allows this structure to emerge sequentially. Orthogonality plays a double role: it simplifies the analysis, matching the original setting of Saxe et al. [44, 45], and makes the model coincide with the unconstrained features model (UFM) [37] 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. Concretely, our contributions are as follows:

• **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*. This approximate diagonalization reduces the matrix-valued CE problem to a lower-dimensional one.

• **Regularization path.** Using this reduction, we cast the ridge-regularized UFM as a four-dimensional optimization which we solve explicitly for *every* regularization strength λ. The solution exhibits a phase structure: as λ decreases, majority activates, then majority–minority, and finally the minority. Activation thresholds and active-mode magnitudes reduce to scalar nonlinear equations, which we give in closed form. From this we prove the vanishing-regularization limit, obtain a logarithmic convergence rate, and expose qualitative gaps with MSE sequential-learning; see Fig. 1.

• **Gradient-flow dynamics.** We perform an analogous study of gradient-flow under a spectrally aligned initialization compatible with the approximate diagonalization. Results deferred to App. C.

Closely related work. Garrod et al. [19] recently showed that the spectral dynamics framework for MSE dynamics by Saxe et al. [44, 45], 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

[1] App. B for the connection between the regularization-path shown here and the gradient-flow analysis of Saxe et al. [44].

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 [19] has no distinct singular modes. *(c) Regularization path.* We apply the spectral framework to the regularization path. This both connects the regularization path to gradient flow and yields the first analytic characterization of a CE regularization path beyond the closed-form MSE case.

Notation. Define $[N] \coloneqq \{0, 1, \ldots, N-1\}$. Let K classes, set $M \coloneqq K/2$, and use R for the imbalance ratio. Lower/upper-case boldface denote vectors/matrices; $\otimes$ and $\odot$ denote Kronecker and Hadamard products. $\boldsymbol{e}_i$ is 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 the basis in $\mathbb{R}^2$. 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. $\boldsymbol{\Phi}_K \in \{\pm 1\}^{K \times K}$ denotes the Sylvester–Hadamard matrix of order $K = 2^m, m \in \mathbb{N}$. 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. $R \in \mathbb{N}$ denotes 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. It is convenient to order class indices so that even numbers $(0, 2, \ldots)$ correspond to majorities and odd numbers $(1, 3, \ldots)$ to minorities. This lets us write the one-hot encoding matrix $\boldsymbol{Y}$, and its centered counterpart the *simplex-encoding label (SEL) matrix* $\widehat{\boldsymbol{Z}}$, as

$$\boldsymbol{Y} \coloneqq \boldsymbol{I}_M \otimes \begin{pmatrix} \mathbf{1}_R^\top & 0 \\ \mathbf{0}_R^\top & 1 \end{pmatrix} \in \mathbb{R}^{K \times n}, \qquad \widehat{\boldsymbol{Z}} \coloneqq \left(\boldsymbol{I}_K - \tfrac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top\right)\boldsymbol{Y}. \tag{1}$$

When $R = 1$, data are balanced and $\boldsymbol{Y} = \boldsymbol{I}_K$, so that $\widehat{\boldsymbol{Z}} = \boldsymbol{I}_K - \frac{1}{K}\mathbf{1}_K\mathbf{1}_K^\top$ is a simplex ETF (up to scaling). Our focus here, however, is the imbalanced case (see Fig. 1, left).

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) [16, 37]. Throughout, we let $d \geq K$. Concretely, 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 $\boldsymbol{e}_i \in \mathbb{R}^n$, 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,

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

models DNN training with weight decay [15, 18, 23, 24, 48, 63]. For $\lambda > 0$ (UFM$_\lambda$) has finite minimizers, which diverge as $\lambda \to 0^+$, but converge in direction [42, 52]. The RP is accessible for analysis thanks to its *tight* convex semidefinite program (SDP) relaxation in logit space $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{H}$:

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

Concretely, for all $\lambda > 0$, there exists orthogonal matrix $\boldsymbol{R}$ ($\boldsymbol{R}^\top \boldsymbol{R} = \boldsymbol{I}_K$) such that the solution of (UFM$_\lambda$) satisfies $\boldsymbol{W}_\lambda = \boldsymbol{U}\sqrt{\boldsymbol{\Sigma}}\boldsymbol{R}^\top$ 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$_\lambda$). In the balanced case ($R = 1$), 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, i.e., $\boldsymbol{W}_\lambda \boldsymbol{W}_\lambda^\top = \boldsymbol{H}_\lambda^\top \boldsymbol{H}_\lambda \propto \widehat{\boldsymbol{Z}}$ [63]. 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 [24, 52]. Here, we ask: *precisely how do the singular modes of $\boldsymbol{Z}_\lambda$ enter and evolve as λ decreases?*

3. Technical tool: Approximate Softmax diagonalization

Both our RP and GF analyses rest on a technical observation about how the softmax operator interacts with the singular subspace of the SEL matrix $\widehat{\boldsymbol{Z}}$. To formalize this, recall $K = 2^m, m \in \mathbb{N}$ and let $\boldsymbol{U} \in \mathbb{R}^{K\times(K-1)}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times(K-1)}$ be the following matrices, parameterized by the Hadamard basis $\overline{\boldsymbol{\Phi}}_M$, whose columns form orthonormal bases for, the left and right singular subspaces of $\widehat{\boldsymbol{Z}}$:

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

Direct computation confirms that $\widehat{\boldsymbol{Z}} = \boldsymbol{U}\,\mathrm{diag}\,(\boldsymbol{\sigma})\,\boldsymbol{V}^\top$ where $\boldsymbol{\sigma} = [\sqrt{\frac{R+1}{2}}, \sqrt{R}\mathbf{1} \otimes \boldsymbol{e}_+ + \mathbf{1} \otimes \boldsymbol{e}_-]$. Thus, for $R > 1$, $\widehat{\boldsymbol{Z}}$ has *three* distinct ordered singular modes: a **majority mode** at value $\sqrt{R}$, a **majority-minority mode** at value $\sqrt{(R+1)/2}$, and, a **minority mode** at value 1; see [60] for semantic interpretation of these modes.

We would like the softmax residual $\boldsymbol{Y} - \mathbb{S}(\cdot)$ to preserve the diagonal SVD structure of $\widehat{\boldsymbol{Z}}$, as the MSE residual does in [44] and the CE residual does for balanced data in [19]. However, applying $\boldsymbol{Y} - \mathbb{S}(\cdot)$ to a matrix of the form $\boldsymbol{U}\,\mathrm{diag}\,(\cdot)\,\boldsymbol{V}^\top$ produces components outside the span of $\boldsymbol{V}$. Nevertheless, we find in Theorem 1 below that the leakage is confined to a *single* extra direction $\widetilde{\boldsymbol{v}} := \frac{1}{\sqrt{MR(R+1)}} \mathbf{1}_M \otimes \left(\begin{smallmatrix} \mathbf{1}_R \\ R \end{smallmatrix}\right)$. To formalize this, set $\widetilde{\boldsymbol{V}} := [\widetilde{\boldsymbol{v}}, \boldsymbol{V}]$, and introduce the operator $\mathrm{diag}_{+1}(\cdot) : \mathbb{R} \times \mathbb{R} \times \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \to \mathbb{R}^{(K-1)\times K}$ defined by

$$\mathrm{diag}_{+1}(\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-) := \begin{bmatrix} \alpha\, \boldsymbol{e}_0 & \boldsymbol{A} \end{bmatrix}, \qquad \boldsymbol{A} := \mathrm{diag}\,(a,\, \boldsymbol{a}_+ \otimes \boldsymbol{e}_+ + \boldsymbol{a}_- \otimes \boldsymbol{e}_-)\,.$$

This combines a $(K-1) \times (K-1)$ diagonal matrix $\boldsymbol{A}$ and a scaled first standard basis vector $\alpha\, \boldsymbol{e}_0 \in \mathbb{R}^{K-1}$. Pre- and post-multiplying by $\boldsymbol{U}$ and $\widetilde{\boldsymbol{V}}$ thus defines an SVD-like decomposition augmented by a rank-one term, the structure used in 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*

$$\boxed{\boldsymbol{Y} - \mathbb{S}\left(\boldsymbol{U}\,\mathrm{diag}_{+1}(\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-)\,\widetilde{\boldsymbol{V}}^\top\right) = \boldsymbol{U}\,\mathrm{diag}_{+1}(\beta, b, \boldsymbol{b}_+, \boldsymbol{b}_-)\,\widetilde{\boldsymbol{V}}^\top.} \tag{2}$$

where the map $(\beta, b, \boldsymbol{b}_+, \boldsymbol{b}_-) := \mathcal{F}((\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-))$ is given in closed-form in (4) in App. A.

The theorem's closure property is a consequence of our specific Hadamard parameterization of $\boldsymbol{U}, \widetilde{\boldsymbol{V}}$, which adopts the Hadamard-basis idea of [19] but unlike the balanced case, requires the rank-one extension $\widetilde{\boldsymbol{v}}$ to preserve the decomposition. The main novelty and challenge in the proof (deferred to App. F) lies in identifying this parameterization and the formulas that dictate the mapping $\mathcal{F}(\cdot)$ from the LHS spectral parameters $(\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-)$ to the RHS spectral parameters $(\beta, b, \boldsymbol{b}_+, \boldsymbol{b}_-)$. The key point is that we derive this mapping in closed form; see (4) in App. A.

The intuition to keep in mind is the following: the spectral parameters $\boldsymbol{a}_\pm$ correspond to the majority/minority modes and the pair (α, a) to the majority–minority mode; the parameters $\boldsymbol{b}_\pm$ and (β, b) correspond to these in turn. The formulas couple them: (β, b) depends on all the LHS parameters, and the majority/minority terms $\boldsymbol{b}_\pm$ depend on the maj–min pair (α, a), so the modes do not decouple. The majority-minority mode is described by *two* parameters (α, a), rather than one, by design. While exact diagonalization is possible using the singular vector $(a\boldsymbol{v}_0 + \alpha\widetilde{\boldsymbol{v}})/\sqrt{a^2+\alpha^2}$ with singular value $\sqrt{\alpha^2+a^2}$, the formulation of Theorem 1 is better suited for the subsequent analysis, since it keeps the right subspace $\widetilde{\boldsymbol{V}}$ fixed.

4. Regularization path

We analytically solve (UFM$_\lambda$) for every $\lambda > 0$. From Sec. 2, it suffices to solve the convex relaxation (Z-UFM$_\lambda$) in logit space. While this is convex and can be solved numerically, there is no a priori reason to expect an analytic solution. We show that one nevertheless exists in two steps.

First, we use Theorem 1 to reduce the SDP 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 $\boldsymbol{Z}_\lambda = \boldsymbol{U}\,\mathrm{diag}_{+1}(\alpha_\lambda, a_\lambda, \boldsymbol{a}_{+,\lambda}, \boldsymbol{a}_{-,\lambda})\,\widetilde{\boldsymbol{V}}^\top$, where $\alpha_\lambda, a_\lambda, \boldsymbol{a}_{+,\lambda}, \boldsymbol{a}_{-,\lambda}$ are unique minimizers of*

$$\min_{\alpha,a,\boldsymbol{a}_\pm} \mathcal{L}\big(\boldsymbol{U}\,\mathrm{diag}_{+1}(\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-)\,\widetilde{\boldsymbol{V}}^\top\big) + \lambda\big(\sqrt{\alpha^2+a^2} + \|\boldsymbol{a}_+\|_1 + \|\boldsymbol{a}_-\|_1\big)\,. \tag{3}$$

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 showing that the global minimizer of (Z-UFM$_\lambda$) lies on the $(\boldsymbol{U}, \widetilde{\boldsymbol{V}})$ subspace for all $\lambda > 0$, and this is where Theorem 1 enters. The optimality condition for (Z-UFM$_\lambda$) is $\boldsymbol{Y} - \mathbb{S}(\boldsymbol{Z}) \in \lambda\,\partial\|\boldsymbol{Z}\|_*$. For candidate $\boldsymbol{Z} = \boldsymbol{U}\,\mathrm{diag}_{+1}(\boldsymbol{a})\,\widetilde{\boldsymbol{V}}^\top$, Theorem 1 allows us to express the optimality condition as $\boldsymbol{U}\,\mathrm{diag}_{+1}(\mathcal{F}(\boldsymbol{a}))\,\widetilde{\boldsymbol{V}}^\top \in \lambda\,\partial\|\boldsymbol{U}\,\mathrm{diag}_{+1}(\boldsymbol{a})\,\widetilde{\boldsymbol{V}}^\top\|_*$, which by orthogonality of $\boldsymbol{U}, \widetilde{\boldsymbol{V}}$ is equivalent to the optimality condition of (3).

Second, we solve the optimality conditions of (3) explicitly. Consider the auxiliary variables $(\gamma_\lambda, \delta_\lambda) := \sqrt{\frac{2}{R+1}}\big(a_\lambda + \frac{\alpha_\lambda}{\sqrt{R}}, a_\lambda - \alpha_\lambda\sqrt{R}\big)$. Let $Q(x,\theta) := \log(x^{M-1}((1+\theta)x+\theta))$, $\theta_+ := M[\sqrt{R}\lambda^{-1} - 1]_+$, $\theta_- := M[\lambda^{-1} - 1]_+$. Finally, let r_λ solve a scalar nonlinear equation and s_λ be a closed-form function of it, both given in App. A.2, and $\gamma_\lambda = Q(s_\lambda, \theta_+)$, $\delta_\lambda = Q(r_\lambda, \theta_-)$..

Theorem 3 (Analytic solution) *Let $\boldsymbol{a}_\lambda = (\alpha_\lambda,\, a_\lambda,\, a_{+,\lambda},\, a_{-,\lambda})$ the minimizer of (3) and $\lambda_{K,R} \in [1, \sqrt{(R+1)/2}]$ the unique solution of $Q\big(2\lambda/\sqrt{2\lambda^2-1} - 1\,,\, \theta_+\big) = 0$. Then, $\boldsymbol{a}_\lambda$ satisfies:*

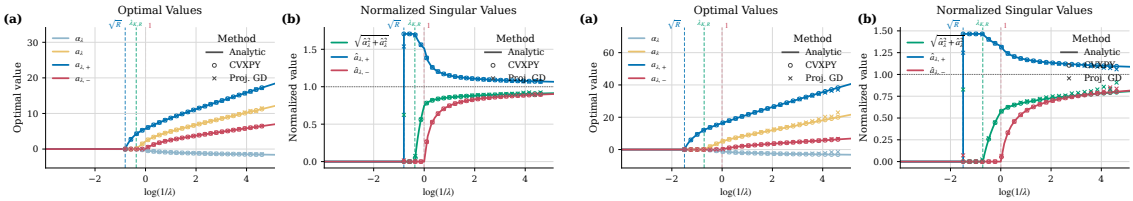
(i) *For $\lambda \geq \sqrt{R}$, $\boldsymbol{a}_\lambda = (0,0,0,0)$.*

(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 $a_{+,\lambda} = \sqrt{R}(\gamma_\lambda - M\log s_\lambda)$, $a_{-,\lambda} = \delta_\lambda - M\log s_\lambda$.*

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 ($a_{-,\lambda}$).



(a) $R = 5$ (b) $R = 20$

Figure 2: Approximate diagonalization components of the RP for $K = 8$ for different imbalance ratios R. In each panel, analytic curves obtained from Theorem. 3 are overlaid with those produced by a direct CVXPY solve of (Z-UFM$_\lambda$) and by projected gradient descent on the parameter sphere $\|\boldsymbol{W}\|_F^2 + \|\boldsymbol{H}\|_F^2 = 2\|\boldsymbol{Z}_\lambda\|_*$ matching the norm of the RP solution at the same λ value.

Emergence thresholds. The theorem identifies the λs at which modes first become nonzero. Majority and minority activate at $\lambda_+ := \sqrt{R}$, $\lambda_- := 1$, 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. App. G.4 and Fig. 3 for details and visualization.

Convergence. More importantly, the theorem identifies the exact evolution of the parameters within each phase in terms of a single scalar nonlinear equation. In App. G.2, we show $a_{\pm,\lambda} \geq 0$ for all λ, with strict positivity once activation occurs. Moreover, the active coordinates increase as λ decreases, and diverge as $\lambda \to 0$. However, their normalized direction converges to the SEL singular direction.

Theorem 4 (Convergence to SEL) *With $\boldsymbol{a}_\lambda$ as in Theorem 3, as $\lambda \to 0$, the normalized spectral modes $\overline{\boldsymbol{a}}_\lambda$ converge to $\overline{\boldsymbol{a}}_\infty = (0, \sqrt{(R+1)/2}, \sqrt{R}, 1)$. Consequently, the logit singular values $\boldsymbol{\sigma}_\lambda := (\sqrt{a_\lambda^2 + \alpha_\lambda^2}, a_{+,\lambda}, a_{-,\lambda})$ converge in direction to the SEL singular values $\boldsymbol{\sigma} = (\sqrt{(R+1)/2}, \sqrt{R}, 1)$. Convergence is at rate $O(1/\log(\lambda^{-1}))$.*

Unlike prior vanishing-regularization convergence results (e.g. [30, 50–52]), which to our knowledge all rely on a non-constructive proof by contradiction following [42], the result above is explicit (it follows directly from the phase-wise formulas of Theorem 3), and gives an explicit convergence rate.

Mode emergence without finite targets. Our analysis highlights a key difference to the classical MSE sequential-learning picture. Under MSE, each mode follows a sigmoidal trajectory, leaving the initialization scale and growing from below monotonically toward its finite target value [20, 44]. In CE, by contrast, the active singular values have no finite targets: they diverge as $\lambda \to 0$, and only their normalized direction converges. Moreover, earlier activation does *not* mean monotone growth from below toward the limiting normalized value: Majority activates first but approaches its normalized SEL value from above; majority-minority mode activates next and approaches from below; and minority activates last, approaches from below and can approach faster than the majority-minority mode depending on K and R. This contrast with the MSE is illustrated in Fig. 1.

5. Conclusion

We analyzed how CE training learns data modes in a deliberately minimal setting, solving the RP and studying GF under a spectral initialization that preserves the subspaces the RP identifies. The resulting dynamics reveal quantitative and qualitative departures from MSE. The key obstacle in extending MSE theory to CE is the loss of closed-form decoupled dynamics; extending the softmax-diagonalization perspective of [19] to a non-symmetric setting shows this is not a dead end. Future extensions include nonlinear and deeper models, richer mode structures, and non-orthogonal inputs.

Acknowledgments

This work was partially funded by the NSERC Discovery Grant No. 2021-03677, the Alliance Grant ALLRP 581098-22, a CIFAR Catalyst grant, and a gift from Google. The authors also acknowledge the use of the Sockeye cluster by UBC Advanced Research Computing. CG is supported by the Charles Coulson Scholarship.

References

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

[2] 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.

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

[4] 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*, 2018.

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

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

[7] 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.

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

[9] 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*, pages 10815–10838. PMLR, 2023.

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

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

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

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

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

[26] 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.

[27] 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.

[28] 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.

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

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

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

[38] Jean-Jacques Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *CR Acad. Sci. Paris Sér. A Math*, 255:2897–2899, 1962.

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

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.

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

[47] 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, pages 52991–53024. Curran Associates, Inc., 2023.

[48] 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.

[49] 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.

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

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

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

[53] 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, pages 106059–106104. Curran Associates, Inc., 2024.

[54] 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.

[55] 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.

[56] 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.

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

[58] Yedi Zhang, Andrew M. Saxe, and Peter E. Latham. Saddle-to-saddle dynamics explains a simplicity bias across neural network architectures. *ArXiv*, abs/2512.20607, 2025. URL https://api.semanticscholar.org/CorpusID:284133040.

[59] 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.

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

[61] 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.

[62] 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.

[63] 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 Conclusion **6**

A Deferred definitions **13**
A.1 The Mapping of theorem 1 and proof outline . . . 13
A.2 Definitions of r_λ, s_λ . . . 13

B Background on MSE loss **14**
B.1 Gradient-flow . . . 14
B.2 Regularization-path . . . 15
B.3 On their relation . . . 15

C Gradient-flow Path **16**
C.1 Background . . . 16
C.2 Results . . . 17
C.2.1 Logit dynamics under balancedness . . . 17
C.2.2 Emergence . . . 18
C.2.3 Convergence . . . 19

D Related Work and Limitations **19**
D.1 Detailed Related Work . . . 19
D.2 Limitations . . . 22

E Notation and basic facts **23**

F Theorem 1 Proof and Discussions **24**
F.1 Proof of Theorem 1 . . . 25
F.2 Simplified form of expressions . . . 30

G Proofs for Regularization-path **31**
G.1 Proof of Theorem 2 and basic properties of the solution . . . 31
G.1.1 Reduction to four variables . . . 35
G.2 Proof of Theorem 3: Analytic solution . . . 37
G.3 Analytic Solution of Theorem 3 . . . 41
G.4 Behavior of the threshold $\lambda_{K,R}$ . . . 43
G.5 Asymptotic behavior of RP solution and Proof of Theorem 4 . . . 45
G.6 Alternative proof of regularization-path limit . . . 46

H Proofs for Gradient-flow path **48**
H.1 Proof of Theorem 6 . . . 48
H.2 Proof of Theorem 10: Majority and Minority tend to Uniformity . . . 49
H.3 Proof of Theorem 8 . . . 50
H.3.1 Auxiliary lemmas . . . 51
H.3.2 Proof of Proposition 42 . . . 52
H.3.3 Proof of Proposition 43 . . . 53
H.4 Lyapunov function counterexample . . . 55

I Numerical Results **57**
I.1 Regularization Path Experiments . . . 57
I.1.1 Regularization Path Comparison for Different K and R . . . 57
I.2 Gradient-flow experiments . . . 61
I.2.1 Setup . . . 61
I.2.2 Results . . . 61
I.2.3 Preservation of block-constant spectral coordinates . . . 63

Appendix A. Deferred definitions

A.1. The Mapping of theorem 1 and proof outline

The map $(\beta, b, \boldsymbol{b}_+, \boldsymbol{b}_-) := \mathcal{F}\left((\alpha, a, \boldsymbol{a}_+, \boldsymbol{a}_-)\right)$ of Theorem 1 is given explicitly by

$$\beta = \sqrt{\tfrac{R}{2(R+1)}}\,(X_+ - X_-)\ , \qquad b = \sqrt{\tfrac{1}{2(R+1)}}\,(RX_+ + X_-)\ , \tag{4a}$$

$$\boldsymbol{b}_+ = \sqrt{R}\big(\mathbf{1}_{M-1} - \tfrac{1}{2M}\,X_+\,\overline{\boldsymbol{\Phi}}_M^\top e^{\boldsymbol{x}_+}\big), \qquad \boldsymbol{b}_- = \mathbf{1}_{M-1} - \tfrac{1}{2M}\,X_-\,\overline{\boldsymbol{\Phi}}_M^\top e^{\boldsymbol{x}_-}\ , \tag{4b}$$

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

$$(\gamma\,,\,\delta) := \sqrt{\tfrac{2}{R+1}}\Big(a + \tfrac{1}{\sqrt{R}}\,\alpha\,,\, a - \alpha\sqrt{R}\Big)\,. \tag{5}$$

As mentioned in Sec. 3 the main difficulty in the proof is identifying the right decomposition and obtaining the formulas above. The key in the proof in App. F was to realize that the residual matrix $\boldsymbol{Q}$ in the LHS of (2) can decompose its $n = (R+1)M$ columns in 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 $\boldsymbol{p}_+ := \boldsymbol{e}_0 - \mathbb{S}\left(\boldsymbol{x}_+ \otimes \boldsymbol{e}_+\right)$ and $\boldsymbol{p}_- := \boldsymbol{e}_1 - \mathbb{S}\left(\boldsymbol{x}_- \otimes \boldsymbol{e}_-\right)$. Finally, 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 $\boldsymbol{p}_+, \boldsymbol{p}_-$.

Note also that (5) defines a one-to-one mapping between (α, a) and (γ, δ) pair; we often find it convenient to work with the latter.

A.2. Definitions of r_λ, s_λ

We give the deferred definition of the pair (r_λ, s_λ) in Theorem 3. Recall from the main body the definitions

$$\theta_+ := M[\sqrt{R}\,\lambda^{-1} - 1]_+, \quad \theta_- := M[\lambda^{-1} - 1]_+, \quad Q(x,\theta) := \log\big(x^{M-1}((1+\theta)x + \theta)\big). \tag{6}$$

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

$$\begin{cases} 1 + r_\lambda = \frac{M}{\lambda(\theta_- + M)} \sqrt{\frac{2(1+s_\lambda)^2}{(1+s_\lambda)^2 - 2}}, \\ (1 + r_\lambda) \cdot Q(r_\lambda, \theta_-) = \sqrt{R} \left(\frac{M}{\lambda(\theta_- + M)} \right) (1 + s_\lambda) \cdot Q(s_\lambda, \theta_+), \end{cases} \tag{7}$$

and set $\gamma_\lambda \coloneqq Q(s_\lambda, \theta_+)$ and $\delta_\lambda \coloneqq Q(r_\lambda, \theta_-)$ for all λ.

Appendix B. 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 [20, 44]. 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* [1, 46]. 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 [10].

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

B.1. Gradient-flow

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

$$\mathcal{L}_{\text{mse}}(\boldsymbol{W}\boldsymbol{H}) \coloneqq \frac{1}{2} \|\widehat{\boldsymbol{Z}} - \boldsymbol{W}\boldsymbol{H}\|_F^2. \tag{8}$$

Under the spectrally aligned initialization of Saxe et al. [44], $\boldsymbol{W}_0 = \boldsymbol{U} \operatorname{diag}(\boldsymbol{w}_0) \boldsymbol{R}^\top$, $\boldsymbol{H}_0 = \boldsymbol{R} \operatorname{diag}(\boldsymbol{h}_0) \boldsymbol{V}^\top$, the singular subspace is invariant for the GF. Concretely, further assuming balancedness ($\boldsymbol{w}_0 = \boldsymbol{h}_0$), the logit singular values

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

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

$$\dot{a}_{i,t} = 2 a_{i,t} (\sigma_i - a_{i,t}), \tag{9}$$

where σ_i are the diagonal entries of $\boldsymbol{\Sigma}$. Thus, for initialization $a_i(0) = e^{-2\delta}$, the modes evolve as [20, Theorem1]

$$a_{i,t} = \sigma_i \cdot \frac{1}{1 + (\sigma_i e^{2\delta} - 1) e^{-2\sigma_i t}}. \tag{10}$$

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 [20, Theorem2]

$$a_{i,\delta t} \to \frac{\sigma_i}{1 + \sigma_i} \cdot \mathbf{1}\{t = 1/\sigma_i\} + \sigma_i \cdot \mathbf{1}\{t > 1/\sigma_i\}, \tag{11}$$

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 [60].

B.2. Regularization-path

We next study the ridge-regularized counterpart of (8):

$$(\boldsymbol{W}_\lambda, \boldsymbol{H}_\lambda) \in \arg\min_{\boldsymbol{W},\boldsymbol{H}} \mathcal{L}_{\mathrm{mse}}(\boldsymbol{W}\boldsymbol{H}) + \frac{\lambda}{2}\|\boldsymbol{W}\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{H}\|_F^2. \tag{12}$$

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

$$\boldsymbol{Z}_\lambda \in \arg\min_{\boldsymbol{Z}} \frac{1}{2}\|\widehat{\boldsymbol{Z}} - \boldsymbol{Z}\|_F^2 + \lambda\|\boldsymbol{Z}\|_*. \tag{13}$$

Since $\widehat{\boldsymbol{Z}} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, the solution is given by singular-value soft-thresholding (e.g., [12, Theorem 2.1]):

$$\boldsymbol{Z}_\lambda = \boldsymbol{U}\,\mathrm{diag}\left((\boldsymbol{\sigma} - \lambda)_+\right)\boldsymbol{V}^\top, \tag{14}$$

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

$$a_{i,\lambda} = (\sigma_i - \lambda)_+, \qquad i \in [K-1]. \tag{15}$$

Thus, mode i is active exactly when $\lambda < \sigma_i$. 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

$$\arg\min_{\boldsymbol{Z}} \frac{1}{2}\|\widehat{\boldsymbol{Z}} - \boldsymbol{Z}\|_F^2 + \frac{\lambda}{2}\|\boldsymbol{Z}\|_F^2 = \frac{1}{1+\lambda}\widehat{\boldsymbol{Z}}. \tag{16}$$

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

B.3. On their relation

Since we have closed-form solutions for both GF (in (10)) and RP (in (15)) we can explicitly relate them. The proposition below is a special case of a more general finding by Berthier [10].

Proposition 5 (RP to GF for MSE loss) *Let $a_{i,\lambda}$ denote the logit singular values of the RP in Eq. (15). Further, let $a_{i,t}^{(\delta)}$ denote the logit singular values of the GF solution to the MSE loss Eq. (10) 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 $\lambda > 0$,*

$$a_{i,\lambda} = \lim_{\delta\to\infty} \hat{a}_{i,\frac{\delta}{\lambda}}^{(\delta)}, \qquad \textit{where} \qquad \hat{a}_{i,t}^{(\delta)} := \frac{1}{t}\int_0^t a_{i,\tau}^{(\delta)}, d\tau. \tag{17}$$

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 (10), 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

$$\hat{a}_{i,\frac{\delta}{\lambda}}^{(\delta)} = \sigma_i + \frac{\lambda}{2\delta} \log\left(\sigma_i^{-1} e^{-2\delta} + \left(1 - \sigma_i^{-1} e^{-2\delta}\right) e^{-2\sigma_i \delta/\lambda}\right) . \tag{18}$$

Taking the limits and invoking the log-sum-exp asymptotic, 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 [10] 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.

Appendix C. Gradient-flow Path

C.1. Background

The gradient flow (GF) dynamics

$$\dot{\boldsymbol{W}}_t = (\boldsymbol{Y} - \mathbb{S}(\boldsymbol{W}_t \boldsymbol{H}_t))\boldsymbol{H}_t^\top , \qquad \dot{\boldsymbol{H}}_t = \boldsymbol{W}_t^\top (\boldsymbol{Y} - \mathbb{S}(\boldsymbol{W}_t \boldsymbol{H}_t)) , \tag{19}$$

is the continuous limit of gradient-descent for optimizing the (UFM). Although simplified, (UFM) remains non-convex, so even convergence of the GF iterates is non-trivial. 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 [28, 36]:

$$\min_{\boldsymbol{W},\boldsymbol{H}} \ \|\boldsymbol{W}\|_F^2 + \|\boldsymbol{H}\|_F^2 \qquad \text{sub. to} \quad (\boldsymbol{e}_{y_i} - \boldsymbol{e}_c)^\top \boldsymbol{W} \boldsymbol{h}_i \geq 1, \ \forall i, c \neq y_i, \tag{20}$$

such a characterization is *not* explicit. There are many KKT points, each with different structural properties, and, in general, convergence can be either at global or local minima and even at saddles of the max-margin problem in (20) [54].

In the binary case ($K = 2$), Ji and Telgarsky [29] 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. [54] establish convergence to a global optimum of (20). But $K = 2$ is both restrictive and trivial in our setting, since the only KKT point is the global minimizer. In the multiclass case, Garrod et al. [19] 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 $R = 1$, and the balanced setting has no distinct modes, so mode emergence cannot be studied there. Here, we ask: *in a setting with non-trivial data modes, how do the iterates $\boldsymbol{W}_t, \boldsymbol{H}_t$ evolve as t increases?*

C.2. Results

To analyze the GF dynamics (19), we follow the program of [19, 20, 44], and use a spectral initialization. Concretely, as dictated by Theorem 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 in [52], namely the SEL matrix $\widehat{\boldsymbol{Z}}$, admits this decomposition; *(b)* The entire regularization path lies exactly on this subspace, as shown in Sec. 4; and *(c)* Allowing us to leverage Theorem 1 ensures that under this initialization, the logits remain in this decomposition at *all* times, turning the matrix into a vector ODE .

Theorem 6 (Invariance of spectral subspace) *Initialize the GF updates in (19) such that $\boldsymbol{W}_0 = \boldsymbol{U}\,\mathrm{diag}\left(w_0, \boldsymbol{w}_{+,0}, \boldsymbol{w}_{-,0}\right)\boldsymbol{R}^\top$ and $\boldsymbol{H}_0 = \boldsymbol{R}\,\mathrm{diag}_{+1}\left(\alpha_{H,0}, h_0, \boldsymbol{h}_{+,0}, \boldsymbol{h}_{-,0}\right)\widetilde{\boldsymbol{V}}^\top$ for a partial orthogonal matrix $\boldsymbol{R} \in \mathbb{R}^{d\times(K-1)}$. Then, for all times t, it holds that*

$$\boldsymbol{W}_t = \boldsymbol{U}\ \mathrm{diag}\left(w_t, \boldsymbol{w}_{+,t}, \boldsymbol{w}_{-,t}\right)\ \boldsymbol{R}^\top \quad \textit{and} \quad \boldsymbol{H}_t = \boldsymbol{R}\,\mathrm{diag}_{+1}\left(\alpha_{H,t}, h_t, \boldsymbol{h}_{+,t}, \boldsymbol{h}_{-,t}\right)\ \widetilde{\boldsymbol{V}}^\top\,, \tag{21a}$$

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

$$\dot{w} = \beta\alpha_H + bh, \quad \dot{\alpha}_H = \beta w, \quad \dot{h} = bw, \quad \dot{\boldsymbol{w}}_\pm = \boldsymbol{b}_\pm \odot \boldsymbol{h}_\pm, \quad \dot{\boldsymbol{h}}_\pm = \boldsymbol{b}_\pm \odot \boldsymbol{w}_\pm\,, \tag{22}$$

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

C.2.1. LOGIT DYNAMICS UNDER BALANCEDNESS

As with the regularization-path, we now focus on the logit matrix $\boldsymbol{Z}_t = \boldsymbol{W}_t\boldsymbol{H}_t$, which under the parameterization in Eq. (21) 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 their dynamics, we adopt the standard balancedness condition at initialization, $\boldsymbol{W}_0^\top\boldsymbol{W}_0 = \boldsymbol{H}_0\boldsymbol{H}_0^\top$ [13, 19, 20, 44, 45], which is preserved by the dynamics; we leave the general case to future work. It will also be convenient to track the maj-min mode in the (γ, δ) parameterization of Eq. (5), motivated by its appearance in the regularization-path.

Corollary 7 *Suppose the spectral initialization of Theorem 6 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 $\boldsymbol{Z}_t = \boldsymbol{U}\,\mathrm{diag}_{+1}\left(\alpha_t, a_t, \boldsymbol{a}_{+,t}, \boldsymbol{a}_{-,t}\right)\widetilde{\boldsymbol{V}}^\top$ *for all* $t \geq 0$. *With* (γ_t, δ_t) *defined as in Eq. (5) (a one-to-one transformation of* (α_t, a_t)*), the logit parameters evolve as*

$$\dot{\gamma} = \frac{(2R\gamma^2 + \delta^2)\,X_+ + \gamma\delta\,X_-}{\sqrt{2(R\gamma^2+\delta^2)}}, \quad \dot{\delta} = \frac{(R\gamma^2 + 2\delta^2)\,X_- + R\gamma\delta\,X_+}{\sqrt{2(R\gamma^2+\delta^2)}}, \quad \dot{\boldsymbol{a}}_\pm = 2\,\boldsymbol{a}_\pm \odot \boldsymbol{b}_\pm. \tag{23}$$

We henceforth focus on the ODE in (23) 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.

C.2.2. EMERGENCE

In analogy with the regularization path, one expects the modes to appear sequentially, with the majority mode emerging first. This is indeed the case, although the statement is necessarily different: along GF the 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 regularization path, since the reduced ODE does not admit an explicit solution. We give some insights below.

Theorem 8 (Majority first) *Fix* $R > 1$. *Consider the reduced logit dynamics (23) 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. H.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 (23). For emergence, however, it suffices to control $\boldsymbol{b}_\pm$ uniformly in the small-coordinate regime. A key observation is that by setting $\boldsymbol{y}_+ := (M\sqrt{R})^{-1}\overline{\boldsymbol{\Phi}}\boldsymbol{a}_+$ and $\boldsymbol{y}_- := M^{-1}\overline{\boldsymbol{\Phi}}\boldsymbol{a}_-$, it holds $|\frac{X_\pm}{K}\overline{\boldsymbol{\Phi}}^\top e^{\boldsymbol{x}_\pm}| \leq \frac{1}{\boldsymbol{1}^\top e^{\boldsymbol{y}_\pm}}|\overline{\boldsymbol{\Phi}}^\top e^{\boldsymbol{y}_\pm}|$; thus the dependence on γ, δ drops out of this bound. Hence, when $\boldsymbol{a}_\pm = O(\varepsilon)$, $\boldsymbol{b}_+ = \sqrt{R}\,\boldsymbol{1} + O(\varepsilon)$ and $\boldsymbol{b}_- = \boldsymbol{1} + O(\varepsilon)$, 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 (23), $\dot{a}_0 = R\gamma X_+ + \delta X_-$. 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 does not close in terms of a_0 alone without controlling the relative sizes of γ and δ. A simple Cauchy–Schwarz bound gives $\dot{a}_0 \approx R\gamma + \delta \leq 2\sqrt{(R+1)/2}\cdot a_0$. 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 Theorem 8, 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 regularization-path 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 $\dot{a}_0 \approx R\gamma + \delta \geq \sqrt{2}\,a_0$ 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 by the relevant time.

C.2.3. CONVERGENCE

To further relate GF equations to the RP, it is useful to view (23) 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 $\|\boldsymbol{a}\|_{+1}$ in (3) 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 $-\mathcal{P}(\boldsymbol{a}_\lambda)\nabla\mathcal{E}(\boldsymbol{a}_\lambda) = 2\lambda\boldsymbol{a}_\lambda$. Therefore, RP solutions are exactly *radial points* of the GF.

Proposition 9 *A positive point $\boldsymbol{a} > 0$ is radial for the ODE in (23), that is satisfies $\dot{\boldsymbol{a}} = \mu\boldsymbol{a}$ for some $\mu \in (0, 2)$, iff it is the unique RP solution of (3) with $\lambda = \mu/2$.*

For (23), 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 Theorem 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 the flow contracts deviations from it measured in KL; App. H.2.

Theorem 10 (Maj. and min. tend to uniformity) *If the GF (23) is initialized so that $\boldsymbol{a}_\pm \propto \mathbf{1}$ then this remains true. More generally, for positive initialization, $\dot{D}_{\mathrm{KL}}(\overline{\mathbf{1}}\|\overline{\boldsymbol{a}}_\pm) \leq 0$, with equality iff $\boldsymbol{a}_\pm \propto \mathbf{1}$.*

Proof of Theorem 10 is in App. H.2. The two preceding results show that GF must converge, after normalization, to a logit matrix with three distinct singular values, in agreement with the regularization path. Moreover, among potential limiting configurations, the regularization path itself is the only full-rank case once the logit norm is sufficiently large. The remaining low-rank critical points correspond to networks whose widths are too narrow for the SEL matrix to emerge. Since rank-constrained minima are known to remain critical points in the unconstrained regime [58], and are typically saddles, we expect these low-rank solutions to be unstable. We therefore anticipate that gradient flow converges to the regularization path, which would lead to the emergence dynamics being inherited more broadly throughout the loss surface. We pose formalizing attractiveness of the regularization path as an important open question. Although it can likely be made rigorous using arguments similar to those of the balanced case [19], we note that the diagonal-plus-rank-one parameterization needed here makes the argument more delicate. As evidence for this added difficulty, App. H.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. [19], need not be a Lyapunov function in our coordinates.

Appendix D. Related Work and Limitations

D.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.

[40]. 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 [25, 27, 47, 49, 56] and empirically [2, 14–17, 22, 33, 40, 52, 61, 62]. 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) [23, 37, 63]. Its predictions extend to imbalanced classes [9, 15, 16, 52], multilabel data [34], language data [57, 61], and other losses [6, 17, 31].

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 [23]. In the symmetric balanced one-hot setting, the limiting neural-collapse geometry is the same for MSE-UFM and CE-UFM [23, 37]. Moreover, in that setting, regularization does not change the qualitative conclusion [19, 28, 63]; subtleties start to appear, for example, in deeper UFM variants [18, 48]. Once one departs from the balanced symmetric setting, however, the geometry can become sensitive to whether regularization is present [14–16, 40]. 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 [14, 15, 24, 63], or characterizing the limiting direction of GD/GF in the unregularized case [19, 28, 52]. Our focus is different: in addition to convergence, we study *mode emergence* along the training path. Our motivation on this is Zhao and Thrampoulidis [60], 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. [44] in terms of technical tools and by Saxe et al. [45] 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 [5, 7, 20, 38, 53]; see also the recent discussion in Bach [5]. 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 step toward overcoming this obstacle was taken very recently by Garrod et al. [19]. Their work inspired ours in its technical core. In the balanced one-hot setting, the authors of [19] showed that if one chooses a particular Sylvester–Hadamard singular basis of the label matrix, then the CE residual $\boldsymbol{Y} - \mathbb{S}(\boldsymbol{W}\boldsymbol{H})$ 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. [19]. 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 [42, 43]. In particular, Park and Hastie [41] 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 [15, 24]. Dang et al. [15] 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 [15] 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. The result closest to our RP analysis is [24]. They study the convexified CE-UFM logit/nuclear-norm problem (Z-UFM$_\lambda$) and characterize the block-support regimes of the global minimizer under class imbalance. In the bias-free two-cluster specialization relevant to our setting, their Theorem 3.2 gives the same qualitative sequence of supports, after scaling the regularization parameter by the number of samples: zero, majority-only, majority plus cross-cluster block, and fully active. In this sense, we can map their intermediate threshold, after a change of coordinates, to our $\lambda_{K,R}$. However, their result is a block-support characterization: it determines which blocks of the mean prediction matrix are active in each regime. First, this does *not* provide a mode-emergence interpretation. Second, they do *not* analyze how this threshold behaves as function of K, R as we do in App. G.4. Third, and more important, it does *not* solve the RP. This distinction is essential for our purposes. Our Hadamard-coordinate reduction turns the path into a four-dimensional ℓ_1-style problem whose coordinates are the majority, majority–minority, and minority singular modes. We then solve the active optimality equations phase by phase, obtaining scalar formulas for the mode amplitudes throughout the path. Our analytic mode-wise solution characterizes the evolution of the singular values, the $\lambda \to 0$ expansion of the normalized singular profile, the logarithmic directional convergence rate, and is also the structure that lets us connect the regularization path to gradient-flow mode emergence. Finally, [52] 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. [16] 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.

D.2. Limitations

We analyze how CE training learns data modes in a deliberately minimal setting: 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. In this setting, we solved the ridge-regularization path and studied GF under a spectral initialization that preserves the subspaces identified by this path. The resulting dynamics 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 [19] to a non-symmetric setting, we further showcase that this obstacle is not necessarily a dead end. 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. Our regularization-path analysis assumes that K is a power of two, although empirically the formulas appear more general (Fig. 7). 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 this connection remains open. Future directions include extending the analysis to nonlinear and deeper

models, richer mode structures, non-orthogonal inputs, and other softmax-based settings such as attention.

Appendix E. Notation and basic facts

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

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

Lemma 11 *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*

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

Then, any Sylvester Hadamard matrix of order M satisfies

$$\mathbf{\Pi}_m \overline{\mathbf{\Phi}}_M \boldsymbol{e}_i = (-1)^{m.(i+1)} \overline{\mathbf{\Phi}}_M \boldsymbol{e}_i, \tag{24}$$

$$\mathbf{\Pi}_{i\oplus j}\left(\mathbf{\Phi}_M \boldsymbol{e}_m \boldsymbol{e}_m^\top \mathbf{\Phi}_M^\top \boldsymbol{e}_j\right) = \mathbf{\Phi}_M \boldsymbol{e}_m \boldsymbol{e}_m^\top \mathbf{\Phi}_M^\top \boldsymbol{e}_i, \tag{25}$$

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

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

$$\sum_{m=0}^{M-1} \mathbf{\Pi}_m = (\boldsymbol{J} + \boldsymbol{I}_2) \otimes \ldots \otimes (\boldsymbol{J} + \boldsymbol{I}_2) = \mathbf{1}_M \mathbf{1}_M^\top. \tag{26}$$

Proof The key to prove (24) is to expand the RHS as a Kronecker product by using the fact that $\overline{\mathbf{\Phi}}_M \boldsymbol{e}_i = \mathbf{\Phi}_M \boldsymbol{e}_{i+1}$:

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

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

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

■

Lemma 12 *The matrix $\boldsymbol{Z} \in \mathbb{R}^{K\times n}$ which is defined below, can be diagonalized as follows:*

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

$$\boldsymbol{U} \coloneqq \begin{bmatrix} \frac{1}{\sqrt{2M}} \mathbf{1}_M \otimes \begin{pmatrix} 1 \\ -1 \end{pmatrix} & \frac{1}{\sqrt{M}} \overline{\mathbf{\Phi}}_M \otimes \boldsymbol{I}_2 \end{bmatrix} \mathbf{\Sigma}_{\boldsymbol{Z}} \coloneqq \begin{bmatrix} \sqrt{\frac{R+1}{2}} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{I}_{M-1} \otimes \mathbf{\Sigma}_2 \end{bmatrix} \tag{27}$$

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

Proof The diagonalization in (27) 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 13 *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. $\boldsymbol{C}\boldsymbol{e}_r = \boldsymbol{e}_s$. ■

Appendix F. Theorem 1 Proof and Discussions

In this section, we prove that the diagonal-plus-rank-one factorization is in fact a special case of the more general block-diagonal-plus-rank-one factorization which is made possible by Sylvester Hadamard matrices. In addition, to highlight the appearance of the rank-one term, we first prove the general permutation structure *without* having this term in the softmax argument, and later show that the results are retained, with minor changes to the expressions.

F.1. Proof of Theorem 1

Lemma 14 *For any $\boldsymbol{A} \in \mathbb{R}^{(K-1)\times(K-1)}$ having the block structure*

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

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

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

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

$$\begin{aligned}
\boldsymbol{x} &\coloneqq \frac{1}{2M}\left(\sqrt{\frac{2}{R+1}} a \boldsymbol{1}_M \otimes \begin{pmatrix} 1 \\ -1 \end{pmatrix} + 2\sum_{m=0}^{M-2} \overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_m \otimes \frac{\boldsymbol{a}_{m,0}}{\sqrt{R}}\right) \\
\boldsymbol{y} &\coloneqq \frac{1}{2M}\left(-\sqrt{\frac{2}{R+1}} a \boldsymbol{1}_M \otimes \begin{pmatrix} 1 \\ -1 \end{pmatrix} + 2\sum_{m=0}^{M-2} \overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_m \otimes \boldsymbol{a}_{m,1}\right).
\end{aligned} \tag{28}$$

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

$$\begin{aligned}
\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}^\top = &\frac{a}{\sqrt{R+1}}\left(\frac{1}{M}\boldsymbol{1}_M \boldsymbol{1}_M^\top\right) \otimes \left[\tfrac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix} \boldsymbol{1}_R^\top \quad -\tfrac{1}{\sqrt{2}}\begin{pmatrix} 1 \\ -1 \end{pmatrix}\right] \\
&+ \frac{1}{2}\sum_{m=0}^{M-2}\left(\overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_m \boldsymbol{e}_m^\top \overline{\boldsymbol{\Phi}}_M^\top\right) \otimes \left(\boldsymbol{A}_m \left[\begin{pmatrix} 1 \\ 0 \end{pmatrix} \tfrac{\boldsymbol{1}_R^\top}{\sqrt{R}} \quad \begin{pmatrix} 0 \\ 1 \end{pmatrix}\right]\right).
\end{aligned} \tag{29}$$

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 (25) 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

$$\boldsymbol{\Phi}_M \boldsymbol{e}_{m+1} \boldsymbol{e}_{m+1}^\top \boldsymbol{\Phi}_M^\top = \left[\boldsymbol{\Phi}_M \boldsymbol{e}_{m+1} \quad \dots \quad \boldsymbol{\Pi}_{M-1}(\boldsymbol{\Phi}_M \boldsymbol{e}_{m+1})\right] \tag{30}$$

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

$$\boldsymbol{U}\boldsymbol{A}\boldsymbol{V}^\top = \left[\boldsymbol{x}\boldsymbol{1}_R^\top \quad \boldsymbol{y} \quad (\widetilde{\boldsymbol{\Pi}}_1 \boldsymbol{x})\boldsymbol{1}_R^\top \quad \widetilde{\boldsymbol{\Pi}}_1 \boldsymbol{y} \quad \dots \quad (\widetilde{\boldsymbol{\Pi}}_{M-1}\boldsymbol{x})\boldsymbol{1}_R^\top \quad \widetilde{\boldsymbol{\Pi}}_{M-1}\boldsymbol{y}\right], \tag{31}$$

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 $\mathbb{S}\left(\boldsymbol{\Pi}\boldsymbol{x}\right) = \boldsymbol{\Pi}\mathbb{S}\left(\boldsymbol{x}\right)$, 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

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

■

Corollary 15 *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*

$$\begin{aligned} \widetilde{\boldsymbol{x}} &\coloneqq \frac{1}{M}\left(\sqrt{\frac{2}{R+1}}a\mathbf{1}_M + \frac{1}{\sqrt{R}}\overline{\boldsymbol{\Phi}}_M\boldsymbol{a}_+\right), \quad \mathbb{S}(\boldsymbol{x}) = \mathbb{S}\left(\widetilde{\boldsymbol{x}} \otimes \begin{pmatrix}1\\0\end{pmatrix}\right) \\ \widetilde{\boldsymbol{y}} &\coloneqq \frac{1}{M}\left(\sqrt{\frac{2}{R+1}}a\mathbf{1}_M + \overline{\boldsymbol{\Phi}}_M\boldsymbol{a}_-\right) \quad \mathbb{S}(\boldsymbol{y}) = \mathbb{S}\left(\widetilde{\boldsymbol{y}} \otimes \begin{pmatrix}0\\1\end{pmatrix}\right), \end{aligned} \tag{32}$$

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 (28) leads to

$$\begin{aligned} \boldsymbol{x} &= \frac{1}{2M}\left(\sqrt{\frac{2}{R+1}}a\mathbf{1}_M \otimes \begin{pmatrix}1\\-1\end{pmatrix} + 2\sum_{m=0}^{M-2}\frac{a_{m,0}}{\sqrt{R}}\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_m \otimes \begin{pmatrix}1\\0\end{pmatrix}\right) \\ \boldsymbol{y} &= \frac{1}{2M}\left(-\sqrt{\frac{2}{R+1}}a\mathbf{1}_M \otimes \begin{pmatrix}1\\-1\end{pmatrix} + 2\sum_{m=0}^{M-2}a_{m,1}\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_m \otimes \begin{pmatrix}0\\1\end{pmatrix}\right). \end{aligned} \tag{33}$$

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. ■

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 16 *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 (27)*

$$\boldsymbol{U}^\top\boldsymbol{Q}\boldsymbol{V} = \boldsymbol{B} = \operatorname{diag}(b, \boldsymbol{B}_0, \ldots \boldsymbol{B}_{M-2}), \quad \boldsymbol{B}_i \in \mathbb{R}^{2\times 2}, \tag{34}$$

$$b = \frac{1}{\sqrt{2(R+1)}}(\overline{\boldsymbol{\Phi}}_K\boldsymbol{e}_0)^\top(R\boldsymbol{p} - \boldsymbol{q}), \quad \boldsymbol{B}_i = \left((\overline{\boldsymbol{\Phi}}_M\boldsymbol{e}_i)^\top \otimes \boldsymbol{I}_2\right)\begin{bmatrix}\sqrt{R}\boldsymbol{p} & \boldsymbol{q}\end{bmatrix} \tag{35}$$

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 (26), the first column of $\boldsymbol{B}$ can be established as

$$\begin{aligned} \boldsymbol{Q}\boldsymbol{v}_0 &= \frac{1}{\sqrt{M(R+1)}}\sum_{m=0}^{M-1}(\boldsymbol{\Pi}_m \otimes \boldsymbol{I}_2)(R\boldsymbol{p} - \boldsymbol{q}) = \frac{1}{\sqrt{M(R+1)}}\left((\mathbf{1}_M\mathbf{1}_M^\top) \otimes \boldsymbol{I}_2\right)(R\boldsymbol{p} - \boldsymbol{q}) \\ \boldsymbol{B}\boldsymbol{e}_0 &= \boldsymbol{U}^\top\boldsymbol{Q}\boldsymbol{v}_0 = \frac{1}{\sqrt{2(R+1)}}(\overline{\boldsymbol{\Phi}}_K\boldsymbol{e}_0)^\top(R\boldsymbol{p} - \boldsymbol{q})\boldsymbol{e}_0 = b\boldsymbol{e}_0. \end{aligned} \tag{36}$$

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

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

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

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

where we have used the symmetry of $\boldsymbol{\Pi}_m, \boldsymbol{\Phi}$, and the mixed-product Kronecker identity. Next, we use the identity (24) for $\boldsymbol{\Pi}_m$ to simplify the expression to

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

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 17 *The matrices $\boldsymbol{U}, \boldsymbol{V}$ as defined in (27), satisfy*

$$\begin{aligned}\boldsymbol{U}\boldsymbol{U}^\top &= \boldsymbol{I}_k - \frac{1}{K}\mathbf{1}_K \mathbf{1}_K^\top \\ \boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{I}_n &= \boldsymbol{I}_M \otimes \begin{bmatrix} \frac{\mathbf{1}_R \mathbf{1}_R^\top}{R} - \boldsymbol{I}_R & \mathbf{0}_R \\ \mathbf{0}_R^\top & 0 \end{bmatrix} - \frac{1}{M(R+1)} (\mathbf{1}_M \mathbf{1}_M^\top) \otimes \begin{bmatrix} \frac{\mathbf{1}_R \mathbf{1}_R^\top}{R} & \mathbf{1}_R \\ \mathbf{1}_R^\top & R \end{bmatrix} .\end{aligned} \tag{38}$$

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:

$$\boldsymbol{U}\boldsymbol{U}^\top \boldsymbol{Q} = \boldsymbol{Q}, \quad \boldsymbol{Q}(\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{I}_k) = -\beta \boldsymbol{u}_0 \widetilde{\boldsymbol{v}}^\top, \tag{39}$$

$$\beta = \sqrt{\frac{R}{2(R+1)}} (\overline{\boldsymbol{\Phi}}_K \boldsymbol{e}_0)^\top (\boldsymbol{p} + \boldsymbol{q}), \quad \boldsymbol{u}_0 = \boldsymbol{U}\boldsymbol{e}_0 = \overline{\boldsymbol{\Phi}}_K \boldsymbol{e}_0, \quad \widetilde{\boldsymbol{v}} = \frac{\mathbf{1}_M \otimes \begin{pmatrix}\mathbf{1}_R \\ R\end{pmatrix}}{\sqrt{MR(R+1)}} \tag{40}$$

Proof Equation (38) 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 $\boldsymbol{Q}\boldsymbol{V}\boldsymbol{V}^\top$:

$$\begin{aligned}\boldsymbol{Q}(\boldsymbol{V}\boldsymbol{V}^\top - \boldsymbol{I}_n) &= \sum_{m=0}^{M-1}\left(\boldsymbol{e}_m^\top \otimes \boldsymbol{0}_{K\times(R+1)} - \frac{1}{M(R+1)}\boldsymbol{1}_M^\top \otimes \left((\boldsymbol{\Pi}_m \otimes \boldsymbol{I}_2)(\boldsymbol{p}+\boldsymbol{q})\begin{bmatrix}\boldsymbol{1}_R^\top & R\end{bmatrix}\right)\right)\\ &= -\frac{1}{M(R+1)}\boldsymbol{1}_M^\top \otimes \left(\left((\boldsymbol{1}_M\boldsymbol{1}_M^\top)\otimes \boldsymbol{I}_2\right)(\boldsymbol{p}+\boldsymbol{q})\begin{bmatrix}\boldsymbol{1}_R^\top & R\end{bmatrix}\right)\\ &= -\frac{1}{M(R+1)}\left(\boldsymbol{1}_M \otimes \left((\boldsymbol{1}_M^\top \otimes \boldsymbol{I}_2)(\boldsymbol{p}+\boldsymbol{q})\right)\right)\left(\boldsymbol{1}_M^\top \otimes \begin{bmatrix}\boldsymbol{1}_R^\top & R\end{bmatrix}\right).\end{aligned} \tag{41}$$

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

$$\begin{aligned}(\boldsymbol{1}_M^\top \otimes \boldsymbol{I}_2)(\boldsymbol{p}+\boldsymbol{q}) &= (\boldsymbol{1}_M^\top \otimes (\boldsymbol{I}_2 - \frac{1}{2}\boldsymbol{1}_2\boldsymbol{1}_2^\top))(\boldsymbol{p}+\boldsymbol{q}) = \frac{1}{2}\left(\boldsymbol{1}_M^\top \otimes \begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix}\right)(\boldsymbol{p}+\boldsymbol{q})\\ &= \frac{1}{2}\begin{pmatrix}1\\ -1\end{pmatrix}\left(\boldsymbol{1}_M^\top \otimes \begin{pmatrix}1 & -1\end{pmatrix}\right)(\boldsymbol{p}+\boldsymbol{q}) = \frac{1}{2}\begin{pmatrix}1\\ -1\end{pmatrix}(\overline{\boldsymbol{\Phi}}_K\boldsymbol{e}_0)^\top(\boldsymbol{p}+\boldsymbol{q}).\end{aligned} \tag{42}$$

Normalizing each row and column vector using the above simplification in (42) would lead to the desired result. ■

Remark 18 *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.*

The following lemma will show how Theorem 16 and Lemma 17 can be used to give a factorization of the matrix $\boldsymbol{Q}$:

Lemma 19 *Suppose a matrix $\boldsymbol{M}$, for given $\boldsymbol{U}, \boldsymbol{V}, \widetilde{\boldsymbol{v}}$ and $\boldsymbol{B} := \boldsymbol{U}^\top\boldsymbol{Q}\boldsymbol{V}$ 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. ■

Since we want the softmax spectral argument and output to share their structure, we add a term involving α to the block matrix,

Theorem 20 *Suppose the matrix $\boldsymbol{A} \in \mathbb{R}^{(K-1)\times(K-1)}$ has the block structure*

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

Then, for $\boldsymbol{U}, \boldsymbol{V}$ defined as in (27), 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

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

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

$$\begin{aligned}\boldsymbol{x} &\coloneqq \frac{1}{2M}\left(\sqrt{\frac{2}{R+1}}(a + \frac{\alpha}{\sqrt{R}})\mathbf{1}_M \otimes \begin{pmatrix}1\\-1\end{pmatrix} + 2\sum_{m=0}^{M-2}\overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_m \otimes \frac{\boldsymbol{a}_{m,0}}{\sqrt{R}}\right)\\ \boldsymbol{y} &\coloneqq \frac{1}{2M}\left(\sqrt{\frac{2}{R+1}}(-a + \alpha\sqrt{R})\mathbf{1}_M \otimes \begin{pmatrix}1\\-1\end{pmatrix} + 2\sum_{m=0}^{M-2}\overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_m \otimes \boldsymbol{a}_{m,1}\right).\end{aligned} \tag{43}$$

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

$$\boldsymbol{B} = \operatorname{diag}(b, \boldsymbol{B}_0, \ldots \boldsymbol{B}_{M-2}), \quad \boldsymbol{B}_i \in \mathbb{R}^{2\times 2}, \tag{44}$$

$$b = \frac{1}{\sqrt{2(R+1)}}(\overline{\boldsymbol{\Phi}}_K \boldsymbol{e}_0)^\top (R\boldsymbol{p} - \boldsymbol{q}), \quad \beta = \sqrt{\frac{R}{2(R+1)}}(\overline{\boldsymbol{\Phi}}_K \boldsymbol{e}_0)^\top (\boldsymbol{p} + \boldsymbol{q}), \tag{45}$$

$$\boldsymbol{B}_i = \left((\overline{\boldsymbol{\Phi}}_M \boldsymbol{e}_i)^\top \otimes \boldsymbol{I}_2\right)\begin{bmatrix}\sqrt{R}\boldsymbol{p} & \boldsymbol{q}\end{bmatrix} \quad \widetilde{\boldsymbol{v}} = \frac{\mathbf{1}_M \otimes \begin{pmatrix}\mathbf{1}_R\\ R\end{pmatrix}}{\sqrt{MR(R+1)}} \tag{46}$$

Proof All parts directly follow from Lemma 14, Theorem 16, Lemma 17, Lemma 19, and Proposition 15. Note that the additional $\alpha \boldsymbol{u}_0 \widetilde{\boldsymbol{v}}^\top$ introduces an additional

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

into the expansion (29), which has been accounted for in the the adjusted $\boldsymbol{x}, \boldsymbol{y}$ expressions. ■

Theorem 21 (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*

$$\begin{aligned}\boldsymbol{p}_+ &\coloneqq \boldsymbol{e}_0 - \mathbb{S}(\boldsymbol{x}_+ \otimes \boldsymbol{e}_+), \qquad \boldsymbol{x}_+ \coloneqq \frac{1}{M}\left(\sqrt{\frac{2}{R+1}}(a + \frac{\alpha}{\sqrt{R}})\mathbf{1}_M + \frac{1}{\sqrt{R}}\overline{\boldsymbol{\Phi}}_M \boldsymbol{a}_+\right),\\ \boldsymbol{p}_- &\coloneqq \boldsymbol{e}_1 - \mathbb{S}(\boldsymbol{x}_- \otimes \boldsymbol{e}_-), \qquad \boldsymbol{x}_- \coloneqq \frac{1}{M}\left(\sqrt{\frac{2}{R+1}}(a - \alpha\sqrt{R})\mathbf{1}_M + \overline{\boldsymbol{\Phi}}_M \boldsymbol{a}_-\right),\end{aligned} \tag{47}$$

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

$$\begin{aligned}\beta &= \sqrt{\frac{R}{2(R+1)}}(\overline{\boldsymbol{\Phi}}_K\boldsymbol{e}_0)^\top(\boldsymbol{p}_+ + \boldsymbol{p}_-), & b &= \frac{1}{\sqrt{2(R+1)}}(\overline{\boldsymbol{\Phi}}_K\boldsymbol{e}_0)^\top(R\boldsymbol{p}_+ - \boldsymbol{p}_-),\\ \boldsymbol{b}_+ &= \sqrt{R}(\overline{\boldsymbol{\Phi}}_M\otimes\boldsymbol{e}_+)^\top\boldsymbol{p}_+, & \boldsymbol{b}_- &= (\overline{\boldsymbol{\Phi}}_M\otimes\boldsymbol{e}_-)^\top\boldsymbol{p}_-.\end{aligned} \tag{48}$$

F.2. Simplified form of expressions

Lemma 22 *For $\beta, b, \boldsymbol{b}_+$, and $\boldsymbol{b}_-$ defined as in (48), the following simplified expressions hold:*

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

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 and $\boldsymbol{b}_+ \leq \sqrt{R}\mathbf{1}_{M-1}$ 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 (48), 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 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 $\boldsymbol{b}_+ = \sqrt{R}\mathbf{1}_{M-1}$, $\boldsymbol{b}_- = \mathbf{1}_{M-1}$. Also,

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

since $\boldsymbol{\Psi}_M$ only has entries in $\{0,2\}^{(M-1)\times(M-1)}$. The same reasoning applies to $\boldsymbol{b}_-$ to show $\boldsymbol{b}_- \geq \mathbf{0}$. Finally, to prove that $\boldsymbol{b}_- \leq \mathbf{1}_{M-1}$ for $\boldsymbol{a}_- \geq \mathbf{0}$, we proceed to show that $\boldsymbol{\Phi}_M^\top \exp(\boldsymbol{x}_-) \geq 0$. 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 for $\boldsymbol{b}_+$ to show that for $\boldsymbol{b}_+ \leq \sqrt{R}\mathbf{1}$ for $\boldsymbol{a}_+ \geq \mathbf{0}$. ■

Appendix G. Proofs for Regularization-path

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

We begin with a proof sketch before giving the formal argument. 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 (3)—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 $\boldsymbol{Z} = \boldsymbol{U}\,\mathrm{diag}_{+1}(\boldsymbol{a})\,\widetilde{\boldsymbol{V}}^\top$, Theorem 1 rewrites the loss gradient as $\boldsymbol{U}\,\mathrm{diag}_{+1}(\mathcal{F}(\boldsymbol{a}))\,\widetilde{\boldsymbol{V}}^\top$, so the condition becomes $\boldsymbol{U}\,\mathrm{diag}_{+1}(\mathcal{F}(\boldsymbol{a}))\,\widetilde{\boldsymbol{V}}^\top \in \lambda\,\partial\|\boldsymbol{U}\,\mathrm{diag}_{+1}(\boldsymbol{a})\,\widetilde{\boldsymbol{V}}^\top\|_*$. 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 (3). Since this is precisely the optimality condition for (3), any of its solution 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 (3) 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 23 (KKT conditions for regularized CE in the diagonal basis) *Consider the optimization problem*

$$\min_{\alpha\in\mathbb{R},\ \boldsymbol{A}\in\mathbb{R}^{(k-1)\times(k-1)}} \mathrm{CE}(\boldsymbol{L}) + \lambda\|\boldsymbol{L}\|_*, \qquad \text{s.t. } \boldsymbol{L} = \boldsymbol{U}\begin{bmatrix}\alpha\boldsymbol{e}_0 & \boldsymbol{A}\end{bmatrix}\widetilde{\boldsymbol{V}}^\top, \tag{49}$$

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 (49) satisfies

$$\boldsymbol{Y} - \mathbb{S}(\boldsymbol{L}) = \boldsymbol{U}\begin{bmatrix}\beta \boldsymbol{e}_0 & \boldsymbol{B}\end{bmatrix}\widetilde{\boldsymbol{V}}^\top, \qquad \boldsymbol{B} = \operatorname{diag}(b, \boldsymbol{b}_+ \otimes \boldsymbol{e}_+ + \boldsymbol{b}_- \otimes \boldsymbol{e}_-), \tag{50}$$

where

$$(\beta, b) \in \lambda\, \partial\sqrt{\alpha^2 + a^2}, \tag{51}$$

$$\boldsymbol{b}_\pm \in \lambda \partial \|\boldsymbol{a}_\pm\|_1, \tag{52}$$

Explicitly, this yields

$$(\beta, b) = \lambda \frac{(\alpha, a)}{\sqrt{\alpha^2 + a^2}} \ \textit{if}\ (\alpha, a) \neq (0,0), \quad \beta^2 + b^2 \leq \lambda^2 \ \textit{otherwise}, \tag{53}$$

$$\boldsymbol{b}_\pm[j] = \lambda \operatorname{sign}(\boldsymbol{a}_\pm[j]) \ \textit{if}\ \boldsymbol{a}_\pm[j] \neq 0, \quad |\boldsymbol{b}_\pm[j]| \leq \lambda \ \textit{otherwise}. \tag{54}$$

Proof Theorem 21 guarantees that (50) 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 (49) is convex in the free parameters $(\alpha, \boldsymbol{A})$, and the KKT condition reduces to first-order subgradient optimality:

$$\boldsymbol{0} \in \nabla_{\alpha,\boldsymbol{A}} \operatorname{CE}(\mathcal{T}(\begin{bmatrix}\alpha\boldsymbol{e}_0 & \boldsymbol{A}\end{bmatrix})) + \lambda\, \partial_{\alpha,\boldsymbol{A}} \|\mathcal{T}(\begin{bmatrix}\alpha\boldsymbol{e}_0 & \boldsymbol{A}\end{bmatrix})\|_*. \tag{55}$$

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 [8, Theorem 3.43 (b)], the above expression can be simplified to

$$\boldsymbol{U}^\top(\boldsymbol{Y} - \mathbb{S}\left(\mathcal{T}(\begin{bmatrix}\alpha\boldsymbol{e}_0 & \boldsymbol{A}\end{bmatrix})\right)\widetilde{\boldsymbol{V}} = \begin{bmatrix}\beta\boldsymbol{e}_0 & \boldsymbol{B}\end{bmatrix} \in \lambda\, \partial \left\|\begin{bmatrix}\alpha\boldsymbol{e}_0 & \boldsymbol{A}\end{bmatrix}\right\|_*. \tag{56}$$

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

$$\left\|\begin{bmatrix}\alpha\boldsymbol{e}_0 & \boldsymbol{A}\end{bmatrix}\right\|_* = \sqrt{\alpha^2 + a^2} + \|\boldsymbol{a}_+\|_1 + \|\boldsymbol{a}_-\|_1, \tag{57}$$

Applying the subgradient formula to (56) 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 (51) and (52). 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 $\boldsymbol{L} = \boldsymbol{U}\begin{bmatrix}\alpha\boldsymbol{e}_0 & \boldsymbol{A}\end{bmatrix}\widetilde{\boldsymbol{V}}^\top$ 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 24 *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 [11, 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 25 *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 [11, 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 26 (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 21. If

$$(\beta, b) \in \lambda\, \partial\sqrt{\alpha^2 + a^2}, \tag{58}$$

$$\boldsymbol{b}_\pm[j] \in \lambda\, \partial|\boldsymbol{a}_\pm[j]|, \qquad j \in [k/2] \tag{59}$$

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

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

Proof Consider the objective

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

Notice that the cross-entropy term is invariant to centering each column; that is, $\mathrm{CE}(\boldsymbol{L}) = \mathrm{CE}(\widehat{\boldsymbol{L}})$, 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 25 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

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

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 24, 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

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

■

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 27 (Basic properties of regulaization path solution) *The optimal solution to* (49) *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

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 (5) 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 $\lambda \geq \sqrt{\frac{R+1}{2}}$.

■

G.1.1. REDUCTION TO FOUR VARIABLES

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

Proposition 28 (Reduction to four variables) *There exist scalars* $a_{+,\lambda}$ *and* $a_{-,\lambda}$ *such that the minimizer of* (3) *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 29 (see also 13), 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 (3) 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 29 (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*

$$\boldsymbol{P}\overline{\boldsymbol{\Phi}}_M = \overline{\boldsymbol{\Phi}}_M \boldsymbol{C}. \tag{60}$$

For example, this is guaranteed to exist by Lemma 13. Then the vectorized objective in (3)

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

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 21 only requires the structure of

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

instead of the full matrix $\boldsymbol{Q}$. Consequently, the block columns $\boldsymbol{Q}'_i$, defined in similar fashion to Theorem 21 satisfy

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

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.

■

G.2. Proof of Theorem 3: Analytic solution

In this subsection we solve the four-dimensional convex problem from (3) 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 30 *Let $0 \le t \le 1,\ \gamma \in \mathbb{R}$, and suppose $\boldsymbol{r} \in \mathbb{R}^M$ satisfies*

$$\boldsymbol{r} = \exp\left(\frac{1}{M}(\gamma \mathbf{1}_M + \overline{\boldsymbol{\Phi}}_M \boldsymbol{\eta})\right), \qquad \overline{\boldsymbol{\Phi}}_M^\top \boldsymbol{r} = t\big(M + \mathbf{1}_M^\top \boldsymbol{r}\big)\mathbf{1}_{M-1}, \tag{61}$$

Then, the solution to this equation satisfies

$$\boldsymbol{\eta} = (\gamma - M\log(s))\,\mathbf{1}_{M-1}, \quad s^{M-1}((1+\theta)s + \theta) = e^\gamma, \quad \theta := \frac{Mt}{1-t}, \tag{62}$$

where $0 < s \le \exp(\gamma/M)$ is the unique solution of the scalar equation in (62)

Proof The first step is to multiply (61) by $\overline{\boldsymbol{\Phi}}_m$ given that $\overline{\boldsymbol{\Phi}}_M \overline{\boldsymbol{\Phi}}_M^\top = M\boldsymbol{I}_M - \mathbf{1}_M\mathbf{1}_M^\top$ to reach

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

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

$$\log(\boldsymbol{r}) = \log(\frac{\kappa + t(1+\kappa)(M-1)}{\kappa - t(1+\kappa)})\boldsymbol{e}_0 + \log(\kappa - t(1+\kappa))\mathbf{1}_M, \tag{63}$$

$$\boldsymbol{\eta} = \overline{\boldsymbol{\Phi}}_M^\top \log(\boldsymbol{r}) = \log(\frac{\kappa + t(1+\kappa)(M-1)}{\kappa - t(1+\kappa)})\mathbf{1}_M, \tag{64}$$

$$\gamma = \mathbf{1}_M^\top \log(\boldsymbol{r}) = \log\left((\kappa + t(1+\kappa)(M-1))\,(\kappa - t(1+\kappa))^{M-1}\right), \tag{65}$$

after which the desired result can be reached by defining $s := \kappa - t(1+\kappa)$ 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:

$$\Psi_\lambda(x) := \left(\frac{M}{\lambda(\theta_{-,\lambda} + M)}\right)\sqrt{\frac{2(1+x)^2}{(1+x)^2 - 2}} - 1, \quad x > \sqrt{2} - 1 \tag{66a}$$

$$H_\lambda(x) := Q(\Psi_\lambda(x), \theta_{-,\lambda}) - \left(\frac{M\sqrt{R}}{\lambda(\theta_{-,\lambda} + M)}\right)\frac{1+x}{1+\Psi_\lambda(x)}Q(x, \theta_{+,\lambda}) \tag{66b}$$

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 31 *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 $\xi_\lambda := \frac{2\lambda}{\sqrt{2\lambda^2-1}} - 1$ and $\theta_\lambda := \theta_{+,\lambda}$. Then, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda} Q(\xi_\lambda, \theta_\lambda) &= \frac{\partial Q}{\partial \xi_\lambda}\frac{\mathrm{d}\xi_\lambda}{\mathrm{d}\lambda} + \frac{\partial Q}{\partial \theta_\lambda}\frac{\mathrm{d}\theta_\lambda}{\mathrm{d}\lambda} \\ &= \frac{\partial Q}{\partial \xi_\lambda}\frac{-2}{(2x^2-1)^{3/2}} + \frac{\partial Q}{\partial \theta_\lambda}\frac{-m\sqrt{R}}{\lambda^2} < 0,\end{aligned} \tag{67}$$

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 32 *Let $\lambda_{K,R} \in [1, \sqrt{\frac{R+1}{2}}]$ be the point defined in Lemma 31. 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 $\lambda \leq \lambda_{K,R}$, (67) 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 33 *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 34 *For each λ, let z_λ denote the unique zero of $Q(x, \theta_+)$, defined in (6). Then, for any $0 < \lambda < \lambda' \le \sqrt{R}$, we have $z_\lambda < z_{\lambda'}$.*

Proof Let $\xi_\lambda \coloneqq \theta_+$ 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.

G.3. Analytic Solution of Theorem 3

Theorem 35 (Phase-wise analytic solution) *Let $\lambda_{K,R} \in [1, \sqrt{\frac{R+1}{2}}]$ be the unique root of*

$$Q(\frac{2\lambda}{\sqrt{2\lambda^2-1}} - 1, \theta_+) \tag{68}$$

Then the unique solution $\boldsymbol{a}_\lambda$ of the minimization in (3) is given as follows:

(i) *For $\lambda \in [\sqrt{R}, \infty)$, $\boldsymbol{a}_\lambda = \begin{bmatrix} 0 & 0 & 0 & 0 \end{bmatrix}^\top$*

(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

$$\gamma_\lambda = Q(s_\lambda, \theta_{+,\lambda}), \quad \delta_\lambda = Q(r_\lambda, \theta_{-,\lambda}). \tag{69}$$

(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

$$\gamma_\lambda = Q(s_\lambda, \theta_{+,\lambda}), \quad \delta_\lambda = Q(r_\lambda, \theta_{-,\lambda}). \tag{70}$$

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 31.

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 23 then guarantees that the proposed formulas hold in every phase.

(i) This case would follow trivially from the sparsity features proved in Lemma 27.

(ii) Based on Lemma 27, 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 $\boldsymbol{b}_+ = \lambda \boldsymbol{1}_{M-1}$, after which Lemma 30 (with $t = 1 - \frac{\lambda}{\sqrt{R}}$) would result in $\boldsymbol{a}_+ = -M\sqrt{R}\log s_\lambda$, where $Q(s_\lambda, \theta_+) = 0$ and $s_\lambda \in (0, 1)$

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 34, we have $z \geq \frac{2\lambda_{K,R}}{\sqrt{2\lambda_{K,R}^2-1}} - 1$, 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 27 would require $\boldsymbol{a}_+ \neq \mathbf{0}$ and $\boldsymbol{a}_- = \mathbf{0}$ for any solution. Assuming all entries in $\boldsymbol{a}_+$ are non-zero would again give $\boldsymbol{b}_+ = \lambda \mathbf{1}_{M-1}$, after which Lemma 30 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 27 guarantees $\gamma, \delta \geq 0$, they can be reformulated equivalently as

$$\lambda^2 = \beta^2 + b^2 = \frac{1}{2}(RX_+^2 + X_-^2), \quad \frac{\gamma}{X_+} = \frac{\delta}{X_-}. \tag{71}$$

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

$$\frac{1}{2} = \frac{1}{(1+s)^2} + \frac{1}{\lambda^2(1+r)^2}, \quad \lambda\delta\,(1+r) = \sqrt{R}\gamma\,(1+s)$$

The existence of $s \in (\sqrt{2}-1, \xi_\lambda)$ and the fact that all terms in (69) are well-defined follow from Proposition 32. 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 (7).

(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 27 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 30 would give

$$\begin{gathered} \boldsymbol{a}_+ = \sqrt{R}(\gamma - M \log s)\mathbf{1}_{M-1}, \quad \boldsymbol{a}_- = (\delta - M \log r)\mathbf{1}_{M-1}, \\ \gamma = Q(s, \theta_+), \quad \delta = Q(r, \theta_-). \end{gathered} \tag{72}$$

In addition, the reformulated KKT conditions for (α, a) of (71) would simplify to the slightly different form

$$\frac{1}{2} = \frac{1}{(1+s)^2} + \frac{1}{(1+r)^2}, \quad \delta\,(1+r) = \sqrt{R}\gamma\,(1+s) \tag{73}$$

The above system of four equations can be reformulated as a scalar equation in s, where Proposition 33 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 (7). ■

Next, we will discuss more high-level characteristics of the regularization path solution given the analytic parametrization for different phases.

Proposition 36 (More characteristics of the regularization path solution) *For all values of λ, the regularization path solution parameters satisfy the following properties:*

$$\forall \lambda \geq 0: \quad a_{+,\lambda} \geq \sqrt{R}\, a_{-,\lambda}, \quad \alpha_\lambda \leq 0 \tag{74}$$

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

$$\begin{aligned} a_+ = \sqrt{R}(\gamma - M \log s), \quad & s^{M-1}((1+\theta_+)s + \theta_+) = e^\gamma \Rightarrow a_+ = \sqrt{R} \log(1 + \theta_+ + \frac{\theta_+}{s}) \\ a_- = (\delta - M \log r), \quad & r^{M-1}((1+\theta_+)r + \theta_-) = e^\delta \Rightarrow a_- = \log(1 + \theta_- + \frac{\theta_-}{r}) \end{aligned} \tag{75}$$

We know that $s \leq r$ from Proposition 33 and $\theta_+ = \sqrt{R}\theta_- + M(\sqrt{R} - 1) \geq \theta_-$. For $\lambda \leq 1$, we therefore conclude that

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

This establishes the first claim for all $\lambda > 0$. 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 $a_- \leq \frac{a_+}{\sqrt{R}}$, the above expression would become negative, contradicting the prior assumption. Hence, $\alpha \leq 0$, with strict inequality in its active phase.

■

G.4. Behavior of the threshold $\lambda_{K,R}$

Proposition 37 *Consider the threshold $\lambda_{K,R} \in [1, \sqrt{\frac{R+1}{2}}]$ between phases (iii) and (iv) in Theorem 35, with its value defined by (68). Then, it has the following properties:*

1. *$\lambda_{K,R}$ is increasing 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

$$1 = \frac{2}{(1+s)^2} + \frac{1}{2\lambda_{K,R}^2}, \quad Q(s, \theta_+) = \log\big(s^{M-1}((1+\theta_+)s+\theta_+)\big) = 0, \quad \theta_+ = M[\sqrt{R}\lambda_{K,R}^{-1} - 1] \tag{76}$$

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

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

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

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

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

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

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

G.5. Asymptotic behavior of RP solution and Proof of Theorem 4

Before analyzing the asymptotic behavior of the regularization path solution, it is useful to restate and refine the upper and lower bounds for s, r for $\lambda \in (0, 1]$ in the following lemma:

Lemma 38 *For the regularization path characterized (72) and (73), the following bounds can be established:*

$$\sqrt{\frac{2(R+1)}{R}} - 1 \leq s \leq 1, \quad 1 \leq r \leq \sqrt{2(R+1)} - 1 \tag{77}$$

Proof Recall that $s \in (\sqrt{2} - 1, 1)$ and $r \in (1, \infty)$ has already been established in Proposition 33. To reach the refined bounds in the lemma, notice that $\alpha \leq 0$ from Theorem 36 results in $\delta \geq \gamma$. Substituting this bound in (73) and simplifying proves the desired bounds. ■

Proposition 39 (Asymptotic direction as $\lambda \to 0^+$) *For each $\lambda \in (0, 1)$, consider the solution of the KKT system in (72) and (73) 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 $\boldsymbol{a}_\infty := (0, \sqrt{\frac{R+1}{2}}, \sqrt{R}\mathbf{1}_{M-1}, \mathbf{1}_{M-1})$. Also define the norm operator $\|\boldsymbol{a}\|_{+1} := \sqrt{a^2 + \alpha^2} + \|\boldsymbol{a}_+\|_1 + \|\boldsymbol{a}_-\|_1$. Then the following limit and rate of convergence hold:

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

Proof We bound the quantities using (73), along with the bounds established in Lemma 38.

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

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

Since s, r are bounded from above and below by Lemma 38,

$$\begin{gathered} \gamma = \log(1+\theta_+) + c_\gamma, \quad M\log(\sqrt{\frac{2(R+1)}{R}} - 1) \leq c_\gamma \leq -\log(1 - \sqrt{\frac{R}{2(R+1)}}) \\ \delta = \log(1+\theta_-) + c_\delta, \quad 0 \leq c_\delta \leq \log(2) + \log(\sqrt{2(R+1)} - 1) \\ \frac{a_+}{\sqrt{R}} = \log(1+\theta_+) + c_{a_+}, \quad 0 \leq c_{a_+} \leq -\log(1 - \sqrt{\frac{R}{2(R+1)}}) \\ \frac{a_-}{\sqrt{R}} = \log(1+\theta_-) + c_{a_-}, \quad 0 \leq c_{a_-} \leq \log(2) \end{gathered} \tag{78}$$

Since $\log(1+\theta_\pm)$ scale as $\mathcal{O}(\log(\lambda^{-1}))$, it is evident from the above bounds in (78) 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:

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

■

Beyond the above formal characterization of the limit of the regularization path, we further conjecture that the singular value associated with the majority class converges to its limit from above; in other words,

$$\frac{a_{+,\lambda}}{\|\boldsymbol{a}_\lambda\|_{+1}} - \frac{a_{+,\infty}}{\|\boldsymbol{a}_\infty\|_{+1}} \geq 0, \quad \lim_{\lambda\to 0^+} \frac{a_{+,\lambda}}{\|\boldsymbol{a}_\lambda\|_{+1}} - \frac{a_{+,\infty}}{\|\boldsymbol{a}_\infty\|_{+1}} = 0.$$

Supporting Argument The limit statement follows directly from Proposition 39. For the inequality claim, we use the following inequality, observed consistently in empirical simulations of the regularization path solution:

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

Next, we simplify the expression as

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

The denominator is positive. The first term in the numerator is nonnegative as a result of Theorem 36, and the second term is nonnegative as a result of the empirical inequality.

G.6. Alternative proof of regularization-path limit

For completeness, we verify here that the max-margin problem that is associated with the regularization path limit of (3) has SEL as its unique solution. Note that Theorem 4 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 [42, 43] and used extensively in the recent literature [30, 32, 50, e.g.], that the solution of (3) directionally converges in the limit $\lambda \to 0$ to the (normalized) max-margin solution below:

$$\begin{aligned} \min_{\boldsymbol{a}_+,\boldsymbol{a}_-,a,\alpha} \quad & \|\boldsymbol{a}_+\|_1 + \|\boldsymbol{a}_-\|_1 + \sqrt{a^2+\alpha^2} \\ \text{s.t.} \quad & \forall i \in [M-1] : \frac{1}{\sqrt{R}}(\mathbf{1}-\overline{\boldsymbol{\phi}}_{i+1})^T\boldsymbol{a}_+ \geq 1, \quad \frac{1}{\sqrt{R}}\mathbf{1}^T\boldsymbol{a}_+ + \sqrt{\frac{2}{R+1}}\left(a + \frac{\alpha}{\sqrt{R}}\right) \geq 1 \\ & \forall i \in [M-1] : (\mathbf{1}-\overline{\boldsymbol{\phi}}_{i+1})^T\boldsymbol{a}_- \geq 1, \quad \mathbf{1}^T\boldsymbol{a}_- + \sqrt{\frac{2}{R+1}}\left(a - \sqrt{R}\alpha\right) \geq 1 \end{aligned} \tag{79}$$

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 4. This also yields an alternative proof of [52, Theorem 1, Proposition 2]. Note that compared to our approach in Theorem 4 the overall proof here is not as explicit (cf. proof by contradiction) and does not yield rates of convergence.

Theorem 40 *The optimization problem of* (79) *admits the unique solution of*

$$\boldsymbol{a}_+ = \frac{1}{M}\sqrt{R}\mathbf{1}_{M-1}, \quad \boldsymbol{a}_- = \frac{1}{M}\mathbf{1}_{M-1}, \quad a = \frac{1}{M}\sqrt{\frac{R+1}{2}}, \quad \alpha = 0 \tag{80}$$

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

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

The KKT conditions in terms of the dual variables are

$$\sqrt{R}\,\mathrm{sign}(\hat{\boldsymbol{a}}_+) = \boldsymbol{\Psi}\boldsymbol{\nu}_+ + \lambda_+\mathbf{1}, \quad \mathrm{sign}(\hat{\boldsymbol{a}}_-) = \boldsymbol{\Psi}\boldsymbol{\nu}_- + \lambda_-\mathbf{1}, \tag{81a}$$

$$\sqrt{\frac{2}{R+1}}(\lambda_+ + \lambda_-, \frac{\lambda_+}{\sqrt{R}} - \lambda_-\sqrt{R}) \in \partial_{(\hat{a},\hat{\alpha})}(\sqrt{\hat{a}^2 + \hat{\alpha}^2}) \tag{81b}$$

$$\boldsymbol{\nu}_\pm \geq \mathbf{0}, \quad \lambda_\pm \geq 0, \quad \boldsymbol{\nu}_\pm \odot (\boldsymbol{\Psi}\hat{\boldsymbol{a}}_\pm - \mathbf{1}) = \mathbf{0}, \tag{81c}$$

$$\lambda_+(\mathbf{1}^T\hat{\boldsymbol{a}}_+ + (\hat{a} + \frac{\hat{\alpha}}{\sqrt{R}}) - 1) = 0, \quad \lambda_-(\mathbf{1}^T\hat{\boldsymbol{a}}_- + (\hat{a} - \hat{\alpha}\sqrt{R}) - 1) = 0, \tag{81d}$$

where the sign function is treated in an extended value form. Notice that since the entries of $\boldsymbol{\Psi}$ are nonnegative, (81a) alongside the dual constraints in (81c) 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 to show 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

$$\boldsymbol{\nu}_+ = \boldsymbol{\Psi}^{-1}(\sqrt{R} - \lambda_+)\mathbf{1} = \frac{1}{M}(\sqrt{R} - \lambda_+)\mathbf{1}, \boldsymbol{\nu}_- = \frac{1}{M}(1 - \lambda_-)\mathbf{1} \tag{82a}$$

$$\sqrt{\frac{2}{R+1}}(\lambda_+ + \lambda_-, \frac{\lambda_+}{\sqrt{R}} - \lambda_-\sqrt{R}) = \frac{1}{\sqrt{\hat{a}^2 + \hat{\alpha}^2}}(\hat{a}, \hat{\alpha}). \tag{82b}$$

Notice that (82b) implies that $\frac{\lambda_+^2}{R} + \lambda_-^2 = \frac{1}{2}$. This constraint eliminates the possibility for $\lambda_+ = \sqrt{R}$ or $\lambda_- = 1$, from which 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 of $\lambda_+ = 0$ or $\lambda_- = 0$ will lead to violation of one of the above simplified constraints; consequently, $\lambda_+, \lambda_- \neq 0$, which leads to the scalar constraints being active. Solving for $\hat{a}, \hat{\alpha}$ will give the solution in the theorem statement.

So far, (80) 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 (79). 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

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

The scalar constraints above hold for (80) and the terms $\mathbf{1}^T\hat{\boldsymbol{a}}_+, \mathbf{1}^T\hat{\boldsymbol{a}}_-$ in the objective function as well. As a result, the scalar constraints are active for every other solution. Use the fact that $\mathbf{1}^T\boldsymbol{\Psi} = M\mathbf{1}^T$ to conclude that while every element of the vector $\boldsymbol{\Psi}(\hat{\boldsymbol{a}}_+ - \frac{1}{M}\mathbf{1})$ is nonnegative, but 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.

■

Appendix H. Proofs for Gradient-flow path

H.1. Proof of Theorem 6

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 $\boldsymbol{H} = \boldsymbol{R}\begin{bmatrix}\alpha_H\boldsymbol{e}_+ & \boldsymbol{W}_\star\end{bmatrix}\widetilde{\boldsymbol{V}}^\top$. Then we have $\boldsymbol{W}\boldsymbol{H} = \boldsymbol{U}\begin{bmatrix}\boldsymbol{W}_\star\alpha_H\boldsymbol{e}_+ & \boldsymbol{W}_\star\boldsymbol{W}_\star\end{bmatrix}\widetilde{\boldsymbol{V}}^\top$ and

$$\dot{\boldsymbol{W}} = \boldsymbol{U}\begin{bmatrix}\beta\boldsymbol{e}_0 & \boldsymbol{B}\end{bmatrix}\underbrace{\widetilde{\boldsymbol{V}}^\top\widetilde{\boldsymbol{V}}}_{\boldsymbol{I}}\begin{bmatrix}\alpha_H e_0 & \widetilde{\boldsymbol{H}}\end{bmatrix}^\top\boldsymbol{R}^\top = \boldsymbol{U}(\beta\alpha_H\boldsymbol{e}_0\boldsymbol{e}_0^\top + \boldsymbol{B}\widetilde{\boldsymbol{H}})\boldsymbol{R}^\top$$

$$\dot{\boldsymbol{H}} = \boldsymbol{R}\boldsymbol{W}_\star\begin{bmatrix}\beta\boldsymbol{e}_0 & \boldsymbol{B}\end{bmatrix}\widetilde{\boldsymbol{V}}^\top = \boldsymbol{R}\begin{bmatrix}\beta(\boldsymbol{W}_\star)_0\boldsymbol{e}_0 & \boldsymbol{W}_\star\boldsymbol{B}\end{bmatrix}\widetilde{\boldsymbol{V}}^\top.$$

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 $\frac{\mathrm{d}}{\mathrm{d}t}(\boldsymbol{W}^\top\boldsymbol{W} - \boldsymbol{H}\boldsymbol{H}^\top) = \mathbf{0}$. Combining this with the factorizations of $\boldsymbol{W}, \boldsymbol{H}$, we have

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

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

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

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

$$\dot{\boldsymbol{w}_\pm} = \boldsymbol{b}_\pm \odot \boldsymbol{h}_\pm, \quad \dot{\boldsymbol{h}_\pm} = \boldsymbol{b}_\pm \odot \boldsymbol{w}_\pm, \quad \dot{w} = \beta\alpha_H + bh, \quad \dot{h} = bw, \quad \dot{\alpha_H} = \beta\widetilde{w} \tag{83}$$

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

$$\boldsymbol{a}_\pm \coloneqq \boldsymbol{w}_\pm \odot \boldsymbol{h}_\pm, \quad \dot{\boldsymbol{a}_\pm} = \boldsymbol{b}_\pm \odot (\boldsymbol{w}_\pm^2 + \boldsymbol{h}_\pm^2), \tag{84}$$

$$a \coloneqq wh, \quad \alpha \coloneqq \alpha_H w, \quad \dot{a} = b(w^2 + h^2) + \beta\alpha_H h, \quad \dot{\alpha} = \beta(w^2 + \alpha_H^2) + bh\,\alpha_H. \tag{85}$$

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

$$\boldsymbol{a}_\pm \coloneqq \boldsymbol{w}_\pm^2 \geq \boldsymbol{0}, \quad \dot{\boldsymbol{a}_\pm} = 2\boldsymbol{a}_\pm \odot \boldsymbol{b}_\pm, \quad w^2 = \sqrt{a^2 + \alpha^2}, \tag{86}$$

$$\dot{a} = b(w^2 + \frac{a^2}{w^2}) + \beta\frac{\alpha a}{w^2} = \frac{b(2a^2 + \alpha^2) + \beta a\alpha}{\sqrt{a^2 + \alpha^2}}, \tag{87}$$

$$\dot{\alpha} = \beta(w^2 + \frac{\alpha^2}{w^2}) + b\frac{a\alpha}{w^2} = \frac{\beta(2\alpha^2 + a^2) + ba\alpha}{\sqrt{a^2 + \alpha^2}}. \tag{88}$$

H.2. Proof of Theorem 10: Majority and Minority tend to Uniformity

Theorem 41 (Restatement of Theorem 10) *Let $\boldsymbol{a}_{\pm,t} \in \mathbb{R}^{M-1}$ be variables initialized with $\boldsymbol{a}_{\pm,0} > \boldsymbol{0}$, and evolve under (23), for arbitrary a_0, α_0 initialization. For every vector $\boldsymbol{a}$, define its normalization as $\overline{\boldsymbol{a}} \coloneqq \boldsymbol{a}/\|\boldsymbol{a}\|_1$. Then, the Kullback-Leibler (KL) divergence $D_{KL}(\overline{\boldsymbol{1}}_{M-1}||\overline{\boldsymbol{a}_\pm})$ is decreasing over time, and $\dot{D}_{KL}(\overline{\boldsymbol{1}}_{M-1}||\overline{\boldsymbol{a}_\pm}) = 0$ iff $\overline{\boldsymbol{a}}_\pm = (M-1)^{-1}\boldsymbol{1}_{M-1}$.*

The proof closely follows the method of [19] for the balanced 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

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

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

where $\boldsymbol{q}_+ := (\sqrt{R}M)^{-1}\boldsymbol{a}_+,\ \boldsymbol{q}_- = \boldsymbol{a}_-$. The above expressions suggest a similar form in the numerator of both derivatives, with strictly positive denominators and coefficients. In the next step, define $\boldsymbol{z} := \overline{\boldsymbol{\Phi}}\boldsymbol{q}$, which directly establishes $\mathbf{1}^\top \boldsymbol{z} = (\mathbf{1}^\top \overline{\boldsymbol{\Phi}})\boldsymbol{q} = 0$. Recall that the first row of $\overline{\boldsymbol{\Phi}}$ is $\mathbf{1}^\top$, from which $z_0 = \mathbf{1}^\top \boldsymbol{q}$ 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 $\boldsymbol{z} = (M\boldsymbol{e}_0 - \mathbf{1})q$ for some scalar q. Using the definition $\boldsymbol{z} = \overline{\boldsymbol{\Phi}}\boldsymbol{q}$ and solving for $\boldsymbol{q}$ gives the unique answer $\boldsymbol{q} = q\mathbf{1}$ for the derivative to vanish.

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

H.3. Proof of Theorem 8

We prove the theorem by separately showing that majority emerges before minority (Proposition 42), and then, we show it emerges before maj-min (Proposition 43).

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. (23) under positive initialization satisfying, for some constants $C, c > 0$,

$$\|\boldsymbol{a}^\varepsilon(0)\|_\infty \leq C\varepsilon, \qquad \text{and} \qquad c\varepsilon \leq \|\boldsymbol{a}_+^\varepsilon(0)\|_\infty. \tag{89}$$

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 $T_+^\varepsilon < T_0^\varepsilon$.

Proposition 42 (Majority emerges first) *Assume (89). 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 43 (Majority mode emerges before the maj-min mode) *Assume (89). 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 $\varepsilon \in (0, \varepsilon_0(\eta))$,*

$$T_+^\varepsilon(\eta) < T_0^\varepsilon(\eta).$$

H.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 44 (Local control of $\boldsymbol{b}_\pm$) *Fix any $\rho > 0$. Then,*

$$\forall t \in [0, T_+(\rho)]\,, \qquad \mathbf{0} \le \sqrt{R}\,\mathbf{1} - \boldsymbol{b}_{+,t}(t) \le \sqrt{R}\,\tanh(\frac{\rho}{\sqrt{R}})\mathbf{1}, \tag{90a}$$

$$\forall t \in [0, T_-(\rho)]\,, \qquad \mathbf{0} \le \mathbf{1} - \boldsymbol{b}_{-,t}(t) \le \tanh(\frac{\rho}{2})\mathbf{1}. \tag{90b}$$

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 of Theorem 41, we have defined

$$\boldsymbol{z}_+ := \frac{1}{M\sqrt{R}}\overline{\boldsymbol{\Phi}}\boldsymbol{a}_+, \qquad \boldsymbol{z}_- := \frac{1}{M}\overline{\boldsymbol{\Phi}}\boldsymbol{a}_-.$$

The non-negativity bounds in (90) are proved in Lemma 22. Here, we prove the upper bounds. We will prove the bound in (90b), and the bound for (90a) follows in the exact same way and is omitted for brevity.

For each $j \in [M-1]$, define the index sets $S_j^\pm := \{m : \overline{\boldsymbol{\Phi}}_{mj} = \pm 1\}$. 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 assumption $\boldsymbol{a}_- \ge \mathbf{0}$ and that again that $\overline{\boldsymbol{\Phi}} \in \{\pm 1\}^{M\times(M-1)}$ has $M/2$ entries of -1 and $M/2 - 1$ entries of 1, we reach

$$\|\boldsymbol{z}_-\|_\infty = \left\|\frac{1}{M}\overline{\boldsymbol{\Phi}}\boldsymbol{a}_+\right\|_\infty \le \frac{M-1}{2M}\|\boldsymbol{a}_+\|_\infty \le \frac{\rho}{2}\,.$$

Substituting the above bound will result in

$$(\mathbf{1} - \boldsymbol{b}_-)_j \le \tanh(\frac{1}{2}\log(\frac{Me^{\rho/2}}{Me^{-\rho/2}})) = \tanh(\frac{\rho}{2}),$$

which is the desired result. ■

Lemma 45 (Local control of $X_\pm$) *Fix $\rho > 0$, and define the local stopping time*

$$\tau_\rho := \inf\Big\{t \ge 0 : \max\big\{\gamma_t(t), \delta_t(t), \|\boldsymbol{a}_{+,t}(t)\|_\infty, \|\boldsymbol{a}_{-,t}(t)\|_\infty\big\} = \rho\Big\}.$$

Then, for every $t \in [0, \tau_\rho)$,

$$\frac{2}{1+e^{2\rho}} =: \underline{X}(\rho) \le X_{+,t}, X_{-,t} \le \overline{X}(\rho) := \frac{2}{1+e^{-2\rho}},$$

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. ■

H.3.2. PROOF OF PROPOSITION 42

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 44, for every $t \in [0, T_+^\varepsilon(\rho)]$,

$$\|\boldsymbol{b}_{+,t} - \sqrt{R}\,\mathbf{1}\|_\infty \le \sqrt{R}\,\tanh\left(\frac{\rho}{2\sqrt{R}}\right) \le \sqrt{R}\,\tanh(\frac{\rho}{2}).$$

Hence, for every coordinate i and every $t \in [0, T_+^\varepsilon(\rho)]$,

$$\sqrt{R}\big(1 - \tanh(\frac{\rho}{2})\big) \le b_{+,i}(t) \le \sqrt{R},$$

and therefore

$$2\sqrt{R}\big(1 - \tanh(\frac{\rho}{2})\big)a_{+,i}(t) \le \dot{a}_{+,i}(t) \le 2\sqrt{R}\,a_{+,i}(t).$$

Applying Gronwall's inequality coordinatewise and taking maximum over i gives, for every $t \in [0, T_+^\varepsilon(\rho)]$,

$$\|\boldsymbol{a}_{+,0}^\varepsilon\|_\infty e^{2\sqrt{R}(1-\tanh(\frac{\rho}{2}))t} \le \|\boldsymbol{a}_{+,t}^\varepsilon\|_\infty \le \|\boldsymbol{a}_{+,0}^\varepsilon\|_\infty e^{2\sqrt{R}t}. \tag{91}$$

In the exact same manner, for every $t \in [0, T_-^\varepsilon(\rho)]$,

$$\|\boldsymbol{a}_{-,0}^\varepsilon\|_\infty e^{2(1-\tanh(\frac{\rho}{2}))t} \le \|\boldsymbol{a}_{-,t}^\varepsilon\|_\infty \le \|\boldsymbol{a}_{-,0}^\varepsilon\|_\infty e^{2t}. \tag{92}$$

Because $\eta < \rho$, we have

$$T_+^\varepsilon(\eta) \le T_+^\varepsilon(\rho), \qquad T_-^\varepsilon(\eta) \le T_-^\varepsilon(\rho),$$

and therefore (91) may be evaluated at $t = T_+^\varepsilon(\eta)$, while (92) 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

$$\frac{1}{2\sqrt{R}} \log \frac{\eta}{\|\boldsymbol{a}_+^\varepsilon(0)\|_\infty} \le T_+^\varepsilon(\eta) \le \frac{1}{2\sqrt{R}\big(1-\tanh(\frac{\rho}{2})\big)} \log \frac{\eta}{\|\boldsymbol{a}_+^\varepsilon(0)\|_\infty}, \tag{93}$$

$$\frac{1}{2} \log \frac{\eta}{\|\boldsymbol{a}_-^\varepsilon(0)\|_\infty} \le T_-^\varepsilon(\eta) \le \frac{1}{2\big(1-\tanh(\frac{\rho}{2})\big)} \log \frac{\eta}{\|\boldsymbol{a}_-^\varepsilon(0)\|_\infty}, \tag{94}$$

Because $R > 1$ and $\tanh(\frac{\rho}{2}) \to 0$ as $\rho \to 0$, we may choose $\eta_0 > 0$ such that for every $\rho \in (0, \eta_0)$, $\sqrt{R}\big(1-\tanh(\frac{\rho}{2})\big) > 1$. Fix any $\eta \in (0, \eta_0)$ and choose any $\rho \in (\eta, \eta_0)$. Then (93), (94), 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

$$T_+^\varepsilon(\eta) \le \frac{1}{2\sqrt{R}(1-\tanh(\frac{\rho}{2}))} \log \frac{\eta}{c\varepsilon}, \qquad T_-^\varepsilon(\eta) \ge \frac{1}{2} \log \frac{\eta}{C\varepsilon}.$$

By the condition $\sqrt{R}\big(1-\tanh(\rho/M)\big) > 1$, 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.

H.3.3. PROOF OF PROPOSITION 43

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

$$\dot a_0(t) = R\gamma(t)X_{+,t} + \delta(t)X_{-,t}. \tag{95}$$

By Lemma 45, 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 $[0, \tau_\rho)$,

$$\dot a_0(t) \le \overline{X}(\rho)\big(R\gamma(t)+\delta(t)\big) \le \sqrt{2(R+1)}\,\overline{X}(\rho)\,a_0(t), \tag{96}$$

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 44, exactly as in the proof of Proposition 42,

$$\|\boldsymbol{a}_{+,t}(t)\|_\infty \ge \|\boldsymbol{a}_+(0)\|_\infty e^{2\sqrt{R}(1-\tanh(\frac{\rho}{2}))t}, \qquad t \in [0, T_+(\rho)]. \tag{97}$$

Let

$$\mu(\rho) := 2\sqrt{R}(1-\tanh(\frac{\rho}{2})), \qquad \nu(\rho) := \sqrt{2(R+1)}\,\overline{X}(\rho).$$

Since $\tanh(\frac{\rho}{2}) \to 0$ 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 $\rho_{\text{rate}} > 0$ such that

$$\mu(\rho) > \nu(\rho) \qquad \text{for all } \rho \in (0, \rho_{\text{rate}}). \tag{98}$$

By Proposition 42, there exists $\rho_{\text{maj}} > 0$ such that its conclusion holds for every threshold $\rho \in (0, \rho_{\text{maj}})$. Define

$$\rho_0 := \min\{\rho_{\text{rate}}, \rho_{\text{maj}}\}, \qquad \eta_0 := \rho_0/2.$$

Fix any $\eta \in (0, \eta_0)$, and set $\rho := 2\eta$. Then $\rho \in (0, \rho_0)$, so both (98) and Proposition 42 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 42, 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

$$T < \tau_\rho\,, \tag{99}$$

and we can plug in T to (96) and (97).

Specifically, By (96), (99), and Gronwall's inequality,

$$a_0(T) \le a_0(0)e^{\nu(\rho)T}.$$

Since $a_0(T) = \eta$, this implies

$$e^T \geq \left(\frac{\eta}{a_0(0)}\right)^{1/\nu(\rho)}. \tag{100}$$

Using (97), (100), and $\|\boldsymbol{a}_+(0)\|_\infty \geq c\varepsilon$, we get $\|\boldsymbol{a}_{+,t}(T)\|_\infty \geq c\varepsilon \left(\frac{\eta}{a_0(0)}\right)^{\mu(\rho)/\nu(\rho)}$. Since

$$a_0(0) \leq C\sqrt{\frac{R+1}{2}}\,\varepsilon,$$

we obtain

$$\|\boldsymbol{a}_{+,T}\|_\infty \geq c\left(C\sqrt{\frac{R+1}{2}}\right)^{-\mu(\rho)/\nu(\rho)} \eta^{\mu(\rho)/\nu(\rho)}\varepsilon^{1-\mu(\rho)/\nu(\rho)}. \tag{101}$$

By (98), $\mu(\rho) > \nu(\rho)$. Hence the exponent $1 - \mu(\rho)/\nu(\rho)$ is negative, so the right-hand side of (101) diverges as $\varepsilon \to 0$. In particular, for all sufficiently small ε,

$$\|\boldsymbol{a}_{+,T}\|_\infty > \eta.$$

But $T = T_0^\varepsilon(\eta) \leq T_+^\varepsilon(\eta)$, so by definition of $T_+^\varepsilon(\eta)$, $\|\boldsymbol{a}_{+,T}\|_\infty \leq \eta$, a contradiction. Therefore $T_+^\varepsilon(\eta) < T_0^\varepsilon(\eta)$ completing the proof.

H.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 [19]. 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 46 (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

$$c := \sqrt{\frac{\kappa^2+1}{2}}. \tag{102}$$

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

$$\dot{V} = \sum_{i=0}^{2} \bar{z}_i r_i - \frac{1}{3}\sum_{i=0}^{2} r_i \tag{103}$$

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 (103)

$$\lim_{\varepsilon\to 0^+} \dot{V} = \frac{c(\kappa+1)/c + 2 + 2}{c+2} - \frac{1}{3}((\kappa+1)/c + 2 + 2) = \frac{\kappa+5}{c+2} - \frac{1}{3}\left(\frac{\kappa+1}{c} + 4\right).$$

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. ■

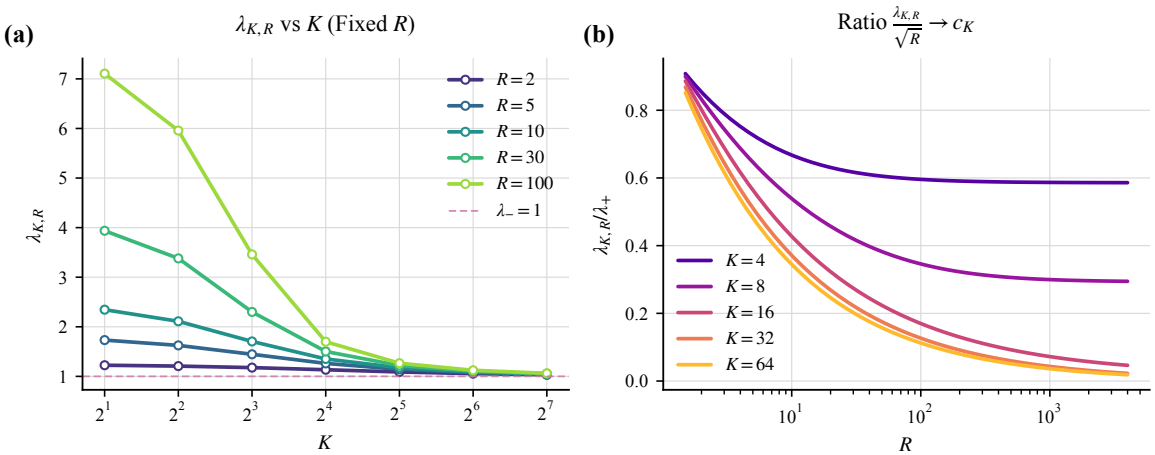


Figure 3: Behavior of regularization-path emergence thresholds for varying number of classes (K) and imbalance ratios (R).

Appendix I. Numerical Results

I.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 profile. Thus, the plots emphasize deviations from the limiting geometry rather than the overall divergence of the singular values.

I.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

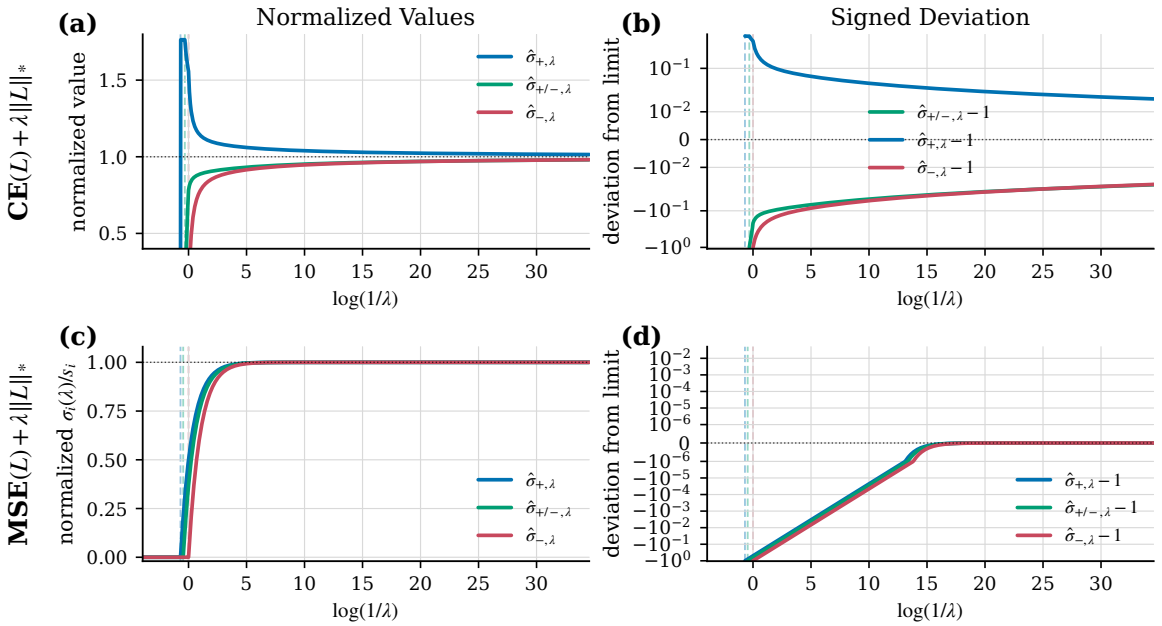
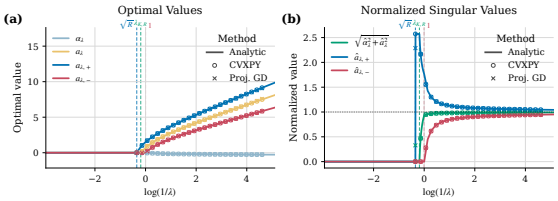


Figure 4: Comparison of the regularization-path trajectories for the three normalized singular coordinates under CE and MSE, shown for $R = 4$ and $K = 8$. The panels track the majority, majority–minority, and minority coordinates as functions of λ; horizontal dashed lines indicate their limiting normalized values

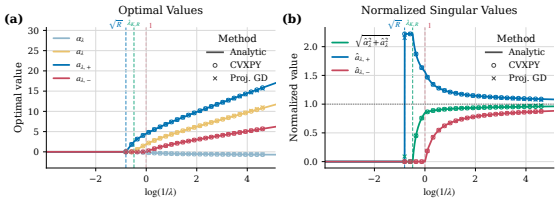
solution, the singular values obtained by solving the convex logit-space problem (Z-UFM$_\lambda$) in CVXPY, and the singular values obtained by projected gradient descent on the sphere of radius B corresponding to that value of λ. This comparison serves two purposes: it verifies the correctness of the 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 Figs. 5 and 6.

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 is not a power of two and the particular Hadamard construction is unavailable. As a numerical check, we compare in Fig. 7 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.

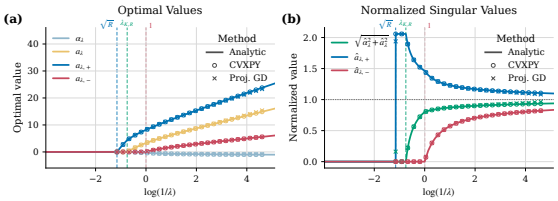
Role of the nuclear norm. The phase transitions identified in the regularization path are not a generic consequence of imbalance alone. To illustrate this, we repeat the same path computation after replacing the nuclear-norm penalty in (Z-UFM$_\lambda$) by a Frobenius-norm penalty in logit space. The resulting trajectories are shown in Fig. 8 for both CE and MSE losses.



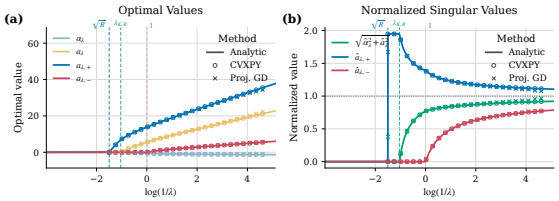
(a) $R = 2$



(b) $R = 5$

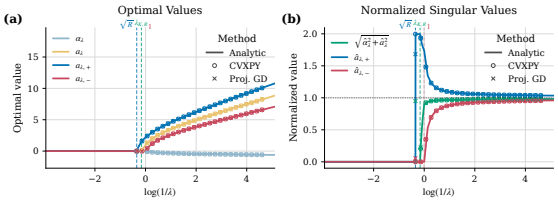


(c) $R = 10$

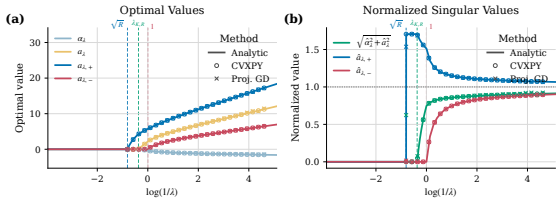


(d) $R = 20$

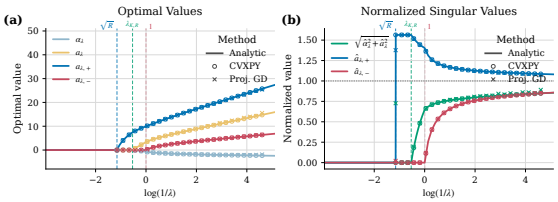
Figure 5: Numerical comparison of the regularization-path singular values for $K = 4$ across different imbalance ratios R. In each panel, the curves obtained from the analytic solution are overlaid with those produced by a direct CVXPY solve of (Z-UFM$_\lambda$) and by projected gradient descent on the sphere with radius B matched to the same value of λ.



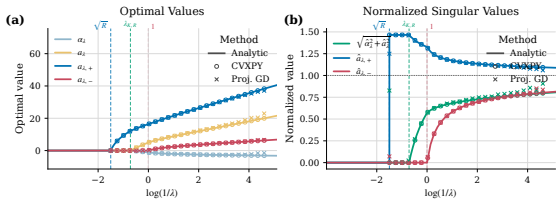
(a) $R = 2$



(b) $R = 5$



(c) $R = 10$



(d) $R = 20$

Figure 6: 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.

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 profile, all coordinates remain essentially constant

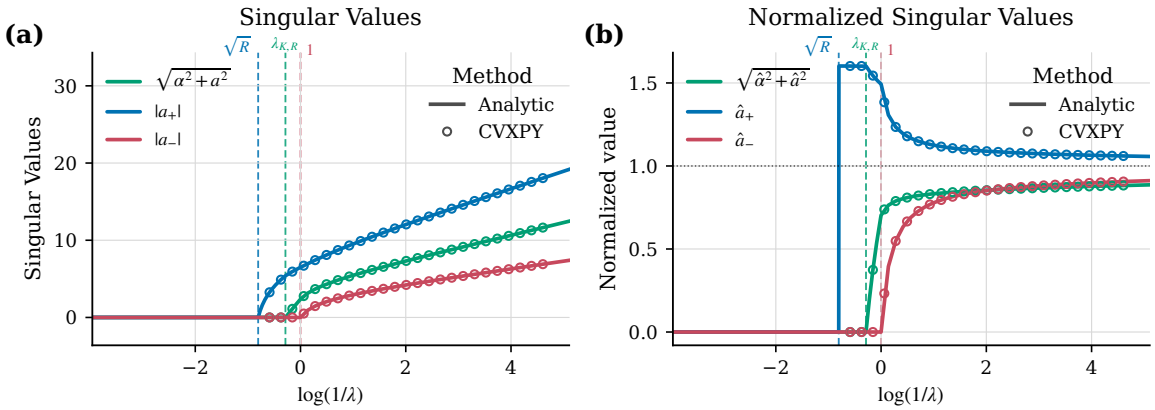


Figure 7: Regularization-path validation for $K = 12$ and $R = 5$. The figure compares the singular-coordinate trajectories predicted by the analytic formulas with those obtained from a direct CVXPY solve of the convex nuclear-norm problem (Z-UFM$_\lambda$). The numerical agreement supports the claim that the regularization-path description is not an artifact of the Sylvester–Hadamard coordinate construction.

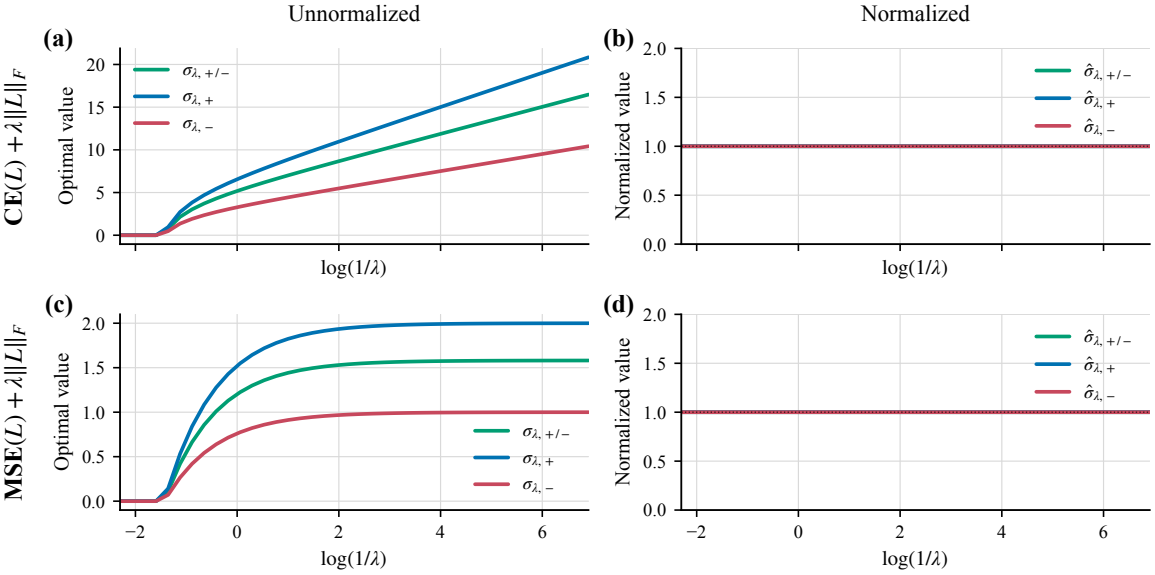


Figure 8: Regularization paths with a Frobenius-norm penalty in logit space. The top row shows the CE objective with Frobenius regularization, and the bottom row shows the corresponding MSE objective. Left panels plot the unnormalized singular coordinates, while right panels plot the same coordinates after normalization by their limiting profile. In contrast to the nuclear-norm path, the normalized coordinates remain aligned throughout the trajectory, indicating that the modes do not emerge sequentially.

at one along the path. Thus, the staged activation observed for the nuclear-norm regularization path disappears.

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.

I.2. Gradient-flow experiments

I.2.1. SETUP

We numerically simulate the gradient flow in (19) 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 6, 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 6. We simulate both the original gradient descent dynamics and the reduced ODE from the same initial coordinates.

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 as a directional discrepancy measure; unlike in the balanced setting, it is not a Lyapunov function for the dynamics.

I.2.2. RESULTS

The results are summarized in Fig. 9. They show three main trends.

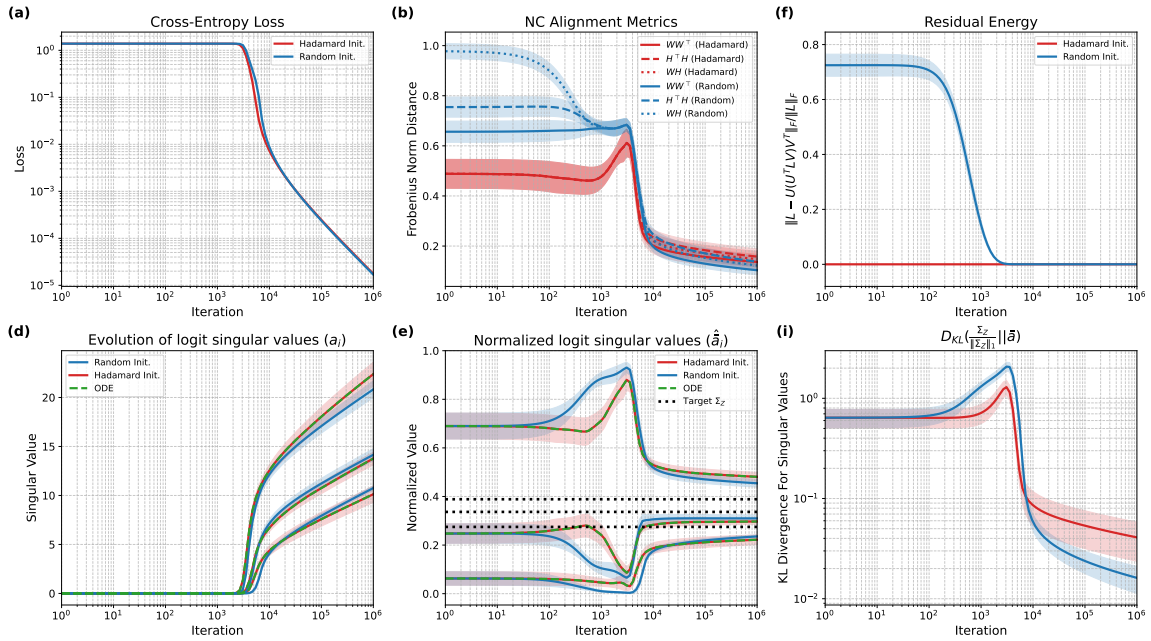


Figure 9: Gradient-flow experiments for $K = 4$ and $R = 2$. We compare Hadamard initialization, random spectral initialization with the same initial singular values, and the reduced vector ODE initialized from the Hadamard coordinates. The panels show the cross-entropy loss, neural-collapse-style alignment 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.

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 6 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

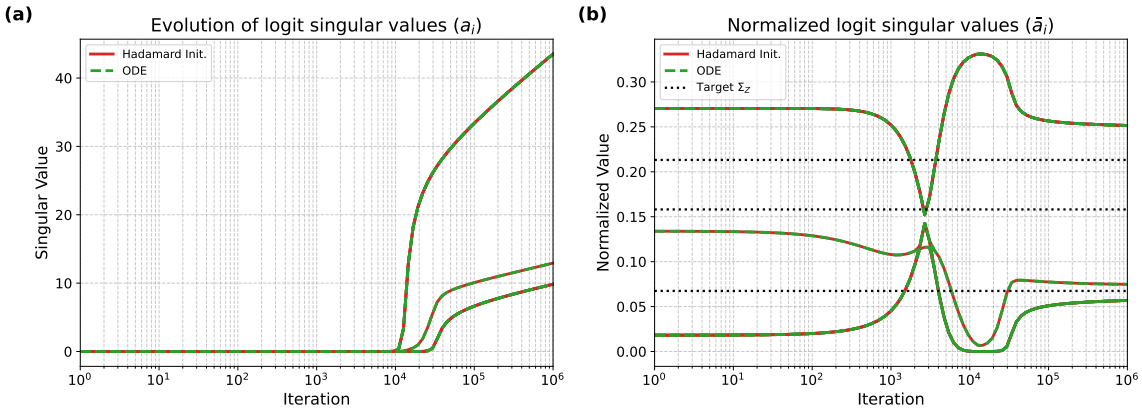


Figure 10: Preservation of block-constant spectral coordinates under gradient flow. The model uses $K = 8$ and $R = 10$, and is initialized with $\boldsymbol{a}_{+,0} = a_{+,0}\mathbf{1}_{M-1}$ and $\boldsymbol{a}_{-,0} = a_{-,0}\mathbf{1}_{M-1}$. The trajectories show that the coordinates within each 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.

exhibits a transient increase before decreasing, which is consistent with the fact that it is not a Lyapunov function in this imbalanced setting.

I.2.3. PRESERVATION OF BLOCK-CONSTANT SPECTRAL COORDINATES

The reduced dynamics in Theorem 6 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 Theorem 10.

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 Fig. 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.