Diagonalizing the Softmax: Hadamard Initialization for Tractable Cross-Entropy Dynamics

Connall Garrod
Mathematical Institute
University of Oxford
connall.garrod@maths.ox.ac.uk

Jonathan P. Keating
Mathematical Institute
University of Oxford

Christos Thrampoulidis
Department of Electrical and Computer Engineering
University of British Columbia
cthrampo@ece.ubc.ca

Abstract

Cross-entropy (CE) loss is central to deep learning, but existing theory often relies on simplifications—such as squared loss or convex models—that miss key aspects of CE optimization. In this work, we study *multi-class* CE dynamics using a two-layer linear network with orthogonal inputs, the simplest non-convex setting for which the CE implicit bias remains unresolved. This model coincides with the unconstrained features model used to study neural collapse (NC). Our analysis is based on a key observation: Hadamard initialization diagonalizes the softmax operator. This allows us to extend the spectral initialization framework that Saxe et al. [61, 62] developed for squared loss. We prove convergence to NC under spectral CE training and give the first finite-time analysis in this setting via an explicit Lyapunov function that decreases monotonically along the trajectory and guarantees convergence to NC despite spurious critical points. We further identify CE-specific phenomena absent under squared loss—coupling, non-monotonic convergence, qualitative dependence on the number of classes, and exponential slowdowns. We also characterize the role of network width and show empirically that spectral dynamics qualitatively model small random initialization.

1 Introduction

The training dynamics of linear networks under mean-squared-error (MSE) loss have led to a range of influential contributions to the theory of deep learning [2, 3, 26, 49, 73]. Among the most elegant and far-reaching of these is the spectral-dynamics framework initiated by Saxe et al. [61, 62]. In these analyses, gradient flow is reduced to a tractable evolution of singular values, yielding closed-form training trajectories. This perspective has shaped our understanding of emergence [62], initialization [11, 43], optimization [23], and lazy-to-rich transitions [15]. A central strength of these works is that they model the full optimization trajectory, rather than only asymptotic characterizations.

No comparable trajectory-level theory is known for cross-entropy (CE), the standard loss for classification. The main obstruction is the softmax, which nonlinearly couples the logit singular modes, preventing the spectral reductions that make MSE dynamics tractable. As a result, even simple non-convex CE models have remained out of reach of explicit trajectory-level analysis. *Can the insights and techniques developed for linear MSE networks be extended to CE with softmax activation?*

This gap is important because CE, not MSE, is the objective used to train modern classifiers. Yet for several phenomena specific to classification networks, we still lack a trajectory-level understanding of

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

how their structure emerges during optimization. The most striking example is neural collapse (NC) [55], a remarkable geometric phenomenon observed near the end of training: class-mean embeddings converge to a simplex equiangular tight frame (ETF), while the final-layer classifier aligns with the same frame. This geometry appears across a wide range of architectures and datasets, despite the absence of any explicit mechanism enforcing it. The central puzzle is therefore: *what implicit mechanism in gradient-based optimization selects this precise geometric configuration from among the many solutions that achieve vanishing training loss?*

This question lies squarely within the framework of *implicit bias*, which seeks to characterize the solutions selected by gradient-based optimization when multiple optima are available. Existing implicit-bias results, however, fall short of fully answering this. For logistic regression, Soudry et al. [63] showed that gradient descent converges in direction to the max-margin solution. This theory has since been extended to deep homogeneous networks, where gradient descent converges to KKT points of a max-margin problem [50, 38, 12]. In contrast to the convex case, however, these results are not fully explicit: they do *not* identify which KKT point is selected [74]. Consequently, even when applied by Ji et al. [35] to the unconstrained features model (UFM), a theoretical model for studying NC, they cannot squarely explain why NC emerges in the absence of explicit regularization.

More broadly, implicit-bias theory for CE-trained models remains largely confined to three regimes: logistic regression (lacking depth), binary classification (omitting multi-class interactions), or KKT characterizations (unable to distinguish between KKT points). This limitation is especially consequential for CE loss, where the infimum is approached only asymptotically. Before reaching its limiting direction, a trajectory may exhibit rich transient behavior, and convergence to the asymptotic regime may be so slow that purely limiting descriptions miss the dynamics that occur at practically relevant timescales. A complete account of implicit bias in this setting therefore requires understanding not only the endpoint selected by optimization, but the entire trajectory by which that selection occurs.

1.1 Contributions

We give the first *trajectory-level* analysis of multiclass CE training dynamics for the CE UFM (see Eq. (2)), a model popular in the NC literature that also represents the simplest non-convex model for which the implicit bias remained unknown: a two-layer linear network with orthogonal inputs.

Softmax diagonalization. Our main technical innovation, Theorem 1, shows that Hadamard initialization diagonalizes the softmax nonlinearity. Concretely, for a *Sylvester-Hadamard matrix* $\Phi \in \mathbb{R}^{K\times K}$(Def. 1), any vector $a \in \mathbb{R}^K$, and the column-wise softmax map $P(\cdot)$:

$$P(\Phi \operatorname{diag}(a)\, \Phi^T) = \Phi \operatorname{diag}(b)\, \Phi^T\,, \qquad \text{with} \quad b := (1/K)\, \Phi^\top P(\Phi a)\,. \tag{1}$$

This enables us to extend the framework of Saxe et al. [61, 62] from MSE to CE, reducing gradient flow to a vector ODE over the singular values a of the logit matrix. This reduced ODE retains the complex couplings of the softmax, yet it is sufficiently structured to allow for a study of its trajectory and implicit bias, exposing several phenomena specific to non-convex multi-class CE training.

Non-monotonicity. Unlike MSE, where singular values converge monotonically, CE can exhibit non-monotonic convergence whose behavior depends sensitively on the number of classes K. For $K = 2, 4$, natural distance-to-convergence metrics decrease monotonically, whereas for $K \geq 8$ we construct explicit initializations where they initially *increase*. Even more strikingly, we construct metrics that are monotonic for $K = 8$ but fail for $K \geq 16$. These phenomena are invisible to KKT or binary analyses and reveal a subtle unrecognized K-dependence in multi-class CE dynamics.

Guaranteed convergence to NC. Despite the non-monotonicity of natural metrics and the presence of competing saddle points, we prove that gradient flow *always* converges to NC under Hadamard initialization. Unlike typical asymptotic implicit-bias proofs, our argument is constructive: The KL divergence is an explicit Lyapunov function that decreases monotonically along the full trajectory.

Role of initialization. To reconcile our theory with the monotonic behavior often observed in practice, we show that standard small random initialization naturally biases the dynamics *of a wide UFM* toward the simplex ETF subspace, steering trajectories away from problematic saddle regions.

The Hadamard trajectories we analyze are genuine gradient-flow paths, so the phenomena above can all arise in multi-class non-convex CE training. Empirically, these trajectories also model small random initialization: both show a two-stage pattern in which singular vectors first align with the target subspace, after which singular values evolve qualitatively similarly (Sec. 7, App. H).

Overall, our analysis shows that even when constrained to spectral initialization, the resulting CE dynamics differ substantially from their MSE counterparts. App. C further identifies the restrictive conditions (linearized regime or removing depth) needed to recover MSE-like dynamics under CE.

Finally, while we focus on the unregularized balanced UFM, we show in App. F that our Hadamard reduction and the KL divergence Lyapunov analysis extend naturally to L_2-regularized training. Extensions to deeper networks and imbalanced classes are left for future work. The diagonalization argument of Theorem 1 is also not specific to softmax: retaining the Sylvester–Hadamard construction, it applies to any entrywise nonlinearity, since such maps are permutation-equivariant. We therefore hope our framework for path-level analysis in CE settings will inspire similar analyses in more general nonlinear networks.

2 Related work

Here we review the most closely related work; a comprehensive discussion appears in App. B.

Neural collapse. Many works study the emergence of NC through the UFM [53, 18] across a range of loss functions [87, 53, 35, 27, 86]. For CE, existing analyses rely on explicit L_2 regularization [88, 21] or establish convergence only to a KKT point [35]. Instead, we provide a dynamical analysis in this setting and prove convergence to NC under a representative spectral initialization scheme.

Implicit bias. It has been shown that separable logistic regression must converge in direction to the max-margin solution [63]. Subsequent work has extended this to homogeneous networks [50, 37, 38], alternative optimizers [24, 7, 68, 57, 82, 17], and other architectures [25, 44, 5, 70, 75, 40]. Much of this work focuses on binary classification. We show that multiclass interactions, central to NC, give rise to phenomena not captured in binary settings.

Spectral dynamics. Saxe et al. [61] showed that deep linear networks trained with MSE loss can admit tractable explicit dynamics over singular modes. This perspective has since been extended to emergence [62], initialization [11, 43], optimization [23], training regimes [15], and linear attention [51]. These analyses rely on the mode-wise decoupling or matrix Riccati reductions enabled by MSE. In contrast, softmax CE couples logits nonlinearly, preventing a direct application of prior spectral-dynamics frameworks. Our work develops such a framework for this central setting.

3 Background

Consider K-class classification with n samples per class. Let $h_{ic} \in \mathbb{R}^d$ denote the embedding, parameterized by a deep network, of sample $i \in [n]$ from class $c \in [K]$, and take $d \geq K$. Stack these by class to form the embedding matrix $H = [h_{1,1}, \dots, h_{n,1}, h_{1,2}, \dots, h_{n,K}] \in \mathbb{R}^{d \times Kn}$, and let $W \in \mathbb{R}^{K \times d}$ denote the last layer classifier head, so that the logits of the training data are $Z = WH$.

Neural collapse (NC) is the empirically observed phenomenon wherein embeddings collapse onto their class means, and these means, together with the classifier W, align with a simplex ETF [55]. That is, $H^\top H \propto S \otimes 1_n 1_n^\top$ and $WW^\top \propto S$, where $S = I_K - \frac{1}{K} 1_K 1_K^\top \in \mathbb{R}^{K \times K}$ is the simplex equiangular tight frame (ETF) matrix, and 1_m is the m-dimensional ones vector. For matrices $A = [a_{ij}] \in \mathbb{R}^{p \times q}$ and $B \in \mathbb{R}^{r \times s}$, their Kronecker product is $A \otimes B = [a_{ij} B]_{i,j} \in \mathbb{R}^{pr \times qs}$.

Unconstrained Features Model. The UFM is a mathematical model used to study NC: it approximates the embedding map of an expressive deep network as mapping training data to arbitrary points in the feature space, treating the embedding h_{ic} of the data as freely optimized vectors. Under CE loss with *no* explicit regularization, the UFM minimizes over $W \in \mathbb{R}^{K \times d}$ and $H \in \mathbb{R}^{d \times Kn}$:

$$\mathcal{L}(W, H) = \sum\nolimits_{c \in [K]} \sum\nolimits_{i \in [n]} \log\left(1 + \sum\nolimits_{c' \neq c} e^{(Wh_{ic})_{c'} - (Wh_{ic})_c}\right). \tag{2}$$

Equivalently, this can be viewed as a two-layer linear network with first-layer weights H, output-layer weights W, and input data the standard basis vectors in $\mathbb{R}^{Kn}$. See App. B for a description of prior UFM analyses, and [22, App. A] for a comprehensive discussion of the UFM's role as a modeling tool.

Let $Y = I_K \otimes 1_n^\top \in \mathbb{R}^{K \times nK}$ denote the one-hot label matrix. Define $P(Z)$ as the *softmax* operation applied column-wise to the logit matrix $Z = WH$: i.e., $\left[P(Z)\right]_{ij} = \exp(Z_{ij}) \,/\, \sum_{k=1}^K \exp(Z_{kj})$. Under gradient flow on the CE loss in Eq. (2), the parameters evolve according to

$$\frac{dW}{dt} = (Y - P(WH))H^T, \qquad \frac{dH}{dt} = W^T(Y - P(WH)). \tag{3}$$

Using the analysis of Lyu & Li [50], Ji et al. [35] showed that this gradient flow converges in direction to a KKT point of the following *non-convex* max-margin problem:

$$(\hat{W}, \hat{H}) \in \arg\min_{W,H} \left\{ \frac{1}{2}\|W\|_F^2 + \frac{1}{2}\|H\|_F^2 \right\}, \quad \text{s.t. } (w_c - w_{c'})^T h_{ic} \geq 1, \forall c' \neq c \in [K], i \in [n]. \tag{4}$$

Here, w_c^T denotes the c-th row of W. They further proved that (i) The global minimizers of this max-margin problem are NC solutions; that is, $\hat{W}$ and $\hat{H}$ satisfy NC. (ii) All other critical points, which correspond to non-NC geometries, are saddles. While useful, this does *not* immediately guarantee convergence to NC: as shown by Vardi et al. [74], gradient flow can, in general, converge to KKT points that are not even local minima of the non-convex max-margin problem. This is because this benign landscape is a property of the asymptotic problem, not of the true loss surface, which has no finite minimizers, and so standard arguments about avoiding saddles [45] do *not* apply here. Thus, despite the canonical nature of this setting, the question is open: *Does gradient flow* (3) *converge in direction to a global minimizer of* (4)*, i.e., to NC solutions?*

Exact dynamics with MSE loss. Replacing CE with MSE loss makes the UFM analytically tractable, thanks to Saxe et al. [61, 62]. Their framework applies generally to whitened inputs, and the UFM is a canonical instance of this setting. Under MSE loss, the gradient flow (3) is instead:

$$\frac{dW}{dt} = (Y - WH)H^T, \qquad \frac{dH}{dt} = W^T(Y - WH). \tag{5}$$

Let the singular value decomposition of Y be $Y = U\Sigma V^T$, with singular values s_c for $c \in [K]$. Restrict initialization to a form of spectral initialization:

$$W = UD_W R^T, \qquad H = RD_H V^T,$$

for some orthogonal matrix $R \in \mathbb{R}^{d \times d}$, where the only nonzero diagonal entries of D_W and D_H are $\alpha_1, ..., \alpha_K$ and $\beta_1, ..., \beta_K$ respectively. Then, at initialization, the label matrix Y and the logit matrix WH are mutually diagonalizable: $Y - WH = U(\Sigma - D_W D_H)V^T$. When put in Eq. (5), and using $U^T U = V^T V = I$, this implies:

$$\frac{dW}{dt} = U\,(\Sigma - D_W D_H)D_H^T\,R^T, \qquad \frac{dH}{dt} = R\,D_W^T(\Sigma - D_W D_H)\,V^T. \tag{6}$$

Thus, only the singular values of W and H evolve, while the singular vectors remain fixed. Specifically, from Eq. (6), the singular values evolve according to

$$\frac{d\alpha_i}{dt} = (s_i - \alpha_i\beta_i)\,\beta_i, \qquad \frac{d\beta_i}{dt} = (s_i - \alpha_i\beta_i)\,\alpha_i\,.$$

These equations decouple into K independent pairs of ODEs, and can be solved to show that the products of the singular values $a_i = \alpha_i\beta_i$ follow sigmoidal trajectories from their initial value to s_i. This implies that the logit singular values a_i converge *monotonically* to the singular values of the target matrix Y. It has also been shown empirically that small random initialization has the same characteristic behavior as these analytic paths [62, 23]. But *can this framework extend to CE loss?*

4 Hadamard initialization

Contrasting the gradient flow equations for CE and MSE loss, given in Eqns. (3) and (5), highlights the main difficulty in extending spectral dynamics arguments to CE. For MSE, it suffices that Y and WH are simultaneously diagonalizable. For CE, however, one must simultaneously diagonalize Y, WH, and $P(WH)$, which demands precise control over how the softmax nonlinearity interacts with the singular value structure of the logit matrix. Nonlinear functions rarely preserve such structure.

4.1 Diagonalizing the softmax

We show that this obstacle can be overcome by initializing with Sylvester Hadamard matrices.

Definition 1. *The **Sylvester Hadamard matrices** $\{\Phi_{2^m} : m \in \mathbb{N}\}$ are defined recursively by*

$$\Phi_1 = 1, \quad \Phi_{2^m} = \begin{bmatrix} \Phi_{2^{m-1}} & \Phi_{2^{m-1}} \\ \Phi_{2^{m-1}} & -\Phi_{2^{m-1}} \end{bmatrix} \in \mathbb{R}^{2^m \times 2^m}.$$

Each Φ_{2^m} is a ± 1-valued matrix, whose first row/column consist entirely of ones. Moreover, they are symmetric $\Phi_{2^m}^T = \Phi_{2^m}$ and orthogonal $\Phi_{2^m}^T \Phi_{2^m} = 2^m I_{2^m}$. Sylvester Hadamard matrices form a subclass of the more general Hadamard matrices, the study of which is left to future work.

We first illustrate our technique in the simplified setting where $n = 1$ (one data point per class), then extend to general n. In this case, the label matrix reduces to $Y = I_K$, whose SVD allows $U = V$ to be any orthogonal matrix. We show that selecting $U = V = \frac{1}{\sqrt{K}}\Phi$, where Φ is a Sylvester Hadamard matrix, achieves mutual diagonalizability despite the softmax activation.

Theorem 1. *For $K = 2^m, m \in \mathbb{N}$, let $U = \frac{1}{\sqrt{K}}\Phi$, where Φ is the $K \times K$ Sylvester Hadamard matrix. Denote the columns of U by u_i, $i \in [K]$. Then, for any $Z = \sum_i a_i u_i u_i^T$, its softmax satisfies*

$$P(Z) = \sum\nolimits_i \nu_i u_i u_i^T, \qquad \text{where } \nu_i = \frac{\sum_j \Phi_{ij} e^{\frac{1}{K}(\Phi a)_j}}{\sum_j e^{\frac{1}{K}(\Phi a)_j}}.$$

Throughout, we refer to this initialization as *Hadamard initialization*. The theorem shows that this spectral structure is preserved: a matrix diagonalizable by a Sylvester–Hadamard matrix remains so after applying softmax. This diagonalization follows from the interplay of two properties: (i) *permutation equivariance of the softmax*, and (ii) *group structure of Hadamard rows*. Together these ensure that applying softmax to Z produces a dyadic matrix: a matrix whose entries are of the form $c_{i \oplus j}$ for some vector c, where $\oplus$ denotes the bitwise XOR operation. Such matrices are known to be diagonalized by Sylvester-Hadamard matrices [31], validating Eq. (1). See App. G.1 for a detailed proof.

4.2 Tractable cross-entropy dynamics

Using Theorem 1, we can adapt the arguments of Saxe et al. [61], detailed in Sec. 3. Specifically, initializing with $W = UD_W R^T$, $H = RD_H U^T$, for any orthogonal matrix $R \in \mathbb{R}^{d \times d}$, with $U = \frac{1}{\sqrt{K}}\Phi$, implies Y, WH, and $P(WH)$ are mutually diagonalizable. This ensures that the singular vectors of W and H are fixed, and all time evolution occurs in their singular values. This is specific to our construction: a generic orthogonal U does not yield mutual diagonalizability.

To extend this construction from one to n samples per class, we exploit the tensor-product structure $Y = I_K \otimes 1_n^T$. Let Q be an orthogonal matrix whose first column is $\frac{1}{\sqrt{n}}1_n$, meaning it is a right singular-vector matrix of 1_n^T. Then $V = U \otimes Q$ combines the Hadamard modes across classes with a basis that separates the class constant direction from the within-class directions. The theorem below summarizes the result for general n. The proof is in App. G.2, with validation in Fig. 1.

Theorem 2. *For $K = 2^m, m \in \mathbb{N}$, let $U = \frac{1}{\sqrt{K}}\Phi$, where Φ is the $K \times K$ Sylvester Hadamard matrix. For the UFM in Eq. (2), with balanced classes of size n, initialize the parameters as $W = UD_W R^T$, and $H = RD_H V^T$, where $R \in \mathbb{R}^{d \times d}$ is orthogonal, $V = U \otimes Q$, where Q is a right singular-vector matrix of 1_n^T, and the only nonzero singular values of D_W and D_H are $\alpha_1, \ldots, \alpha_K$ and $\beta_1, \ldots, \beta_K$, respectively. Then the gradient flow updates in Eq. (3) reduce to*

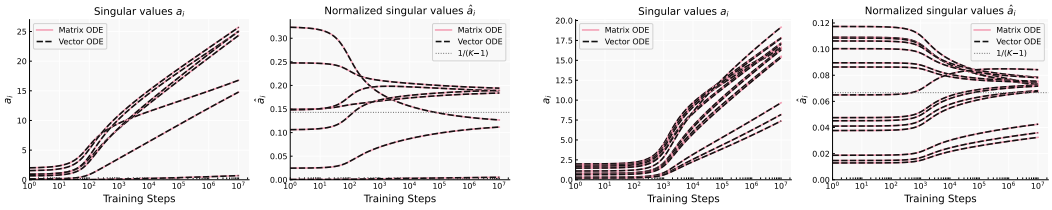
$$\frac{d\alpha_i}{dt} = \frac{\sqrt{n}}{D} b_i \beta_i, \quad \frac{d\beta_i}{dt} = \frac{\sqrt{n}}{D} b_i \alpha_i,$$

where, for $a_i = \alpha_i \beta_i$:

$$b_i = \sum\nolimits_{j \in [K]} [1_K 1_K^T - \Phi]_{ij}\, e^{-\frac{1}{K\sqrt{n}}([1_K 1_K^T - \Phi]a)_j}, \quad D = \sum\nolimits_{j \in [K]} e^{-\frac{1}{K\sqrt{n}}([1_K 1_K^T - \Phi]a)_j}. \tag{7}$$

We index singular values by their associated Hadamard singular vectors rather than by decreasing magnitude, so their indices remain fixed throughout training even if singular values cross.

Recall that the first row and column of Φ consist of all ones. Since $[1_K 1_K^T - \Phi]_{1j} = 0$ for all j, it follows that $b_1 = 0$, and therefore α_1, β_1 remain constant. Moreover, since $[1_K 1_K^T - \Phi]_{j1} = 0$ for all j, these trivial singular values do not appear in the evolution equations for the others. We may thus, without loss of generality, set $\alpha_1 = \beta_1 = 0$ and henceforth take $\alpha, \beta \in \mathbb{R}^{K-1}$ to be the vectors of non-trivial singular values, with indices shifted accordingly. Consequently, define $\Psi \in \mathbb{R}^{K-1 \times K-1}$ by deleting the first column and row of $1_K 1_K^T - \Phi$, so that $\Psi = 1_{K-1} 1_{K-1}^T - X$, where $X = \Phi[2:K, 2:K]$ is the core of the Sylvester Hadamard matrix.



(a) $K = 8$ (b) $K = 16$

Figure 1: Numerical illustration of the Hadamard reduction for $K = 8, 16$. The trajectories of the full matrix dynamics (Eq. (3)) and the reduced vector ODE (Eq. (9)) are indistinguishable, illustrating Thms. 1 and 2. The growth of the raw singular values motivates studying the normalized ones.

These dynamics correspond to gradient flow (w.r.t parameters α_i, β_i) on the energy

$$E(\alpha, \beta) = Kn \log \Big(1 + \sum\nolimits_{i=1}^{K-1} \exp\big(-\frac{1}{K\sqrt{n}}(\Psi a)_i\big)\Big), \tag{8}$$

where $a_i = \alpha_i \beta_i$. A consequence of this variational structure is the symmetry $\alpha_i \to \gamma\alpha_i$, $\beta_i \to \gamma^{-1}\beta_i$, from which Noether's theorem yields the conserved quantities $\alpha_i^2 - \beta_i^2 = C_i$. For simplicity, and following common practice [13, 61, 62, 23], we adopt the balancedness condition that sets $C_i = 0$, thus $\alpha_i = \beta_i$ throughout. Under this mild assumption, the dynamics can be reduced to the evolution of the *logit singular values* $a_i = \alpha_i \beta_i$; see App. D for details. Additionally, absorbing $2\sqrt{n}$ into the time variable and $1/(K\sqrt{n})$ into a yields the following reduced dynamics for $a \in \mathbb{R}^{K-1}$:

$$\frac{da_i}{dt} = \frac{1}{D} b_i a_i\,, \quad \text{with } b_i = \sum_{j=1}^{K-1} \Psi_{ij} e^{-(\Psi a)_j} \text{ and } D = 1 + \sum_{j=1}^{K-1} e^{-(\Psi a)_j}. \tag{9}$$

We now list basic properties of the dynamics that we use in our subsequent analysis in Sec. 5.

Coupled dynamics. Unlike the MSE case, the terms b_i/D depend on the complete set of singular values $a_1, \ldots, a_{K-1}$. The equations therefore do not decouple, and each singular value influences the trajectory of the others. Because of this, there are not enough conserved quantities to guarantee closed-form dynamics. Nevertheless, we show below that they are sufficiently tractable.

Positive singular values. The singular values $a_i(t)$ are monotonically increasing in time. Therefore, if $a_i(0) > 0$, then $a_i(t) > 0$ for all $t \geq 0$. Moreover, if $a_i(0) = 0$, then it remains zero throughout training. We henceforth assume all singular values are initialized to positive values.

Mirror-flow interpretation. Since $a_i(t)$ remains positive for all future times, we can write the updates (absorbing the K, n-dependent factors) as $\mathrm{d}(\log(a_i))/\mathrm{d}t = -\nabla_{a_i} E(a)$, where $E = E(a)$ is the energy functional in Eq. (8). In this form, the dynamics are viewed as mirror-flow updates with respect to the entropy potential for the energy functional in Eq. (8); see App. D for details.

Normalized singular values. From Eq. (9), it is straightforward to show that the norm of the logit singular values (e.g., the L_1-norm $\|a\|_1 = \sum_i a_i$) diverges to infinity. This is not specific to our reduction: unregularized CE loss on linearly separable data generally drives logits to unbounded norm as the loss approaches zero [38, 50, 63]. This divergence implies that there are no finite minimizers of the UFM objective on the optimization surface. We therefore study the dynamics of the *normalized logit singular values* $\hat{a}_i = a_i/\|a\|_1$, which evolve as:

$$\frac{d}{dt}(\hat{a}_i) = \frac{\hat{a}_i}{D}\left[b_i - \bar{b}\right], \quad \text{where } \bar{b} = \sum\nolimits_{j \in [K-1]} \hat{a}_j b_j. \tag{10}$$

Note that the normalized logit singular value vector $\hat{a}$ lies in the interior of the $\mathbb{R}^{K-1}$ probability simplex, since with positive initialization all entries remain positive throughout the dynamics.

5 Cross-entropy dynamics and implicit bias

5.1 NC as a candidate among stable directions

Note from Eq. (8) that any direction $\hat{a}$ satisfying $(\Psi\hat{a})_i > 0$ for all $i \in [K-1]$ can attain arbitrarily small loss by taking $\|a\|_1 \to \infty$. The key question is therefore which of these zero-loss directions gradient flow in Eq. (9) selects. To gain insight into the structure of potential convergence directions, we inspect the *stable directions*, i.e., those satisfying $\frac{\mathrm{d}\hat{a}}{\mathrm{d}t} = 0$. If a stable direction satisfies $\hat{a}_i > 0$ for

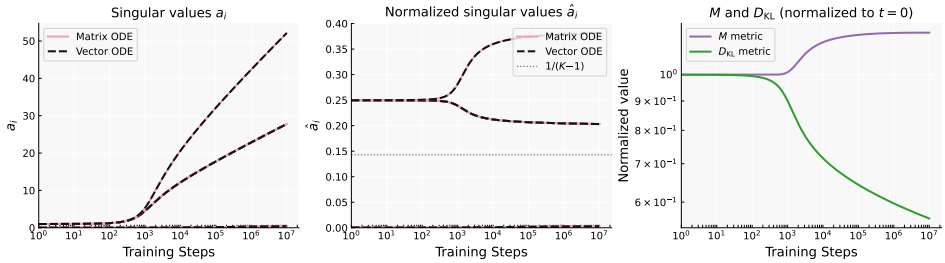


Figure 2: Training with $K = 8$ and initialization $a(0) = [1, 1, 1, 1, 0.001, 0.001, 0.001]$. **Left/Middle:** Evolution of raw/normalized singular values. **Right:** Evolution of the metric M in Eq. (11) and the KL Lyapunov function of Thm. 3, both normalized by their initial values.

all i (meaning all $K-1$ logit singular values are positive), then the corresponding logit matrix is full-rank; otherwise, the stable direction corresponds to a low-rank matrix and such sparse stable directions exist in our problem. Among full-rank stable directions, we show in App. D.3 that the unique solution is $\hat{a} = \frac{1}{K-1}1_{K-1}$, which corresponds to NC. To see why this corresponds to the NC solution, note that equal logit singular values (i.e., $a_i = a_j$ for all $i, j \in [K-1]$) imply that the singular values of W and H are equal as well: $\alpha_i = \beta_i = \alpha_j = \beta_j$ for all $i, j \in [K-1]$. Recall that, without loss of generality, the singular value corresponding to the all-ones eigenspace is zero for both matrices. Together, these conditions yield $W \propto \Phi\,\mathrm{diag}(0, 1, \ldots, 1)R^T$ and $H \propto R\,\mathrm{diag}(0, 1, \ldots, 1)\Phi^T$, which in turn give $WW^T \propto S$ and $H^TH \propto S$, where $S = \frac{1}{K}\Phi\,\mathrm{diag}(0, 1, \ldots, 1)\Phi^T$ is the ETF matrix.

5.2 Convergence to neural collapse via a Lyapunov function

Does gradient flow exhibit an implicit preference for the full-rank NC solution $\hat{a} = \frac{1}{K-1}1_{K-1}$ among all other stable points? We show that this is indeed the case. Moreover, we establish a stronger result than asymptotic convergence: we identify a Lyapunov function, measuring distance to NC, that decreases monotonically along the gradient-flow trajectory. The proof is in App. G.3.

Theorem 3. *Let $K = 2^m, m \in \mathbb{N}$. Let $a(t) \in \mathbb{R}^{K-1}$ be a variable initialized with $a_i(0) > 0$ for all i that evolves under the time evolution of Eq.* (9)*. Then the metric $D_{\mathrm{KL}}(\frac{1}{K-1}1_{K-1} \,\|\, \hat{a})$, where D_{KL} is the Kullback-Leibler divergence, serves as a Lyapunov function of the flow. Specifically, the metric is decreasing throughout training and ensures that $\hat{a} \to 1/(K-1) \cdot 1_{K-1}$ as $t \to \infty$.*

This result is notable in two respects. First, regarding *implicit bias*, existing results typically establish only asymptotic convergence rather than monotonic behavior throughout the full trajectory via a Lyapunov function. Even when restricted to asymptotic convergence, the only prior result applicable to the CE UFM [37] is limited to binary classification ($K = 2$). However, the binary case is trivial in our framework: the unique stable direction (of rank $K - 1 = 1$) is NC itself. In contrast, for $K \geq 4$, multiple stable directions exist, making the implicit bias question nontrivial. Indeed, we show below that, even the cases $K = 4$ vs. $K = 8$ vs. $K > 8$ exhibit qualitatively different optimization properties: several candidate Lyapunov functions work for smaller K but admit counterexamples for larger K. Second, regarding *neural collapse*, all previous claims about convergence to NC either used explicit finite regularization [88], analyzed the regularization path [71], or characterized the KKT points of the corresponding non-convex max-margin problem in Eq. (4) [35]. Theorem 3 is the first result proving that NC occurs purely as a result of gradient flow implicit bias under CE loss.

5.3 Non-monotonic behavior

The monotonic convergence via the KL divergence as a Lyapunov function is a special property. We now show that other natural distance metrics to NC do not exhibit monotonic behavior.

We begin with the squared L2-distance metric, which is commonly used in empirical work on NC:

$$M = \frac{1}{2}\sum\nolimits_{i\in[K-1]} \left(\hat{a}_i - \frac{1}{K-1}\right)^2. \tag{11}$$

For $K = 2, 4$, this metric decreases monotonically; however, monotonicity breaks down for $K \geq 8$.

Theorem 4. *For $K = 2^m, m \in \mathbb{N}$, let $a(t) \in \mathbb{R}^{K-1}$ be a variable initialized with $a(0) > 0$ that evolves as in Eq.* (9)*. If $K = 2$ or $K = 4$, then the metric M in Eq.* (11) *decreases monotonically to zero. If $K \geq 8$, then there exist initializations $a(0)$ such that the metric M is not monotonic.*

The proof appears in App. G.4. This result reveals a fundamental qualitative difference from MSE [61, 62]: here, coupled singular values can induce non-monotonic behavior in natural distance

metrics. Fig. 2 illustrates the severity of this phenomenon, showing the evolution of $\hat{a}_i$ and M for an initialization leading to non-monotonic behavior: M increases monotonically over the displayed trajectory, even though it must eventually converge to zero. At first sight, Fig. 2 may appear inconsistent with Theorem 3, since several normalized singular values remain close to zero over the displayed trajectory. This is a finite-time effect: from Eq. (10), $\frac{d\hat{a}_i}{dt}$ is proportional to $\hat{a}_i$, so singular values that are initially very small can remain visually close to zero for long periods even while increasing. The KL divergence is sensitive to these small modes through its $\log(\hat{a}_i)$ terms, and therefore decreases even while the largest singular values temporarily move away from uniformity. Further discussion appears in App. E.

The role of non-uniform stable points ($K \geq 8$). This behavior occurs because the dynamics first approach a low-rank stable point with non-uniform support—that is, a point where $\hat{a}_i = 0$ for some $i \in [K-1]$ and among the non-zero entries, $\hat{a}_i \neq \hat{a}_j$ for some i, j. By moving toward a non-uniform configuration, metrics such as M measuring uniformity increase. The existence of such non-uniform stable points arises from the structure of the matrix Ψ. Our energy functional (Eq. (8)) depends on the quantities $(\Psi a)_i$, which play the role of margins. Each singular mode $\hat{a}_j$ contributes to enlarging $K/2$ of these margins—precisely those indices $i \in [K-1]$ for which $\Psi_{ij} \neq 0$. As $\|a\|_1$ becomes large, the energy is determined, up to exponentially small terms, by the smallest normalized margins $\arg\min_i(\Psi\hat{a})_i$. While NC maximizes this quantity in the full-rank setting, in rank-constrained regimes we do not necessarily have an equal number of singular values contributing to each margin, making non-uniform solutions preferable. Further explanation appears in App. E.

For $K = 4$, all stable points are uniform on their support, ensuring standard metrics monotonically decrease. For $K \geq 8$, when non-monotonicity occurs, the trajectory remains near low-rank stable points for extended periods before transitioning to higher-rank saddles, eventually reaching NC. Similar behavior has been observed empirically [33], albeit in settings with finite-norm critical points.

Nuanced optimization behavior with respect to K. Beyond this specific example, the K-dependence of convergence behavior is more nuanced. The dynamics clearly depend on K through the maximum number of rank transitions along trajectories. Theorem 4 captures how the nature of convergence changes qualitatively with K: from monotonic ($K = 4$) to non-monotonic ($K \geq 8$). However, this K-dependence manifests differently across metrics. Some metrics exhibit monotonic convergence for $K = 8$ but break down for $K \geq 16$; see App. D.4 for examples. These findings suggest that many other non-trivial behaviors likely emerge as the class number increases.

Exponential slowdown. The potentially large timescale of convergence is due to lazy training dynamics that emerge in the large norm regime. To see this, consider the quantities $b_i = \sum_{j=1}^{K-1} \Psi_{ij} e^{-\|a\|_1(\Psi\hat{a})_j}$ in terms of the variables $\|a\|_1, \hat{a}$. Clearly, as $\|a\|_1$ becomes large, these terms decay exponentially; consequently, the derivatives of a_i become small and network evolution becomes slow. Consequently, convergence to uniform normalized singular values may occur on timescales far beyond those visible in the numerical trajectories. In settings where the loss does not take its minimal value at infinity, such as in the MSE case, these large-norm exponential slowdown effects do not occur. Further comparison to MSE dynamics is provided in App. C.

6 Standard initialization schemes lead to neural collapse

The evidence presented in Sec. 5.3 raises a natural question: why are such peculiar behaviors not observed in typical training of UFMs [35] or in standard neural networks? To provide an explanation, we first demonstrate that even though we do not generally see monotonic convergence to NC, the standard metric M of Eq. (11) must decrease when it has hit a small enough value.

Theorem 5. *For $K = 2^m, m \in \mathbb{N}$, let $a(t) \in \mathbb{R}^{K-1}$ be a variable initialized with $a_i(0) > 0$ for all i that evolves under the time evolution of Eq.* (9)*. Then the metric M, defined in Eq.* (11)*, satisfies the following implication: $M \leq \min\left(\frac{1}{8K^2}, \frac{1}{18K^3\|a\|_1^2}\right) \implies \frac{dM}{dt} \leq 0$.*

The proof appears in App. G.5. This demonstrates that M will decrease for a finite time training trajectory, provided the initialization is sufficiently close to the NC solution.

We can now explore how typical initialization schemes, featuring random i.i.d. entries in the parameter matrices, impact the singular values of the logit matrix. We show below that, with small-scale Gaussian initialization, the linearized dynamics bias a sufficiently wide system towards the simplex ETF.

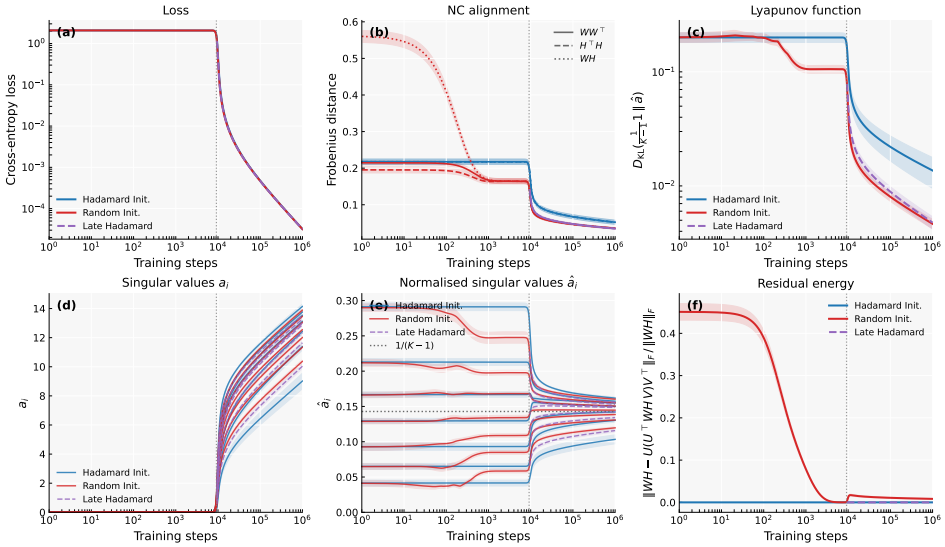


Figure 3: Comparison between gradient flow under small random initialization, Hadamard initialization with the same singular values, and a late-Hadamard trajectory initialized at the end of the alignment phase (vertical dotted line). Parameters $K = 8, d = 40$. Average metrics are reported along with standard error bars over ten runs. For random initialization, an initial alignment phase is followed by normalized singular-value dynamics that qualitatively resemble Hadamard dynamics.

Theorem 6. *Evolve the parameters of the UFM defined in Eq.* (2) *under the gradient flow of Eq.* (3)*, with initialization* $W(0) = \epsilon d^{-\frac{1}{2}} A_1$*,* $H(0) = \epsilon d^{-\frac{1}{2}} A_0$*, where the matrices* A_1, A_0 *have entries drawn i.i.d. standard Gaussian, and* d *is the network width. Defining* $t^* = \frac{1}{2\sqrt{2Kn}} \log(\frac{1}{\epsilon})$*, and taking the limit* $\epsilon \to 0$*, the leading order term of the logit matrix at time* t^* *is given by* $Z(t^*) \sim \frac{1}{2d\sqrt{n}} \epsilon^2 e^{2t^*\sqrt{n}} \left[S X X^T (S \otimes 1_n^T) \right]$*, where* $X \in \mathbb{R}^{K \times d}$ *has entries i.i.d. standard Gaussians. Hence additionally taking the* $d \to \infty$ *limit, we recover asymptotically almost surely*

$$Z(t^*) \sim \frac{1}{2\sqrt{n}} \epsilon^2 e^{2t^*\sqrt{n}} \left(S \otimes 1_n^T \right).$$

The theorem shows that for small random initialization, early-stage dynamics naturally drive the logit matrix toward the NC structure, with the quality of the approximation improving as initialization becomes smaller and the network width increases. Since the NC solution is also attractive at all scales, this means that standard initialization schemes naturally lead the trajectory to avoid regions of non-monotonicity that occur around the saddles in the surface.

7 Numerical experiments

Hadamard initialization confines gradient flow to an invariant spectral subspace. We now test whether these analytic trajectories capture the qualitative behavior of standard small random initialization. Fig. 3 reports, for both random and Hadamard initialization, the metrics studied in the preceding sections, together with the residual energy defined in App. H, which measures the extent to which the singular vectors of the logit matrix are aligned with the Hadamard basis.

Random trajectories exhibit a clear two-stage structure. First, during an initial alignment phase, the loss and singular values change little while the off-subspace residual energy rapidly decreases, indicating alignment with the Hadamard spectral subspace. This behavior is consistent with the MSE case [61, 62] and the "silent alignment effect" observed in prior work [6, 11]. This interpretation is also consistent with our theoretical results: Theorem 6 shows that small random initialization biases the early dynamics toward the simplex-ETF subspace, while App. C.1 shows that the small-logit regime exhibits MSE-like linearized dynamics. After this phase, the random and Hadamard trajectories become qualitatively similar: both show rapid loss decay and growth of the logit singular values, while the normalized singular values exhibit comparable qualitative trends as they move toward the NC equilibrium. The KL divergence is monotone for Hadamard initialization, as guaranteed by Theorem 3, and becomes monotone for random initialization after the alignment phase is completed.

To isolate the post-alignment dynamics, in Fig. 3 we further compare the random trajectory with a late-Hadamard trajectory initialized at the end of the alignment phase, using singular values matched to those of the random trajectory at that time. This late-Hadamard trajectory captures the qualitative evolution of both the loss and the normalized singular values after alignment. The match is not exact, since random trajectories retain small off-subspace components. In addition, small differences among growing singular values can translate into larger changes in the normalized singular values. This sensitivity is a feature of CE training that makes the dynamics richer than in the MSE setting [61, 62, 23]. Overall, these experiments suggest that Hadamard initialization provides a tractable representative trajectory for the qualitative behavior of training from small random initialization after the alignment phase. Additional details and experiments appear in App. H.

8 Conclusion

CE training dominates modern machine learning yet remains far less understood than MSE, even in simple settings. Under a canonical CE setup, we show that parameters grow unboundedly, exponential slowdowns near saddles persist for extended periods, natural convergence metrics behave non-monotonically, and the qualitative nature of convergence depends nontrivially on both initialization and the number of classes. These features contrast sharply with the decoupled, monotone spectral dynamics obtained under MSE. We uncover these features by adapting the spectral-dynamics framework of Saxe et al. to the softmax setting. We view demonstrating that such an extension is feasible as our central contribution, and expect it to serve as a foundation for future CE analyses, similar to what has occurred for MSE.

Future work. Natural next steps include removing the $K = 2^m$ restriction and extending the analysis to alternative optimizers and deeper networks. More broadly, extending the analysis to more general initializations, likely by combining our techniques with tools developed in the MSE setting [11, 15, 43], remains an important open direction. Finally, the diagonalization mechanism of Theorem 1 is not specific to softmax: with the Sylvester–Hadamard construction fixed, the same argument applies to any entrywise nonlinearity through permutation equivariance. Whether this property can yield tractable spectral dynamics for other objectives or architectures is an interesting direction for future work.

Limitations. Our work uses spectral initialization to explore finite-time CE dynamics. These spectral paths are exact gradient-flow trajectories, meaning all the phenomena we identify are genuine features of the CE loss surface. Moreover, we provide empirical evidence that spectral trajectories demonstrate behavior shared with small random initialization, matching the evidence available in the MSE literature. Extending beyond this regime represents valuable future work. Our Hadamard-based analysis also requires $K = 2^m$. We view this as a first step beyond binary classification, which remains the focus of most implicit-bias works. Importantly, our experiments indicate that small random initializations exhibit similar behavior and can be used for arbitrary K. We expect the restriction $K = 2^m$ can be relaxed using Fourier-type constructions due to the nature of circulant matrices, though this introduces additional technical challenges that we leave to future work. Lastly, our analysis assumes balancedness, as described in Section 4.2 and commonly assumed in prior work [13, 61, 62, 23]. Extending the analysis beyond balanced initialization is an important direction for future work.

Acknowledgments and Disclosures

We thank the anonymous reviewers for their thoughtful feedback and constructive suggestions, which helped improve the clarity and presentation of the paper. CG is supported by the Charles Coulson Scholarship. The authors also acknowledge support from His Majesty's Government in the development of this research. CT acknowledges generous support by NSERC through grants NSERC Discovery Grant No. 2021-03677 and Alliance Grant ALLRP 581098- 22. CT also thanks Yize Zhao for discussions on very early phases of this project. The authors are grateful to the organizers of the 2024 workshop "Beyond classical regimes in statistical inference and machine learning" in CIMI Toulouse where this collaboration was initiated. For the purpose of Open Access, the authors have applied a CC BY public copyright license to any Author Accepted Manuscript (AAM) version arising from this submission.

References

[1] 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.

[2] 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.

[3] Sanjeev Arora, Nadav Cohen, and Elad Hazan. On the optimization of deep networks: Implicit acceleration by overparameterization. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 244–253. PMLR, 10–15 Jul 2018.

[4] Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[5] Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. *Advances in neural information processing systems*, 36:48314–48362, 2023.

[6] Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. *ArXiv*, abs/2111.00034, 2021.

[7] Navid Azizan and Babak Hassibi. Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. *arXiv preprint arXiv:1806.00952*, 2018.

[8] Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, 2022.

[9] Daniel Beaglehole, Peter Súkeník, Marco Mondelli, and Mikhail Belkin. Average gradient outer product as a mechanism for deep neural collapse. *arXiv preprint arXiv:2402.13728*, 2024.

[10] 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.

[11] Lukas Braun, Clémentine Dominé, James Fitzgerald, and Andrew Saxe. Exact learning dynamics of deep linear networks with prior knowledge. In *Advances in Neural Information Processing Systems*, volume 35, pages 6615–6629. Curran Associates, Inc., 2022.

[12] Yuhang Cai, Kangjie Zhou, Jingfeng Wu, Song Mei, Michael Lindsey, and Peter L Bartlett. Implicit bias of gradient descent for non-homogeneous deep networks. *arXiv preprint arXiv:2502.16075*, 2025.

[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 Minh Nguyen, Tho Tran, Hung The Tran, Hung Tran, and Nhat Ho. Neural collapse in deep linear networks: From balanced to imbalanced data. In *International Conference on Machine Learning*, 2023.

[15] Cl'ementine Carla Juliette Domin'e, Nicolas Anguita, Alexandra M Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew Saxe. From lazy to rich: Exact learning dynamics in deep linear networks. *ArXiv*, abs/2409.14623, 2024.

[16] Mathieu Even, Scott Pesme, Suriya Gunasekar, and Nicolas Flammarion. (s) gd over diagonal linear networks: Implicit bias, large stepsizes and edge of stability. *Advances in Neural Information Processing Systems*, 36:29406–29448, 2023.

[17] Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. *arXiv preprint arXiv:2502.04664*, 2025.

[18] 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):e2103091118, 2021.

[19] Tomer Galanti, András György, and Marcus Hutter. On the role of neural collapse in transfer learning. *arXiv preprint arXiv:2112.15121*, 2021.

[20] Peifeng Gao, Qianqian Xu, Yibo Yang, Peisong Wen, Huiyang Shao, Zhiyong Yang, Bernard Ghanem, and Qingming Huang. Towards demystifying the generalization behaviors when neural collapse emerges. *ArXiv*, abs/2310.08358, 2023.

[21] Connall Garrod and Jonathan Keating. The persistence of neural collapse despite low-rank bias. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 8860–8916. Curran Associates, Inc., 2025.

[22] Connall Garrod and Jonathan P. Keating. Unifying low dimensional spectra in deep learning. In *Forty-third International Conference on Machine Learning*, 2026.

[23] Gauthier Gidel, Francis Bach, and Simon Lacoste-Julien. Implicit regularization of discrete gradient dynamics in linear neural networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

[24] Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pages 1832–1841. PMLR, 2018.

[25] Suriya Gunasekar, Jason D. Lee, Daniel Soudry, and Nathan Srebro. Implicit bias of gradient descent on linear convolutional networks. In *Neural Information Processing Systems*, 2018.

[26] 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.

[27] X.Y. Han, Vardan Papyan, and David L. Donoho. Neural collapse under MSE loss: Proximity to and dynamics on the central path. In *International Conference on Learning Representations*, 2022.

[28] Hangfeng He and Weijie Su. A law of data separation in deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, 120, 2022.

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

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

[31] Kathy J Horadam. *Hadamard matrices and their applications*. Princeton university press, 2012.

[32] Like Hui, Mikhail Belkin, and Preetum Nakkiran. Limitations of neural collapse for understanding generalization in deep learning. *arXiv preprint arXiv:2202.08384*, 2022.

[33] 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.

[34] 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.

[35] 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.

[36] Ziwei Ji, Miroslav Dudík, Robert E. Schapire, and Matus Telgarsky. Gradient descent follows the regularization path for general losses. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2109–2136. PMLR, 09–12 Jul 2020.

[37] Ziwei Ji and Matus Telgarsky. Gradient descent aligns the layers of deep linear networks. In *International Conference on Learning Representations*, 2019.

[38] Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 17176–17186. Curran Associates, Inc., 2020.

[39] 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.

[40] Addison Kristanto Julistiono, Davoud Ataee Tarzanagh, and Navid Azizan. Optimizing attention with mirror descent: Generalized max-margin token selection. *arXiv preprint arXiv:2410.14581*, 2024.

[41] Vignesh Kothapalli, Ebrahim Rasromani, and Vasudev Awatramani. Neural collapse: A review on modelling principles and generalization. *Trans. Mach. Learn. Res.*, 2023, 2022.

[42] Vignesh Kothapalli and Tom Tirer. Kernel vs. kernel: Exploring how the data structure affects neural collapse. *arXiv preprint arXiv:2406.02105*, 2024.

[43] Daniel Kunin, Allan Ravent'os, Cl'ementine Carla Juliette Domin'e, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. *ArXiv*, abs/2406.06158, 2024.

[44] Hannah Lawrence, Kristian Georgiev, Andrew Dienes, and Bobak T Kiani. Implicit bias of linear equivariant networks. *arXiv preprint arXiv:2110.06084*, 2021.

[45] Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht. First-order methods almost always avoid strict saddle points. *Mathematical programming*, 176(1):311–337, 2019.

[46] Jianing Li and Vardan Papyan. Residual alignment: uncovering the mechanisms of residual networks. *Advances in Neural Information Processing Systems*, 36, 2024.

[47] 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.

[48] Xiao Li, Sheng Liu, Jinxin Zhou, Xinyu Lu, Carlos Fernandez-Granda, Zhihui Zhu, and Qing Qu. Principled and efficient transfer learning of deep models via neural collapse. *arXiv preprint arXiv:2212.12206*, 2022.

[49] 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.

[50] Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

[51] Nischal Mainali and Lucas Teixeira. Exact learning dynamics of in-context learning in linear transformers and its application to non-linear transformers. *ArXiv*, abs/2504.12916, 2025.

[52] Hancheng Min, Zhihui Zhu, and René Vidal. Neural collapse under gradient flow on shallow relu networks for orthogonally separable data. *ArXiv*, abs/2510.21078, 2025.

[53] Dustin G. Mixon, Hans Parshall, and Jianzong Pi. Neural collapse with unconstrained features. *Sampling Theory, Signal Processing, and Data Analysis*, 20, 2020.

[54] Mor Shpigel Nacson, Jason Lee, Suriya Gunasekar, Pedro Henrique Pamplona Savarese, Nathan Srebro, and Daniel Soudry. Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR, 2019.

[55] Vardan Papyan, Xuemei Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences of the United States of America*, 117:24652 – 24663, 2020.

[56] Liam Parker, Emre Onal, Anton Stengel, and Jake Intrater. Neural collapse in the intermediate hidden layers of classification neural networks. *ArXiv*, abs/2308.02760, 2023.

[57] Scott Pesme, Radu-Alexandru Dragomir, and Nicolas Flammarion. Implicit bias of mirror flow on separable data. *Advances in Neural Information Processing Systems*, 37:114094–114121, 2024.

[58] Akshay Rangamani, Marius Lindegaard, Tomer Galanti, and Tomaso A Poggio. Feature learning in deep classifiers through intermediate neural collapse. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28729–28745. PMLR, 23–29 Jul 2023.

[59] Noam Razin and Nadav Cohen. Implicit regularization in deep learning may not be explainable by norms. *Advances in neural information processing systems*, 33:21174–21187, 2020.

[60] Saharon Rosset, Ji Zhu, and Trevor Hastie. Margin maximizing loss functions. *Advances in neural information processing systems*, 16, 2003.

[61] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *CoRR*, abs/1312.6120, 2013.

[62] 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.

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

[64] Jingtong Su, Ya Shi Zhang, Nikolaos Tsilivis, and Julia Kempe. On the robustness of neural collapse and the neural collapse of robustness. *arXiv preprint arXiv:2311.07444*, 2023.

[65] Peter S'uken'ik, Christoph H. Lampert, and Marco Mondelli. Neural collapse is globally optimal in deep regularized resnets and transformers. *ArXiv*, abs/2505.15239, 2025.

[66] 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.

[67] 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.

[68] Haoyuan Sun, Kwangjun Ahn, Christos Thrampoulidis, and Navid Azizan. Mirror descent maximizes generalized margin and can be implemented efficiently. *Advances in Neural Information Processing Systems*, 35:31089–31101, 2022.

[69] Salma Tarmoun, Guilherme Franca, Benjamin D Haeffele, and Rene Vidal. Understanding the dynamics of gradient flow in overparameterized linear models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 10153–10161. PMLR, 18–24 Jul 2021.

[70] Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023.

[71] Christos Thrampoulidis, Ganesh Ramachandra Kini, Vala Vakilian, and Tina Behnia. Imbalance trouble: Revisiting neural-collapse geometry. *Advances in Neural Information Processing Systems*, 35:27225–27238, 2022.

[72] Tom Tirer and Joan Bruna. Extended unconstrained features model for exploring deep neural collapse. In *International Conference on Machine Learning*, 2022.

[73] 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.

[74] 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.

[75] Bhavya Vasudeva, Puneesh Deora, and Christos Thrampoulidis. Implicit bias and fast convergence rates for self-attention. *Transactions on Machine Learning Research*, 2024.

[76] Jingfeng Wu, Vladimir Braverman, and Jason D Lee. Implicit bias of gradient descent for logistic regression at the edge of stability. *Advances in Neural Information Processing Systems*, 36:74229–74256, 2023.

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

[78] Yingwen Wu, Ruiji Yu, Xinwen Cheng, Zhengbao He, and Xiaolin Huang. Pursuing feature separation based on neural collapse for out-of-distribution detection. *arXiv preprint arXiv:2405.17816*, 2024.

[79] Yibo Yang, Shixiang Chen, Xiangtai Li, Liang Xie, Zhouchen Lin, and Dacheng Tao. Inducing neural collapse in imbalanced learning: Do we really need a learnable classifier at the end of deep neural network? *Advances in neural information processing systems*, 35:37991–38002, 2022.

[80] Can Yaras, P. Wang, Wei Hu, Zhihui Zhu, Laura Balzano, and Qing Qu. The law of parsimony in gradient descent for learning deep linear networks. *ArXiv*, abs/2306.01154, 2023.

[81] Emanuele Zangrando, Piero Deidda, Simone Brugiapaglia, Nicola Guglielmi, and Francesco Tudisco. Neural rank collapse: Weight decay and small within-class variability yield low-rank bias. *arXiv:2402.03991*, 2024.

[82] Chenyang Zhang, Difan Zou, and Yuan Cao. The implicit bias of adam on separable data. *Advances in Neural Information Processing Systems*, 37:23988–24021, 2024.

[83] Jiawei Zhang, Yufan Chen, Cheng Jin, Lei Zhu, and Yuantao Gu. Epa: Neural collapse inspired robust out-of-distribution detector. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6515–6519. IEEE, 2024.

[84] Yize Zhao, Tina Behnia, Vala Vakilian, and Christos Thrampoulidis. Implicit geometry of next-token prediction: From language sparsity patterns to model representations. *ArXiv*, abs/2408.15417, 2024.

[85] Yize Zhao and Christos Thrampoulidis. On the geometry of semantics in next-token prediction. *arXiv preprint arXiv:2505.08348*, 2025.

[86] Jinxin Zhou, Xiao Li, Tianyu Ding, Chong You, Qing Qu, and Zhihui Zhu. On the optimization landscape of neural collapse under MSE loss: Global optimality with unconstrained features. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 27179–27202. PMLR, 17–23 Jul 2022.

[87] Jinxin Zhou, Chong You, Xiao Li, Kangning Liu, Sheng Liu, Qing Qu, and Zhihui Zhu. Are all losses created equal: A neural collapse perspective. In *Advances in Neural Information Processing Systems*, volume 35, pages 31697–31710. Curran Associates, Inc., 2022.

[88] 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*, 34:29820–29834, 2021.

Table of Contents

1 Introduction **1**

1.1 Contributions 2

2 Related work **3**

3 Background **3**

4 Hadamard initialization **4**

4.1 Diagonalizing the softmax 4

4.2 Tractable cross-entropy dynamics 5

5 Cross-entropy dynamics and implicit bias **6**

5.1 NC as a candidate among stable directions 6

5.2 Convergence to neural collapse via a Lyapunov function 7

5.3 Non-monotonic behavior 7

6 Standard initialization schemes lead to neural collapse **8**

7 Numerical experiments **9**

8 Conclusion **10**

A Summary of notation **18**

B Further discussion on related work **18**

C Further comparisons to mean squared error dynamics **19**

C.1 Linearized dynamics 20

C.2 Comparison to logit optimization 20

D Further theoretical results **21**

D.1 Reduction to logit singular values 21

D.2 Logit singular value dynamics as mirror flow 21

D.3 NC is the only full-rank stable direction 21

D.4 Additional examples of non-monotonic behavior 22

D.4.1 Reverse-KL metric 22

D.4.2 Frobenius distance 22

D.4.3 Log-ratio of max-min singular values 22

E Non-uniform convergence directions **24**

F Extension to L_2-regularized training **25**

F.1 Hadamard reduction 25

F.2 Normalized dynamics and the KL divergence Lyapunov function 26

G Proofs **26**

G.1 Proof of Theorem 1 26
G.2 Proof of Theorem 2 27
G.3 Proof of Theorem 3 28
G.4 Proof of Theorem 4 31
G.5 Proof of Theorem 5 33
G.6 Proof of Theorem 6 36
G.7 Proof of Theorem 7 38
G.8 Proof of Theorem 8 38

H Further numerical experiments **40**

H.1 Further details on figures in main body 40
H.1.1 Fig. 1 40
H.1.2 Fig. 3 40
H.2 Metrics used in the experiments 41
H.3 Additional random-versus-Hadamard experiments for $K = 4$ and $K = 16$ 41
H.4 The effect of narrow width 41
H.5 The effect of initialization scale 43
H.6 Normalized gradient flow 43

A Summary of notation

Table 1: Summary of Key Notations

Notation	Description	Context / Definition
K	Number of classes.	Assumed to be $K = 2^m$.
n	Number of samples per class.	Classes are assumed to be balanced.
W	Last-layer weight matrix.	$\mathbb{R}^{K\times d}$
H	Feature matrix.	$\mathbb{R}^{d\times Kn}$
Z	Logit matrix.	$Z = WH$
Y	One-hot label matrix.	$Y = I_K \otimes 1_n^T$
$P(Z)$	Softmax matrix.	Softmax applied columnwise to Z.
S	Simplex ETF matrix.	$S = I_K - \frac{1}{K}1_K 1_K^T$
Φ	Sylvester Hadamard matrix.	$\mathbb{R}^{K\times K}$ matrix (Def. 1).
Ψ	Reduced dynamics matrix.	$1_{K-1}1_{K-1}^T - \Phi[2:K, 2:K]$.
U, V	Left/Right singular vector matrices.	$U = V = \frac{1}{\sqrt{K}}\Phi$ for Hadamard Init.
α_i, β_i	Singular values.	Non-trivial singular values of W and H.
a	Logit singular values (vector).	$a = (a_1, \ldots, a_{K-1})$, $a_i = \alpha_i\beta_i$.
$\hat{a}$	Normalized logit singular values.	$\hat{a}_i = a_i/\|a\|_1$.
b, D	Terms in reduced dynamics.	Defined in Eq. 9.
$E(a)$	Energy functional.	Reduced energy functional for a (Eq. 8).
M	L_2-based metric for NC.	Measures distance from $\hat{a}$ to uniform (Eq. 11).
D_{KL}	KL divergence.	Used as a Lyapunov function (Thm. 3).

Throughout, $\mathbb{R}$ denotes the reals and $\mathbb{N}$ the natural numbers. For a positive integer j, we write $[j] = \{1, \ldots, j\}$, and $e_\ell \in \mathbb{R}^j$ denotes the ℓ-th standard basis vector. I_k represents the $k \times k$ identity matrix, and 1_k denotes an all-ones vector of dimension k. For a given matrix A, A^T is its transpose, and $A[i_1 : i_2, j_1 : j_2]$ indicates the submatrix sliced from rows i_1 to i_2 and columns j_1 to j_2. The operator $\mathrm{diag}(v)$ forms a diagonal matrix using the elements of vector v. We use $\otimes$ for the Kronecker product, $\odot$ for the Hadamard (elementwise) product and $\oplus$ for the bitwise XOR. Finally, $\|\cdot\|_F$ and $\|\cdot\|_1$ denote the Frobenius norm and the L_1 norm, respectively. Some key notation is summarized in Table 1.

B Further discussion on related work

This section provides additional context beyond the discussion of the most closely related prior work in the main text.

Neural collapse. First identified in deep classification networks by Papyan et al. [55], NC has since become a central focus of theoretical analysis. Researchers have extensively studied the conditions under which NC emerges, primarily using the UFM [53, 18] with explicit L_2 regularization. Within this framework, NC has been studied under a variety of loss functions [87, 53, 35, 27], consistently emerging as a globally optimal structure. Complementary landscape analyses have revealed favorable optimization properties, where the loss surface consists only of global minima and non-degenerate saddles [86, 88, 35]. Beyond the UFM, NC has also been investigated in other theoretical settings [34, 81]. Extensions of NC research have explored its behavior in settings with a large number of classes [39] and imbalanced class distributions [79, 71, 10, 18, 14, 29], where generalized forms of implicit geometries beyond the simplex ETF have been observed, beginning with work by Thrampoulidis et al. [71]. Generalizations from the hard label setting have also been considered for the multilabel setting [47], soft-labeled language setting [84], and multi-class regression [1]. The influence of dataset properties has also been explored [30, 42]. Researchers have further asked whether NC extends beyond the final layer, with evidence coming both from empirical studies [28, 58, 56] and from analyses of deep UFMs [72, 14, 22, 67, 66, 21], as well as from work beyond the UFM framework altogether [9].

NC has also been connected to a range of practical aspects of deep learning, including adversarial robustness [64], transfer learning [19, 48], generalization [32, 20], and out-of-distribution detection

[78, 83]. The impact of network architecture has been studied in ResNets [46, 65] and large language models (LLMs) [77, 85]. For an overview of NC see the survey by Kothapalli et al. [41].

A recent work by Min et al. [52] considers NC on orthogonally separable data using CE loss. We differ in important ways. First, orthogonal separability requires same-class inputs to be positively correlated and different-class inputs to be strictly negatively correlated, which differs from the canonical UFM studied here, where different examples are orthogonal, rather than negatively correlated. Second, their multiclass analysis requires a semi-local, class-aligned initialization assumption, which they explicitly attribute to the difficulty of characterizing the relevant regions of attraction. This further highlights the core difficulty associated with multiclass CE dynamics that motivates our work. Our approach to this problem is different: we work directly in the orthogonal-input setting and obtain an exact path-level reduction by identifying a Hadamard spectral subspace preserved by the softmax.

Despite this extensive literature, ours is the first work to prove that spectral gradient flow on the unregularized CE UFM converges to NC.

Implicit bias. A large body of work studies how gradient descent or gradient flow induces structure in the absence of explicit regularization. For logistic regression, Soudry et al. [63] showed convergence to the max-margin solution. Ji & Telgarsky [36] related the implicitly regularized path to explicitly regularized solutions in binary classification, which had been earlier studied by Rosset et al. [60]. Analyses of limiting behavior have since been extended to homogeneous [50, 38] and non-homogeneous networks [12], alternative optimizers [24, 7, 68, 57, 82, 17], large step sizes [16, 76], as well as convolutional architectures [25, 44] and transformers [5, 70, 75, 40].

Many of these works focus on binary settings, where no interaction arises between different singular modes. This includes work by Ji & Telgarsky [37, 38], which shows that the singular vectors of the parameter matrices align, a result that implies NC in the binary case. But the NC geometry is practically relevant in multiclass settings, as the literature surveyed above attests.

A complementary thread investigates implicit bias in linear networks with MSE loss, including the effects of gradient descent and network depth [3, 80, 73, 8], matrix factorization [49, 4, 59, 26], and convergence properties [2]. Owing to the simplicity of MSE, however, these analyses differ substantially from our setting.

Implicit bias has also been studied directly in the unregularized CE UFM. Ji et al. [35] proved convergence to a KKT point of a constrained norm minimization problem. Thrampoulidis et al. [71] characterized the solution found by the regularization path under imbalanced classes. Our results are the first to explicitly characterize the implicit bias of gradient-descent dynamics under spectral initialization in unregularized settings, and thus provide the first definitive link between implicit bias theory and neural collapse geometries.

Spectral dynamics. Spectral dynamics analyses study deep learning phenomena by reducing gradient flow to evolution equations for the singular values of the parameter matrices and solving these equations explicitly. To date, such works focus exclusively on linear networks trained with MSE loss. The approach was pioneered by Saxe et al. [61, 62] and extended by Gidel et al. [23] to weaker assumptions and to discrete gradient descent. These studies revealed that networks learn the hierarchical structure of singular values sequentially. Subsequent work generalized the technique to alternative initialization schemes [43, 11, 69] and to investigations of the lazy-to-rich regime transition [11, 15]. Recent efforts have sought to broaden the scope beyond linear architectures. Mainali et al. [51] adapted these methods to transformers, though with simplifications including linear attention mechanisms in place of softmax. Similarly, Zhao et al. [85] employed exact dynamics within the UFM to investigate semantic emergence from language data. However, all existing exact dynamics analyses—including these recent extensions—remain limited to MSE loss. Our work breaks this barrier by extending the framework to the CE loss, a more challenging yet practically essential setting.

C Further comparisons to mean squared error dynamics

We have seen that CE dynamics are fundamentally different from MSE dynamics via non-monotonic convergence and slow convergence rates. Here, we further emphasize this difference by identifying the simplifications the CE setting would need to exhibit the simple behavior of MSE.

C.1 Linearized dynamics

The MSE behavior of independent sigmoidal singular value evolution can be recovered in CE dynamics, but only under a severe restriction: the small $\|a\|_1$ regime where linearized dynamics apply. As the following theorem shows, the singular values decouple in this limit and follow, to leading orders, the same sigmoidal dynamics as in the MSE case. However, even with small initialization, the system quickly escapes this regime, after which coupling effects introduce the complex behaviors discussed in the main text. The proof appears in App. G.7.

Theorem 7. *Let $K = 2^m$, where $m \in \mathbb{N}$. Let $a(t) \in \mathbb{R}^{K-1}$ evolve under the dynamics of Eq.* (9)*, with initialization $a_i(0) > 0$ for all $i = 1, ..., K-1$. Expanding to second order in the scale $\|a\|_1$, the dynamics reduce to*

$$\frac{da_i}{dt} = a_i \left[1 - a_i\right] + O\left(\|a\|_1^3\right).$$

In particular, the linearized dynamics decouple, and can therefore be solved exactly. The resulting trajectories, which are accurate in the linearized regime, are

$$a_i(t) = \frac{a_i(0)}{a_i(0) + (1 - a_i(0))e^{-t}}.$$

The key contrast with MSE is that this decoupled behavior persists at all scales for MSE, whereas for CE it is confined to the small $\|a\|_1$ regime. Moreover, the linearized equations imply $a_i/a_j \to 1$ monotonically, pushing the early-stage dynamics toward NC—though the system exits this regime too quickly for this effect alone to guarantee monotonic convergence to NC.

C.2 Comparison to logit optimization

The second simplification that recovers MSE-like behavior is removing the factorization layer: optimizing logits directly rather than through the product WH. This is analogous to the convex one-layer network analyzed by Saxe et al. [61]. The loss in this setting is

$$\mathcal{L}(Z) = \sum_{c=1}^{K} \sum_{i=1}^{n} \log\left(1 + \sum_{c' \neq c}^{K} e^{(z_{ic})_{c'} - (z_{ic})_c}\right), \tag{12}$$

where z_{ic} are the columns of the logit matrix Z arranged in class order. With Hadamard initialization of the logit matrix Z, we have the following theorem showing the corresponding reduction in the dynamics, and consequently monotonic convergence to NC.

Theorem 8. *Let $K = 2^m$ for $m \in \mathbb{N}$, and let $U = \frac{1}{\sqrt{K}}\Phi$, where Φ is the $K \times K$ Sylvester Hadamard matrix. Consider the optimization problem described in Eq.* (12)*, and initialize the logit matrix as $Z = UD_Z V^T$, with $V = U \otimes Q$, where Q is a right singular-vector matrix of 1_n^T, and D_Z has its nonzero singular values given by $a_1, ..., a_K$. Then the gradient flow equation on this objective, after absorbing a factor of $\frac{1}{K\sqrt{n}}$ into the a variable, a factor of $\frac{1}{K}$ into the t variable, and dropping the trivial singular value, reduces to*

$$\frac{da_i}{dt} = \frac{1}{D} b_i, \qquad i = 1, ..., K-1,$$

where the quantities D and b are defined as in Eq. (9)*. Moreover, the metric M in Eq.* (11) *decreases monotonically along trajectories of gradient flow.*

This result parallels that of Soudry et al. [63] for logistic regression, since NC corresponds to the max-margin solution of the UFM. We see that, when optimizing logits directly, the system converges monotonically to the NC solution—just as in the MSE case studied by Saxe et al. [61, 62]. Taken together, this highlights the role of depth: it is precisely the additional layer that disrupts monotonicity and introduces the more intricate behaviors observed in the CE UFM.

D Further theoretical results

D.1 Reduction to logit singular values

Consider the dynamics derived in Theorem 2. Under the balancedness condition $\alpha_i = \beta_i$, the logit singular values evolve as

$$\frac{da_i}{dt} = \frac{d\alpha_i}{dt}\beta_i + \alpha_i \frac{d\beta_i}{dt} = \frac{\sqrt{n}}{D} b_i(\alpha_i^2 + \beta_i^2) = \frac{2\sqrt{n}}{D} b_i a_i.$$

Then redefining

$$a' = \frac{1}{K\sqrt{n}} a, \quad t' = 2t\sqrt{n},$$

gives the gradient flow equations given in Eq. (9), once primes are dropped, since this represents a simple rescaling that does not change the qualitative dynamics.

D.2 Logit singular value dynamics as mirror flow

Recall the dynamics of the logit singular values from Eq. (9) are given as

$$\frac{da_i}{dt} = a_i \cdot \frac{\sum_{j=1}^{K-1} \Psi_{ij} e^{-(\Psi a)_j}}{1 + \sum_{j=1}^{K-1} e^{-(\Psi a)_j}}, \qquad i \in [K-1]. \tag{13}$$

Define the energy functional

$$E(a) = \log\left(1 + \sum_{j=1}^{K-1} \exp\left(-(\Psi a)_j\right)\right).$$

Then, the dynamics in Eq. (13) can be expressed conveniently in vector form as

$$\frac{\mathrm{d}a}{\mathrm{d}t} = -a \odot \nabla_a E(a).$$

For positive initialization, $a > 0$ in the entire trajectory, and thus it holds equivalently that

$$\frac{\mathrm{dlog}(a)}{\mathrm{d}t} = -\nabla_a E(a).$$

Moreover, $\frac{\mathrm{dlog}(a)}{\mathrm{d}t} = \frac{\mathrm{d}\nabla_a \phi(a)}{\mathrm{d}t}$ where $\phi(a) = \sum_{i=1}^{K-1} a_i \log(a_i)$ is the negative entropy potential. Overall, we have expressed the dynamics of the logit singular values in the form of mirror-flow updates with respect to the entropy potential:

$$\frac{\mathrm{d}\nabla_a \phi(a)}{\mathrm{d}t} = -\nabla_a E(a).$$

D.3 NC is the only full-rank stable direction

Given the dynamics of Eq. (10), if we look for a stable direction, meaning $\frac{d\hat{a}_i}{dt} = 0$ for $i = 1, ..., K-1$, that also satisfies full-rank, meaning $\hat{a}_i > 0$ for $i = 1, \ldots, K-1$, then this reduces to the condition:

$$b_i = \bar{b}, \quad \text{for } i = 1, ..., K-1.$$

This means that the vector b can be written as $b = \bar{b} 1_{K-1}$, which imposes the following condition on the vector a

$$\Psi e^{-(\Psi a)} = \bar{b} 1_{K-1}.$$

Then using that Ψ is invertible and $\Psi^{-1} 1_{K-1} = \frac{1}{K} 1_{K-1}$ gives:

$$a = -\frac{1}{K} \log\left(\frac{1}{K}\bar{b}\right) 1_{K-1},$$

and if $a \propto 1_{K-1}$, it implies $\hat{a} = \frac{1}{K-1} 1_{K-1}$. Hence, the only full-rank stable direction is $\hat{a} = \frac{1}{K-1} 1_{K-1}$.

D.4 Additional examples of non-monotonic behavior

D.4.1 Reverse-KL metric

One can consider the reverse-KL divergence, defined as

$$D_{\mathrm{KL}}\left(\hat{a} \,\|\, \frac{1}{K-1}1_{K-1}\right) = \log(K-1) + \sum_i \hat{a}_i \log(\hat{a}_i).$$

We will denote it by $\tilde{D}_{\mathrm{KL}}$ for this section so as to differentiate it from the KL divergence of the main text. It has derivative

$$\frac{d\tilde{D}_{\mathrm{KL}}}{dt} = \frac{1}{D}\sum_i [b_i - \bar{b}]\hat{a}_i \log(\hat{a}_i).$$

In $K = 8$, a counterexample demonstrating that $\tilde{D}_{\mathrm{KL}}$ is not monotonically decreasing is

$$a = [1, 1, 1, 1.25, 0.0001, 0.0001, 0.0001],$$

where $\frac{d\tilde{D}_{\mathrm{KL}}}{dt} \approx 0.0037$.

D.4.2 Frobenius distance

Typically in other works, the rate of convergence to NC is measured through the following metric:

$$\tilde{M} = \left\| \frac{Z}{\|Z\|_F} - \frac{S}{\|S\|_F} \right\|_F^2,$$

which, when expressed in terms of our logit singular values (after absorbing constants as in Eq. (9)), is given by

$$\tilde{M} = \sum_{i=1}^{K-1} \left[\frac{a_i}{\sqrt{\sum_j a_j^2}} - \frac{1}{\sqrt{K-1}} \right]^2.$$

In the main text, to avoid introducing multiple normalizations, we normalized with the L_1 norm, but the metric $\tilde{M}$ displays the same behavior as the metric M. First note that its derivative is given by

$$\frac{d\tilde{M}}{dt} = \frac{2}{D\|Z\|_F^3\sqrt{K-1}} \left[\sum_{ij} a_i a_j b_j (a_j - a_i) \right],$$

where we continue to denote $\sqrt{\sum_j a_j^2} = \|Z\|_F$. For this to be non-positive we require that:

$$\sum_{ij} a_i a_j b_j [a_j - a_i] \leq 0.$$

A counterexample occurs for $a = [1, 1, 1, 1.25, 0.0001, 0.0001, 0.0001]$, where the LHS of the above inequality is ≈ 0.088. Note the existence of this counterexample for $K = 8$ is sufficient for counterexamples to exist for all $K \geq 8$, using the same reasoning as provided for the proof of Theorem 4 in App. G.4.

D.4.3 Log-ratio of max-min singular values

Define the following metric:

$$M_{\mathrm{LR}} = \max_i(\log(a_i)) - \min_j(\log(a_j)).$$

This quantity is monotonically decreasing for $K = 8$, but non-monotonic for $K = 16$.

We first prove this quantity is monotonically decreasing for $K = 8$:

Proposition 1. *Let $K = 8$, and $a(t) \in \mathbb{R}^{K-1}$ be a variable initialized with $a_i(0) > 0$ for all i that evolves under the time evolution of Eq.* (9)*. Then the metric M_{LR} is monotonically decreasing.*

Proof. Recall our dynamics, given in Eq. (9). By multiplying the numerator and denominator by $\exp(2\sum_j a_j)$, we can rewrite this as

$$\frac{d}{dt}(a_i) = \frac{1}{D} b_i a_i,$$

where

$$b_i = \sum_j \Psi_{ij} \exp\left(\left(\left[2(1_{K-1} 1_{K-1}^T) - \Psi \right] a \right)_j \right),$$
$$D = \exp(2\sum_j a_j) + \sum_j \exp\left(\left(\left[2(1_{K-1} 1_{K-1}^T) - \Psi \right] a \right)_j \right).$$

Expanding this out directly produces the following expressions for the variables b_i

$$b_1 = \exp(2(a_2+a_4+a_6)) + \exp(2(a_3+a_4+a_7)) + \exp(2(a_2+a_5+a_7)) + \exp(2(a_3+a_5+a_6))$$
$$b_2 = \exp(2(a_1+a_4+a_5)) + \exp(2(a_3+a_4+a_7)) + \exp(2(a_1+a_6+a_7)) + \exp(2(a_3+a_5+a_6))$$
$$b_3 = \exp(2(a_2+a_4+a_6)) + \exp(2(a_1+a_4+a_5)) + \exp(2(a_2+a_5+a_7)) + \exp(2(a_1+a_6+a_7))$$
$$b_4 = \exp(2(a_1+a_2+a_3)) + \exp(2(a_2+a_5+a_7)) + \exp(2(a_1+a_6+a_7)) + \exp(2(a_3+a_5+a_6))$$
$$b_5 = \exp(2(a_2+a_4+a_6)) + \exp(2(a_3+a_4+a_7)) + \exp(2(a_1+a_2+a_3)) + \exp(2(a_1+a_6+a_7))$$
$$b_6 = \exp(2(a_1+a_4+a_5)) + \exp(2(a_3+a_4+a_7)) + \exp(2(a_1+a_2+a_3)) + \exp(2(a_2+a_5+a_7))$$
$$b_7 = \exp(2(a_2+a_4+a_6)) + \exp(2(a_1+a_4+a_5)) + \exp(2(a_1+a_2+a_3)) + \exp(2(a_3+a_5+a_6))$$

From this one can note the following properties immediately:

1. There are seven triples of the singular values a_i that appear in the exponents.
2. The triples that feature a_i do not appear in the expression of b_i.
3. Each a_i appears with every other a_j in exactly one triple.
4. Every pair of b_i, b_j shares exactly two terms, and differ in two. The two they share are precisely the two not featuring either of their corresponding a_i.

Now suppose that at some time t we have that the largest mode is given by a_i and the smallest by a_j, then the log ratio of these singular values has derivative given by

$$\frac{d}{dt}\left(\log(a_i) - \log(a_j)\right) = \frac{1}{D}[b_i - b_j].$$

Because of the properties of the expressions for the variables b_i described above, the difference $b_i - b_j$ takes the following form:

$$b_i - b_j = x_j x_p x_q + x_j x_r x_s - x_i x_r x_p - x_i x_s x_q,$$

where we have defined $x_i = \exp(2a_i)$ for the purpose of this section. We can alternatively write this expression as

$$b_i - b_j = x_p[x_j x_q - x_i x_r] + x_s[x_j x_r - x_i x_q],$$

and further as

$$b_i - b_j = x_p x_i x_q \left[\frac{x_j}{x_i} - \frac{x_r}{x_q} \right] + x_s x_i x_r \left[\frac{x_j}{x_i} - \frac{x_q}{x_r} \right].$$

Now note that, because x_j is the smallest, and x_i is the largest, we must have:

$$x_j \leq \text{any other choice of } x, \qquad \frac{1}{x_i} \leq \frac{1}{\text{any other choice of } x}.$$

This implies that:

$$\frac{x_j}{x_i} \leq \frac{x_r}{x_q}, \qquad \frac{x_j}{x_i} \leq \frac{x_q}{x_r},$$

and hence the difference $b_i - b_j$ is non-positive, and only takes value zero when

$$\frac{x_j}{x_i} = \frac{x_r}{x_q} = \frac{x_q}{x_r},$$

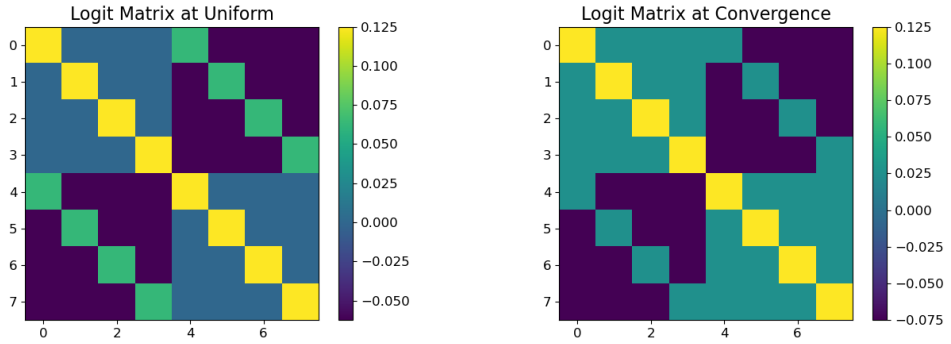


Figure 4: Plots of the normalized logit matrices generated by the vector $a^* = [1, 1, 1, 1, 0, 0, 0]$, which is uniform on its support, and by the non-uniform vector $a_{\text{conv}} = [1, 1, 1, 2, 0, 0, 0]$. The smallest normalized margin achieved by a_{conv} is visibly larger, indicating that this configuration outperforms the uniform one as the logit norm diverges.

but this can only happen if $x_j = x_i$, meaning the largest and smallest mode must be equal, which implies $\hat{a} = \frac{1}{K-1} 1_{K-1}$. Hence the quantity M_{LR} always has a negative derivative, except when we have a neural collapsed solution.

The above shows that, whenever the maximizing and minimizing indices are fixed, $dM_{\text{LR}}/dt \leq 0$. At times when these indices are non-unique, the same argument applies to every maximizing–minimizing pair, so the upper right derivative of M_{LR} is also non-positive. Hence M_{LR} is globally non-increasing.

□

It remains to show a counterexample for $K = 16$. Let

$$a = [5.01, 0.99, 2, 5, 1, 5, 1, 2, 1, 5, 3, 5, 1, 5, 1],$$

then the largest singular value is a_1, and the smallest is a_2, but

$$\frac{d}{dt} \left(\log(a_1) - \log(a_2) \right) \approx 1.8 \times 10^{-13} .$$

Hence this metric is not monotonically decreasing for $K = 16$.

E Non-uniform convergence directions

Here we provide an expanded discussion of the example shown in Figure 2, in order to build intuition for why non-uniform convergence directions can arise in the loss surface for general K. Figure 4 displays the normalized logit matrix with singular-values vector $a^* = [1, 1, 1, 1, 0, 0, 0]$ that lies very close to the initialization used in Figure 2. Since the set of indices for which $a_i \neq 0$ is invariant under the flow, we might expect that the convergence direction reached from the initialization $a(0) = a^*$ is the minimizer among all configurations with at most the first four indices nonzero.

It is straightforward to show that the following quantity decreases monotonically along the flow:

$$E(a) = \log \left(1 + \sum_{j=1}^{K-1} \exp \left(-\|a\|_1 (\Psi \hat{a})_j \right) \right),$$

where we have chosen to express it in terms of $\hat{a}$ and $\|a\|_1$. This coincides with the energy functional of Eq. (8) up to absorbed constants. In particular, as $\|a\|_1 \to \infty$, the energy is asymptotically:

$$E \simeq \sum_{j \in \text{argmin}_u \{(\Psi \hat{a})_u\}} \exp \left(-\|a\|_1 (\Psi \hat{a})_j \right) .$$

Thus, the smallest normalized margin $(\Psi \hat{a})_i$ dominates the energy, and we therefore expect the flow as $t \to \infty$ to maximize this quantity. The normalized margins correspond to the difference between the diagonal and off-diagonal entries of the normalized logit matrix.

Comparing the normalized logit matrix at the uniform-on-support configuration a^* with that of the configuration a_{conv}, as shown in Figure 4, we see that a^* has a smaller minimal margin.

Numerically, what occurs in Figure 2 of the main text is that the last three singular values are initially so small that the dynamics treat them effectively as zero. The remaining singular values, therefore, evolve toward the minimizer supported on only the first four indices. This minimizer is not the uniform point, as indicated by the worse margin of the uniform configuration in Figure 4; instead, it is the configuration a_{conv}, in which one singular value is larger than the others. This is the structure toward which the flow in Figure 2 trends over the considered time interval. Because the flow approaches a non-uniform point, most common measures of uniformity increase, and thus fail to act as Lyapunov functions for the dynamics. The KL divergence avoids this issue by effectively accounting for the growth of the three small singular values, which offsets the non-uniform behavior of the remaining ones.

F Extension to L_2-regularized training

The main text considers the unregularized CE UFM, for which the parameters and logit norms diverge as the loss approaches zero. We now show that the Hadamard reduction extends directly to the L_2-regularized setting. In particular, regularization preserves the invariant spectral subspace of Theorem 2, and the KL divergence Lyapunov argument continues to govern the normalized singular-value dynamics. Unlike in the unregularized setting, however, the singular values converge to a finite-norm NC solution.

For regularization strength $\lambda > 0$, consider the regularized objective

$$\mathcal{L}_\lambda(W, H) = \mathcal{L}(W, H) + \frac{1}{2}\lambda\|W\|_F^2 + \frac{1}{2}\lambda\|H\|_F^2,$$

where $\mathcal{L}$ is the CE objective defined in Eq. (2). Its gradient flow dynamics are

$$\frac{dW}{dt} = (Y - P(WH))H^T - \lambda W, \quad \frac{dH}{dt} = W^T(Y - P(WH)) - \lambda H.$$

F.1 Hadamard reduction

For the regularized case, we can again produce reduced dynamics.

Proposition 2. *Consider the context and initialization of Theorem 2. From this initialization the singular vectors of the parameters remain invariant under gradient flow on $\mathcal{L}_\lambda$. The singular values evolve according to*

$$\frac{d\alpha_i}{dt} = \frac{\sqrt{n}}{D}b_i\beta_i - \lambda\alpha_i, \quad \frac{d\beta_i}{dt} = \frac{\sqrt{n}}{D}b_i\alpha_i - \lambda\beta_i,$$

where b_i, D are defined as in Eq. (7)*.*

Proof.

The CE component of the flow is exactly that considered in Theorem 2. The additional regularization terms are of the form

$$-\lambda W = U(-\lambda D_W)R^T, \quad -\lambda H = R(-\lambda D_H)V^T,$$

and therefore have the same singular vectors as W and H. Hence they only change the singular values and leave singular vectors preserved. Adding these terms to the updates in Theorem 2 then clearly gives the stated dynamics.

□

One minor difference in this case is that the balancedness quantities are no longer conserved. Instead, we have

$$\frac{d}{dt}(\alpha_i^2 - \beta_i^2) = -2\lambda(\alpha_i^2 - \beta_i^2),$$

and hence

$$\alpha_i^2(t) - \beta_i^2(t) = e^{-2\lambda t}(\alpha_i^2(0) - \beta_i^2(0)).$$

This ensures that balancedness emerges asymptotically, since the error is exponentially decreasing in time. Crucially, a balanced initialization $\alpha_i(0) = \beta_i(0)$ remains balanced throughout training. Again, assuming balancedness we can define $a_i = \alpha_i\beta_i$ and produce the evolution equation for the logit singular values

$$\frac{da_i}{dt} = 2\sqrt{n}\frac{1}{D}b_i a_i - 2\lambda a_i.$$

Applying the same rescaling of the time and a variables as used in Sec. 4.2, and defining the effective regularization strength $\rho = \frac{1}{\sqrt{n}}\lambda$ then gives

$$\frac{da_i}{dt} = a_i\left(\frac{1}{D}b_i - \rho\right).$$

Here, b_i and D have the simpler forms given in Eq. (9). Thus L_2-regularization modifies the reduced dynamics only through the shrinking term $-\rho a_i$.

F.2 Normalized dynamics and the KL divergence Lyapunov function

Defining, as in the main text,

$$\hat{a}_i = \frac{a_i}{\|a\|_1}, \quad \bar{b} = \sum_{j=1}^{K-1}\hat{a}_j b_j,$$

the $-\rho a_i$ terms cancel with the corresponding contributions from the derivative of $\|a\|_1$, giving

$$\frac{d\hat{a}_i}{dt} = \frac{\hat{a}_i}{D}(b_i - \bar{b}).$$

This is precisely the normalized flow in Eq. (10). Thus, the explicit L_2 regularization term cancels from the directional dynamics. We emphasize, however, that b_i and D depend on the scale $\|a\|_1$, and so regularization can still affect the directional dynamics indirectly through its effect on the norm.

Since the normalized flow has the same form as in the unregularized setting, the monotonicity conclusion of Theorem 3 continues to hold: the KL divergence is non-increasing along the flow. However, for sufficiently strong regularization the zero solution is the global minimum. In this regime $\|a\|_1 \to 0$, so the limiting direction $\hat{a}$ is not well defined, and KL divergence monotonicity does not imply convergence to the NC direction. Conversely, if the trajectory remains bounded away from zero norm, the strict KL Lyapunov property implies convergence of $\hat{a}$ to the NC direction.

G Proofs

G.1 Proof of Theorem 1

Let ϕ_i, $i \in \{1, 2, \ldots, K\}$, be the rows of the $K \times K$ Sylvester–Hadamard matrix Φ. We use the following "closure" property [31, Lemma 3.6], where $i \oplus j$ denotes bitwise XOR:

Fact 1. *For any $i, j \in \{1, \ldots, K\}$, it holds that $\phi_i \odot \phi_j = \phi_{i\oplus j}$.*

We note that bitwise XOR is naturally defined on zero-based indices. Accordingly, whenever $i \oplus j$ appears for $i, j \in [K]$, we mean $i \oplus j := 1 + ((i-1)\text{XOR}(j-1))$, i.e. XOR is applied to the corresponding indices in $\{0, ..., K-1\}$, and the result is shifted back to $\{1, ..., K\}$. This allows us to maintain one-based indexing throughout the manuscript.

For all $i \in [K]$, define matrices $\Phi_i := \Phi\,\mathrm{diag}\,(\phi_i) = \left[\phi_j^T \odot \phi_i^T\right]_{j\in[K]}$. Note that $\Phi_1 = \Phi$. Using Fact 1, we can express each Φ_i in terms of a corresponding permutation matrix $\Pi_i = \left[e_{i\oplus j}^T\right]_{j\in[K]}$, where $e_\ell \in \mathbb{R}^K$ denotes the ℓ-th standard basis vector:

$$\Phi_i := \Phi\,\mathrm{diag}\,(\phi_i) = \Pi_i\Phi, \quad \text{for all } i \in [K]. \tag{14}$$

Recall that softmax is applied columnwise. Thus, to prove the advertised identity in Eq. (1), it suffices for any column $i \in [K]$ that

$$P(\frac{1}{K}\Phi\,\mathrm{diag}\,(a)\,\phi_i) = \frac{1}{K}\Phi\,\mathrm{diag}\left(\Phi P(\frac{1}{K}\Phi a)\right)\phi_i\,.$$

Using the property $\mathrm{diag}\,(u)\cdot v = \mathrm{diag}\,(v)\cdot u$ for any u, v on both sides of the equality, and applying Eq. (14), this is equivalent to

$$P(\frac{1}{K}\Pi_i \Phi a) = \frac{1}{K}\Pi_i \Phi\Phi P(\frac{1}{K}\Phi a)\,.$$

Since $\Phi\Phi = \Phi\Phi^T = KI_K$, this claim reduces to $P(\frac{1}{K}\Pi_i\Phi a) = \Pi_i P(\frac{1}{K}\Phi a)$, which holds by permutation equivariance of the softmax.

G.2 Proof of Theorem 2

Recall $U = \frac{1}{\sqrt{K}}\Phi$, where Φ is the $K \times K$ Sylvester Hadamard matrix, and that we initialize our parameter matrices as $W = UD_W R^T$, $H = RD_H V^T$, where $R \in \mathbb{R}^{d\times d}$ is orthogonal, and $V = U \otimes Q$ is orthogonal, with Q being a right singular-vector matrix of 1_n^T, meaning it obeys the relation: $1_n^T = \sqrt{n} e_1^{(n)T} Q^T$, where $e_1^{(n)}$ is the first standard basis vector in $\mathbb{R}^n$. Also recall that $D_W \in \mathbb{R}^{K\times d}$ has its only non-zero singular values given by $\alpha_1, \dots \alpha_K$, and $D_H \in \mathbb{R}^{d\times Kn}$ has its only non-zero singular values given by $\beta_1, \dots, \beta_K$. Note as a consequence of the definitions of U, V, we can write $Y = \sqrt{n} U[I_K, 0_{K\times K(n-1)}]V^T$.

We also have at initialization that the logit matrix is given by

$$Z = UD_W D_H V^T.$$

Note that $D_W \in \mathbb{R}^{K\times d}$ and $D_H \in \mathbb{R}^{d\times Kn}$. Hence, defining $D_Z = D_W D_H \in \mathbb{R}^{K\times Kn}$, and $D_a = \mathrm{diag}(a_1, ..., a_K)$, where $a_i = \alpha_i\beta_i$, this matrix can be written as:

$$D_Z = [D_a, 0_{K\times K(n-1)}] = D_a \otimes e_1^{(n)T}.$$

Hence we can write

$$Z = U\left(D_a \otimes e_1^{(n)T}\right)(U\otimes Q)^T = U\left[(D_a U^T)\otimes (e_1^{(n)T}Q^T)\right].$$

Next, using that $1_n^T = \sqrt{n} e_1^{(n)T} Q^T$, this becomes:

$$Z = \frac{1}{\sqrt{n}} U\left[(D_a U^T)\otimes 1_n^T\right].$$

Then using that this has repeated columns, this becomes:

$$Z = (UD_{\tilde{a}}U^T)\otimes 1_n^T,$$

where we define $\tilde{a} = \frac{1}{\sqrt{n}}a$. Thus, although D_Z contains a zero block, the logit matrix Z itself consists of n repeated copies of each column of $UD_{\tilde{a}}U^T$. Since softmax acts independently on each repeated column, Theorem 1 gives

$$P(Z) = (UD_{\tilde{\nu}}U^T)\otimes 1_n^T,$$

where

$$\tilde{\nu}_i = \frac{\sum_j \Phi_{ij} e^{\frac{1}{K}(\Phi\tilde{a})_j}}{\sum_j e^{\frac{1}{K}(\Phi\tilde{a})_j}}.$$

Now consider the gradient flow equations of our model

$$\frac{dW}{dt} = (Y-P)H^T, \qquad \frac{dH}{dt} = W^T(Y-P).$$

Consider the change of variable given by $W = U\tilde{W}$, $H = \tilde{H}V^T$, these have evolution equations given by:

$$\frac{d}{dt}(\tilde{W}) = U^T(Y-P)V\tilde{H}^T, \qquad \frac{d}{dt}(\tilde{H}) = \tilde{W}^T U^T(Y-P)V.$$

Note that, by construction, both the matrices Y, P can be written in SVD form through the matrices U, V

$$Y = \sqrt{n}UD_Y V^T, \qquad P = \sqrt{n}UD_P V^T,$$

where $D_Y = [I_K, 0_{K\times K(n-1)}] \in \mathbb{R}^{K\times Kn}$ and $D_P = [\text{diag}(\tilde{\nu}), 0_{K\times K(n-1)}] \in \mathbb{R}^{K\times Kn}$, and so

$$U^T(Y-P)V = \sqrt{n}[\text{diag}(1_K - \tilde{\nu}), 0_{K\times K(n-1)}],$$

label this matrix as B. Now returning to the evolution equations for $\tilde{H}$ and $\tilde{W}$, we label the rows of $\tilde{W}$ by $\tilde{w}_a$, and the columns of $\tilde{H}$ by $\tilde{h}_a$. These have evolution equations given by:

$$\frac{d}{dt}((\tilde{w}_i)_j) = \sum_x B_{ix}(\tilde{h}_x)_j, \qquad \frac{d}{dt}((\tilde{h}_j)_i) = \sum_x B_{xj}(\tilde{w}_x)_i.$$

Now using that $B_{ix} = \sqrt{n}(1-\tilde{\nu}_i)\delta_{ix}$ gives:

$$\frac{d}{dt}((\tilde{w}_i)_j) = \sqrt{n}(1-\tilde{\nu}_i)(\tilde{h}_i)_j, \qquad \frac{d}{dt}((\tilde{h}_j)_i) = \sqrt{n}(1-\tilde{\nu}_j)(\tilde{w}_j)_i, \tag{15}$$

where in the second of these equations if $j > K$ the derivative is zero.

Note that using our initialization scheme, we have at initialization $\tilde{w}_i = \alpha_i r_i$, $\tilde{h}_i = \beta_i r_i$, for $i = 1, \ldots, K$, and 0 otherwise, where r_i are the columns of R. One observes from Eq. (15) that the derivatives also point in the same direction. Hence this property is maintained at all times, meaning the vectors $\tilde{w}_i, \tilde{h}_i$ only evolve in their coefficients, not in their directions. These coefficients then evolve as:

$$\frac{d}{dt}(\alpha_i) = \sqrt{n}(1-\tilde{\nu}_i)\beta_i, \qquad \frac{d}{dt}(\beta_i) = \sqrt{n}(1-\tilde{\nu}_i)\alpha_i.$$

Translating back to our original parameter matrices W, H, these are the evolution equations for their singular values, and the singular vectors are frozen for all times.

It remains to demonstrate that these equations take the exact form in the theorem statement. Note from the expression of $\tilde{\nu}_i$ we have

$$1 - \tilde{\nu}_i = \frac{1}{\sum_j e^{\frac{1}{K}(\Phi\tilde{a})_j}} \left[\sum_j (1-\Phi_{ij}) e^{\frac{1}{K}(\Phi\tilde{a})_j}\right].$$

Dividing the numerator and denominator by $\exp(\frac{1}{K}\sum_j \tilde{a}_j)$, gives:

$$1 - \tilde{\nu}_i = \frac{1}{\sum_j e^{-\frac{1}{K}([1_K 1_K^T - \Phi]\tilde{a})_j}} \left[\sum_j [1_K 1_K^T - \Phi]_{ij} e^{-\frac{1}{K}([1_K 1_K^T - \Phi]\tilde{a})_j}\right] = \frac{b_i}{D},$$

and hence this gives the desired form, once we recall that $\tilde{a}_i = \frac{1}{\sqrt{n}} a_i$.

G.3 Proof of Theorem 3

The KL divergence $D_{\text{KL}}(p \,\|\, q) = \sum_i p_i \log\left(\frac{p_i}{q_i}\right)$ between two categorical distributions with probabilities p and q is non-negative and equals zero if and only if $p = q$. In our setting, we have

$$D_{\text{KL}}\Big(\frac{1}{K-1}1_{K-1} \,\|\, \hat{a}\Big) = -\log(K-1) - \frac{1}{K-1}\sum_i \log(\hat{a}_i). \tag{16}$$

For the remainder of the proof we denote this by D_{KL}. Differentiating and using Eq. (10), we obtain

$$\frac{d}{dt}\left(D_{\text{KL}}\right) = -\frac{1}{K-1}\sum_i \frac{1}{\hat{a}_i}\frac{d}{dt}(\hat{a}_i) = -\frac{1}{(K-1)D}\left[\left(\sum_i b_i\right) - (K-1)\bar{b}\right], \tag{17}$$

We now define $x_i = (\Psi a)_i$ and rewrite the derivative in terms of these variables, which, recall, play the role of margins. First observe that

$$\bar{b} = \sum_i \hat{a}_i b_i = \sum_{ij} \hat{a}_i \Psi_{ij} e^{-(\Psi a)_j} = \frac{1}{\|a\|_1}\sum_{ij} a_i \Psi_{ij} e^{-(\Psi a)_j} = \frac{1}{\|a\|_1}\sum_j x_j e^{-x_j}. \tag{18}$$

Second, note that

$$\sum_i b_i = \sum_{ij} \Psi_{ij} e^{-(\Psi a)_j} = K \sum_j e^{-(\Psi a)_j} = \left(\sum_i (\Psi \hat{a})_i \right) \sum_j e^{-(\Psi a)_j}, \tag{19}$$

where we used $\sum_i \Psi_{ij} = K = \sum_i (\Psi \hat{a})_i$. Hence, we obtain

$$\sum_i b_i = \frac{1}{\|a\|_1} \Big(\sum_i x_i \Big) \Big(\sum_j e^{-x_j} \Big). \tag{20}$$

Substituting Eqns. (18) and (20) into Eq. (17) yields

$$\frac{d}{dt}(D_{\mathrm{KL}}) = \frac{1}{(K-1)D\|a\|_1} \left[(K-1) \sum_j x_j e^{-x_j} - \left(\sum_i x_i \right) \left(\sum_j e^{-x_j} \right) \right].$$

Now, the bracketed term can be rewritten as $\frac{1}{2} \sum_{ij} (x_i - x_j) \left(e^{-x_i} - e^{-x_j} \right)$, yielding

$$\frac{d}{dt}(D_{\mathrm{KL}}) = \frac{1}{2(K-1)D\|a\|_1} \sum_{ij} (x_i - x_j) \left(e^{-x_i} - e^{-x_j} \right). \tag{21}$$

For each pair (i, j), the terms $x_i - x_j$ and $e^{-x_i} - e^{-x_j}$ have opposite signs unless $x_i = x_j$. Hence, every summand is non-positive, vanishing only when $x_i = x_j$. Hence the full sum is always non-positive, and equal to zero only if x is a uniform vector. Because $D > 0$ and $\|a\|_1 > 0$, the prefactor is strictly positive. Moreover, since Ψ is invertible and has 1_{K-1} as an eigenvector, the condition $x \propto 1_{K-1}$ implies $a \propto 1_{K-1}$, and hence $\hat{a} = \frac{1}{K-1} 1_{K-1}$.

Thus, D_{KL} is non-negative and non-increasing, and both it and its derivative vanish only at $\hat{a} = \hat{u} := \frac{1}{K-1} 1_{K-1}$. It remains to prove that its limiting value is zero.

We proceed in steps as follows.

Normalized logits remain bounded away from zero. Rearranging Eq. (16) we have that

$$\prod_i \hat{a}_i(t) = \frac{e^{-(K-1)D_{\mathrm{KL}}(t)}}{(K-1)^{K-1}}. \tag{22}$$

For every $j \in [K-1]$, $\hat{a}_j(t) \geq \prod_i \hat{a}_i(t)$. Moreover, since $D_{\mathrm{KL}}(t)$ is non-increasing, we have $D_{\mathrm{KL}}(t) \leq D_{\mathrm{KL}}(0)$ for all t. Thus,

$$\forall j \in [K-1], t > 0 \; : \; \hat{a}_j(t) \geq \delta := \frac{e^{-(K-1)D_{\mathrm{KL}}(0)}}{(K-1)^{K-1}}. \tag{23}$$

Relating KL decay and norm growth. Define the normalized margins $\hat{x}_i = (\Psi \hat{a})_i = \frac{x_i}{r}$ where $r := r(t) := \|a(t)\|_1$. From Eq. (21) we have

$$\frac{d}{dt}(D_{\mathrm{KL}}) = \frac{1}{2(K-1)D} \sum_{ij} (\hat{x}_i - \hat{x}_j) \cdot (e^{-r\hat{x}_i} - e^{-r\hat{x}_j}). \tag{24}$$

We will lower bound the rate at which D_{KL} decreases with respect to the margin difference

$$m(t) = \max_i \hat{x}_i - \min_i \hat{x}_i \geq 0.$$

We have already shown that every summand in Eq. (24) is non-positive. Thus, keeping only the (i, j) pair corresponding to the $\max$ and $\min$ values of $\hat{x}_i$ we have

$$-\frac{d}{dt}(D_{\mathrm{KL}}) \geq \frac{1}{2(K-1)D} \cdot m(t) \cdot e^{-r(t) \cdot \min_i \hat{x}_i} \cdot (1 - e^{-r(t) \cdot m(t)}). \tag{25}$$

But, also,

$$\frac{\mathrm{d}r(t)}{\mathrm{d}t} = \sum_i \frac{\mathrm{d}a_i(t)}{\mathrm{d}t} = \frac{1}{D} \sum_i a_i(t) b_i(t) = \frac{r(t)}{D} \sum_i \hat{a}_i(t) b_i(t) = \frac{1}{D} \sum_j x_j e^{-x_j}, \tag{26}$$

where the last equality follows from Eq. (18). Now write

$$\sum_i x_i e^{-x_i} \leq r e^{-r(t)\min_i \hat{x}_i} \sum_i \hat{x}_i = Kr e^{-r(t)\min_i \hat{x}_i},$$

where the last equality used $K = \sum_i \hat{x}_i$ as in Eq. (19). Putting together the above two displays yields

$$\frac{\mathrm{d}r(t)}{\mathrm{d}t} \leq \frac{Kr}{D} e^{-r\min_i \hat{x}_i}. \tag{27}$$

Combining Eq. (25) with (27) yields

$$\begin{aligned} -\frac{d}{dt}(D_{\mathrm{KL}}) &\geq \frac{1}{2K(K-1)} \cdot m(t) \cdot (1 - e^{-r(t)\cdot m(t)}) \cdot \frac{1}{r(t)} \frac{\mathrm{d}r(t)}{\mathrm{d}t} \\ &\geq \frac{1}{2K(K-1)} \cdot \frac{\big(m(t)\big)^2}{1 + r(t)\cdot m(t)} \cdot \frac{\mathrm{d}r(t)}{\mathrm{d}t} \\ &\geq \frac{1}{2K(K-1)} \cdot \big(m(t)\big)^2 \cdot \frac{1}{1+2r(t)} \frac{\mathrm{d}r(t)}{\mathrm{d}t} \end{aligned} \tag{28}$$

where in the second line we use $1 - e^{-x} \geq \frac{x}{1+x}$ for all $x \geq 0$ and in the last line we use $m(t) \leq 2$. The latter is true since $(\Psi)_{ij} \in \{0, 2\}$ and $\hat{a}$ is a probability vector, thus $\hat{x}_i \in [0, 2]$.

Logit norm diverges. Since $\Psi_{ij} \in \{0, 2\}$ and $a \geq 0$, we have $x_j = (\Psi a)_j \leq 2r$. Moreover, because $x_j \geq 0$, it holds $D = 1 + \sum_j e^{-x_j} \leq K$. When combined with Eq. (26) these give

$$\frac{\mathrm{d}r}{\mathrm{d}t} \geq \frac{e^{-2r(t)}}{K} \sum_j x_j = r(t) \cdot e^{-2r(t)}, \tag{29}$$

where the equality uses $\sum_j x_j = 1^\top \Psi a = K 1^\top a = Kr$. In particular, using Eq. (29) we have $\frac{\mathrm{d}e^{2r(t)}}{\mathrm{d}t} = 2\frac{\mathrm{d}r(t)}{\mathrm{d}t} e^{2r(t)} \geq 2r(t) \geq 2r(0)$, so

$$r(t) \;\geq\; \tfrac{1}{2}\log\left(e^{2r(0)} + 2r(0)\cdot t\right) = r(0) + \frac{1}{2}\log(1 + 2r(0)e^{-2r(0)}\cdot t) \;\to\; \infty\,. \tag{30}$$

KL lower bounds the margin difference. We will now prove that for all $t \geq 0$:

$$\big(m(t)\big)^2 \geq \delta \cdot D_{\mathrm{KL}}(t)\,. \tag{31}$$

For this, we rewrite the margin difference as

$$m(t) = \max_i (\Psi v)_i - \min_i (\Psi v)_i,$$

where we set $v := v(t) := \hat{a}(t) - \hat{u}$ with $\hat{u} := \frac{1}{K-1} 1_{K-1}$. This is true because $\Psi\hat{u} = \frac{K}{K-1} 1_{K-1}$.

Moreover, $1_{K-1}^T \Psi v = K 1_{K-1}^T v = 0$, hence the entries of Ψv sum to zero, so that $\max_i (\Psi v)_i \geq 0 \geq \min_i (\Psi v)_i$ and therefore $m(t) \geq \|\Psi v\|_\infty$. Combining this with $\|\Psi v\|_2^2 = v^T \Psi^2 v = K\|v\|_2^2$, where we used $\Psi^2 = K I_{K-1} + K 1 1^T$ and $1^T v = 0$, gives

$$m(t) \;\geq\; \|\Psi v\|_\infty \;\geq\; \tfrac{1}{\sqrt{K-1}} \|\Psi v\|_2 = \sqrt{\tfrac{K}{K-1}}\, \|v\|_2 \;\geq\; \|v\|_2\,.$$

But, using $\log(x) \leq x - 1$, $\sum_i v_i = 0$, and finally $\hat{a}_i \geq \delta$ from Eq. (23), yields

$$D_{\mathrm{KL}}(t) = \sum_i \hat{u}_i \log \frac{\hat{u}_i}{\hat{a}_i} \leq \sum_i \hat{u}_i \frac{\hat{u}_i - \hat{a}_i}{\hat{a}_i} = \sum_i (\hat{a}_i - v_i) \frac{-v_i}{\hat{a}_i} = \sum_i \frac{v_i^2}{\hat{a}_i} \leq \frac{1}{\delta} \|v\|_2^2\,.$$

The two displays combined show the desired Eq. (31).

Put things together. Using Eq. (31) in Eq. (28) we have shown for $D_{\mathrm{KL}}(t) > 0$ (If $D_{\mathrm{KL}}(t) = 0$ at some t then it stays zero by monotonicity.) that

$$-\frac{\mathrm{dlog}(D_{\mathrm{KL}}(t))}{\mathrm{d}t} \geq \frac{\delta}{4K(K-1)} \cdot \frac{d}{dt}\left[\log(1+2r(t))\right]. \tag{32}$$

Integrating this from 0 to t, and using the expression for δ given in Eq. (23), yields

$$D_{\mathrm{KL}}(t) \;\leq\; D_{\mathrm{KL}}(0)\left(\frac{1+2r(0)}{1+2r(t)}\right)^{\rho}, \qquad \rho := \frac{\exp(-(K-1)D_{\mathrm{KL}}(0))}{4K(K-1)^K} > 0\,. \tag{33}$$

Since $r(t) \to \infty$, the right-hand side vanishes and therefore $D_{\mathrm{KL}}(t) \to 0$.

G.4 Proof of Theorem 4

Recall our choice of metric is given by:

$$M = \frac{1}{2}\sum_{i=1}^{K-1}\left(\hat{a}_i - \frac{1}{K-1}\right)^2.$$

A simple calculation, using the derivatives provided in Eq. (10), gives

$$\frac{dM}{dt} = \frac{1}{D}\sum_i \hat{a}_i^2[b_i - \bar{b}]\,. \tag{34}$$

By Theorem 3, $M(t) \to 0$ for every positive initialization. It therefore remains to establish monotonicity for $K = 2, 4$, and to construct counterexamples for $K \geq 8$.

First note that $M = 0$ in the $K = 2$ case for all positive initializations, and so trivially is monotonic. We begin by proving monotonicity for $K = 4$.

Note in $K = 4$ we have

$$\Psi = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & 2 \\ 2 & 2 & 0 \end{bmatrix} \implies \Psi a = \begin{bmatrix} 2a_1 + 2a_3 \\ 2a_2 + 2a_3 \\ 2a_1 + 2a_2 \end{bmatrix} \implies \exp(-\Psi a) = \begin{bmatrix} e^{-2a_1-2a_3} \\ e^{-2a_2-2a_3} \\ e^{-2a_1-2a_2} \end{bmatrix}.$$

Hence

$$\frac{1}{D}b = \frac{2}{1 + e^{-2a_1-2a_3} + e^{-2a_2-2a_3} + e^{-2a_1-2a_2}} \begin{bmatrix} e^{-2a_1-2a_3} + e^{-2a_1-2a_2} \\ e^{-2a_2-2a_3} + e^{-2a_1-2a_2} \\ e^{-2a_1-2a_3} + e^{-2a_2-2a_3} \end{bmatrix}.$$

Multiplying numerator and denominator by $\exp(2\sum_i a_i)$ gives

$$\frac{1}{D}b = \frac{2}{e^{2(a_1+a_2+a_3)} + e^{2a_1} + e^{2a_2} + e^{2a_3}} \begin{bmatrix} e^{2a_2} + e^{2a_3} \\ e^{2a_1} + e^{2a_3} \\ e^{2a_1} + e^{2a_2} \end{bmatrix}.$$

Writing $x_i = \exp(2a_i)$, this is then

$$\frac{1}{D}b = \frac{2}{x_1x_2x_3 + \sum_i x_i} \begin{bmatrix} \sum_i x_i - x_1 \\ \sum_i x_i - x_2 \\ \sum_i x_i - x_3 \end{bmatrix},$$

and then $\bar{b}$ is given by:

$$\frac{1}{D}\bar{b} = \frac{2}{x_1x_2x_3 + \sum_i x_i}[\sum_i x_i - \hat{a}\cdot x].$$

Hence using the expression for the derivative of our metric, given in Eq. (34), we arrive at:

$$\frac{dM}{dt} = \frac{2}{x_1x_2x_3 + \sum_i x_i}\sum_i \hat{a}_i^2[\hat{a}\cdot x - x_i].$$

The coefficient is strictly positive, and so to argue non-increasing we can drop it. Hence we simply need to show:

$$(\hat{a} \cdot x) \left(\sum_i \hat{a}_i^2 \right) - \sum_i \hat{a}_i^2 x_i \leq 0,$$

or equivalently:

$$\sum_i \hat{a}_i^2 \leq \frac{\sum_i \hat{a}_i^2 x_i}{\sum_i \hat{a}_i x_i}.$$

Define the quantities:

$$q_i = \frac{\hat{a}_i x_i}{\sum_i \hat{a}_i x_i},$$

noting that they form a probability distribution. Then our inequality can be viewed as comparing the expectation of $\hat{a}$ over two different probability distributions, one with probabilities $\hat{a}$ and the other with probabilities q.

It is then simple to note that, since for any $\|a\|_1 > 0$, x_i is an increasing function of $\hat{a}_i$, the second distribution places more weight on larger values of $\hat{a}$, and so must be at least as large as the other expectation. They are only equal when there are no larger values of $\hat{a}_i$, meaning $\hat{a} = \frac{1}{K-1} 1_{K-1}$. Hence the inequality is satisfied, and the metric is monotonically decreasing.

It remains to show that the metric is not necessarily monotonically decreasing for $K \geq 8$. To show this for $K = 8$, it is sufficient to note that for $a = [1, 1, 1, 1.25, 0.01, 0.01, 0.01]$ the derivative of our metric takes value $\frac{dM}{dt} \approx 7.3 \times 10^{-4}$, and so the metric is increasing for these values. Call this specific counterexample v^*, and denote its corresponding b vector and its weighted average by b^* and $\bar{b}^*$.

For $K > 8$, define our initialization as $a = [v^*, 0_{K-8}]$, then the derivative of our metric becomes

$$\frac{dM}{dt} = \frac{1}{D} \sum_{i=1}^{7} (\hat{v}_i^*)^2 [b_i - \bar{b}], \qquad \text{where } \bar{b} = \sum_{i=1}^{7} v_i^* b_i.$$

Note that only $b_1, ..., b_7$ are necessary for the computation of the sign. We now use the lemma stated and proven at the end of this section, which gives that when a is only non-zero on at most the first $\frac{K}{2} - 1$ entries, then the first $\frac{K}{2} - 1$ entries of the b vector are just equal to double the values they would be in the $\frac{K}{2}$ dimensional case. Applying this iteratively, we get that if $K = 2^m > 8$, then

$$b_i = 2^{m-3} b_i^*, \qquad \bar{b} = 2^{m-3} \bar{b}^*, \qquad \text{for } i = 1, ..., 7,$$

and so

$$\frac{dM}{dt} = \frac{2^{m-3}}{D} \sum_{i=1}^{7} (\hat{v}_i^*)^2 [b_i^* - \bar{b}^*],$$

but this clearly has the same sign as the counterexample in $K = 8$, and so gives a positive derivative.

This provides a counterexample in any K, but it has entries that are zero, which is not permitted. However if we consider the vector

$$a = [v^*, \epsilon 1_{K-8}],$$

then by continuity, as $\epsilon \to 0$ this tends to the positive value achieved by $a = [v^*, 0_{K-8}]$. Hence for any K there exists a non-zero value of ϵ such that $\frac{dM}{dt} > 0$ for $a = [v^*, \epsilon 1_{K-8}]$. Hence counterexamples exist in all $K \geq 8$.

It only remains to prove the following lemma.

Lemma 1. *Let $K = 2^m$, and $a^* \in \mathbb{R}^{K-1}$, with $b(a^*) = b^* \in \mathbb{R}^{K-1}$ and $\bar{b}^* = \sum_i \hat{a}_i^* b_i^* \in \mathbb{R}$. Let $a' \in \mathbb{R}^{2K-1}$ be given by $a' = [a^*, 0_K]$, and consider the evolution under the dynamics detailed in Eq.* (9)*, but for dimension $2K$, initialized at a'. Denote $b(a') = b' \in \mathbb{R}^{2K-1}$ and $\bar{b}(a') = \bar{b}' \in \mathbb{R}$. Then for $i = 1, ..., K-1$, we have*

$$b_i' = 2b_i^*, \qquad \bar{b}' = 2\bar{b}^*.$$

Proof.

Let Ψ_{K-1} and Ψ_{2K-1} denote the two matrices derived from the Hadamard matrix, as defined in Sec. 4, with the subscript denoting their dimension. Using the construction in terms of Sylvester Hadamard matrices, as defined in Definition 1, these are related in the following way:

$$\Psi_{2K-1} = \begin{bmatrix} \Psi_{K-1} & 0_{K-1} & \Psi_{K-1} \\ 0_{K-1}^T & 2 & 21_{K-1}^T \\ \Psi_{K-1} & 21_{K-1} & 21_{K-1}1_{K-1}^T - \Psi_{K-1} \end{bmatrix}.$$

Hence

$$\Psi_{2K-1}a' = \begin{bmatrix} \Psi_{K-1} & 0_{K-1} & \Psi_{K-1} \\ 0_{K-1}^T & 2 & 21_{K-1}^T \\ \Psi_{K-1} & 21_{K-1} & 21_{K-1}1_{K-1}^T - \Psi_{K-1} \end{bmatrix} \begin{bmatrix} a^* \\ 0 \\ 0_{K-1} \end{bmatrix} = \begin{bmatrix} \Psi_{K-1}a^* \\ 0 \\ \Psi_{K-1}a^* \end{bmatrix},$$

and so

$$\exp(-\Psi_{2K-1}a') = \begin{bmatrix} \exp(-\Psi_{K-1}a^*) \\ 1 \\ \exp(-\Psi_{K-1}a^*) \end{bmatrix},$$

giving

$$b' = \Psi_{2K-1}\exp(-\Psi_{2K-1}a') = \begin{bmatrix} \Psi_{K-1} & 0_{K-1} & \Psi_{K-1} \\ 0_{K-1}^T & 2 & 21_{K-1}^T \\ \Psi_{K-1} & 21_{K-1} & 21_{K-1}1_{K-1}^T - \Psi_{K-1} \end{bmatrix} \begin{bmatrix} \exp(-\Psi_{K-1}a^*) \\ 1 \\ \exp(-\Psi_{K-1}a^*) \end{bmatrix}.$$

From this it is clear that the top $K-1$ entries of b' are given by the vector $2\Psi_{K-1}\exp(-\Psi_{K-1}a) = 2b^*$. Also note

$$\bar{b}' = \sum_j \hat{a}'_j b'_j = \sum_{j=1}^{K-1} \hat{a}^*_j b'_j = 2\bar{b}^*.$$

G.5 Proof of Theorem 5

Recall our metric M, given by:

$$M = \frac{1}{2}\sum_{i=1}^{K-1}\left(\frac{a_i}{\|a\|_1} - \frac{1}{K-1}\right)^2.$$

Throughout the proof we make use of the following properties of $\Psi = 1_{K-1}1_{K-1}^\top - X$:

$$\begin{aligned} \Psi\, 1_{K-1} &= K\, 1_{K-1}, \\ \Psi^2 &= K\, I_{K-1} + K\, 1_{K-1}1_{K-1}^\top, \\ |\Psi_{ij}| &\le 2, \qquad \forall\, i,j, \\ \|\psi_u\|_2 &= \sqrt{2K}, \qquad \forall \text{ columns } \psi_u \text{ of } \Psi. \end{aligned}$$

Additionally, unless otherwise stated, all indices run in $1,\dots,K-1$ and are omitted to simplify notation.

We choose to write our singular modes in the following form:

$$a_i = \mu + \mu\epsilon_i, \qquad \sum_i \epsilon_i = 0, \qquad \mu = \frac{1}{K-1}\sum_i a_i\,.$$

In this form M is given by

$$M = \frac{1}{2(K-1)^2}\sum_i \epsilon_i^2\,.$$

Recall its derivative as given in Eq. (34):

$$\frac{dM}{dt} = \frac{1}{D}\sum_i \hat{a}_i^2[b_i - \bar{b}], \qquad \text{where } \bar{b} = \sum_j \hat{a}_j b_j\,.$$

This we can write alternatively as

$$\frac{dM}{dt} = \frac{1}{D}\left[\left(\sum_i \hat{a}_i^2 b_i\right) - \sum_{ij} \hat{a}_i^2 \hat{a}_j b_j\right] = \frac{1}{D\|a\|_1^3}\left[\left(\sum_i a_i^2 b_i\right)\left(\sum_j a_j\right) - \sum_{ij} a_i^2 a_j b_j\right]$$
$$= \frac{1}{D\|a\|_1^3}\sum_{ij} a_i^2 a_j (b_i - b_j)\,.$$

Now writing this in terms of ϵ and μ gives

$$\frac{dM}{dt} = \frac{1}{D\|a\|_1^3}\sum_{iju}(\mu+\mu\epsilon_i)^2(\mu+\mu\epsilon_j)[\Psi_{iu}-\Psi_{ju}]e^{-\mu(\Psi[1_{K-1}+\epsilon])_u}$$
$$= \frac{e^{-K\mu}}{D(K-1)^3}\sum_{iju}(1+\epsilon_i)^2(1+\epsilon_j)[\Psi_{iu}-\Psi_{ju}]e^{-\mu(\Psi\epsilon)_u}\,,$$

where we used that $\mu = \frac{1}{K-1}\|a\|_1$, and $\Psi 1_{K-1} = K1_{K-1}$.

Expand the exponential as:

$$\exp(-\mu(\Psi\epsilon)_u) = 1 - \mu(\Psi\epsilon)_u + \sum_{n=2}^{\infty}\frac{1}{n!}\left[-\mu(\Psi\epsilon)_u\right]^n\,.$$

Using $\sum_u[\Psi_{iu}-\Psi_{ju}] = 0$, the first term in $\frac{dM}{dt}$ gives

$$\sum_{iju}(1+\epsilon_i)^2(1+\epsilon_j)[\Psi_{iu}-\Psi_{ju}] = 0\,.$$

Further using that $\Psi^2 = KI_{K-1} + K1_{K-1}1_{K-1}^T$, the second term becomes

$$-\mu\sum_{iju}(1+\epsilon_i)^2(1+\epsilon_j)[\Psi_{iu}-\Psi_{ju}](\Psi\epsilon)_u = -\mu\sum_{ij}(1+\epsilon_i)^2(1+\epsilon_j)[(\Psi^2\epsilon)_i - (\Psi^2\epsilon)_j]$$
$$= -K\mu\sum_{ij}(1+\epsilon_i)^2(1+\epsilon_j)[\epsilon_i-\epsilon_j]$$
$$= -K\mu\big[\sum_i(1+\epsilon_i)^2\epsilon_i\sum_j(1+\epsilon_j) - \sum_i(1+\epsilon_i)^2\sum_j(1+\epsilon_j)\epsilon_j\big]$$
$$= -K\mu\left[(K-1)\sum_i\epsilon_i^2 + (K-1)\sum_i\epsilon_i^3 - \left(\sum_i\epsilon_i^2\right)^2\right],$$

where in the last line we used $\sum_i \epsilon_i = 0$.

Again because $\sum_i \epsilon_i = 0$ and $\sum_u \Psi_{iu} = K$, we can also simplify the remainder term by using

$$\sum_{ij}(1+\epsilon_i)^2(1+\epsilon_j)(\Psi_{iu}-\Psi_{ju}) = \sum_i[(K-1)\Psi_{iu}\epsilon_i + (K-1)\Psi_{iu}\epsilon_i^2 - K\epsilon_i^2 - \epsilon_i^2(\Psi\epsilon)_u]\,.$$

Putting all the above together, our remaining expression for the full derivative is given by

$$\frac{dM}{dt} = -\frac{K\mu e^{-K\mu}}{D(K-1)^2}\sum_i\epsilon_i^2 - \frac{K\mu e^{-K\mu}}{D(K-1)^2}\sum_i\epsilon_i^3 + \frac{K\mu e^{-K\mu}}{D(K-1)^3}\left(\sum_i\epsilon_i^2\right)^2$$
$$+\frac{e^{-K\mu}}{D(K-1)^3}\sum_{iu}\left[(K-1)\Psi_{iu}\epsilon_i + (K-1)\Psi_{iu}\epsilon_i^2\underbrace{-K\epsilon_i^2 - \epsilon_i^2(\Psi\epsilon)_u}_{(*)}\right]\left[\sum_{n=2}^{\infty}\frac{1}{n!}[-\mu(\Psi\epsilon)_u]^n\right].$$

We want to show that when M is small enough, this quantity is negative. First let us assume that $|(\Psi\epsilon)_u| \le K$, in which case we can drop the two terms highlighted as $(*)$, since they contribute a

non-positive value, where we used that $e^x - 1 - x \geq 0$. Hence, also dropping a positive multiplicative constant, it remains to show

$$0 \geq -K\mu \sum_i \epsilon_i^2 - K\mu \sum_i \epsilon_i^3 + \frac{K\mu}{K-1} \left(\sum_i \epsilon_i^2 \right)^2 + \sum_{iu} (\Psi_{iu}\epsilon_i + \Psi_{iu}\epsilon_i^2) \left[\sum_{n=2}^{\infty} \frac{1}{n!} [-\mu(\Psi\epsilon)_u]^n \right] .$$

Now we use the following properties to bound the remaining terms:

$$\begin{aligned} \sum_i \epsilon_i^2 &= 2(K-1)^2 M, \\ |\sum_i \epsilon_i^3| \leq \left(\sum_i \epsilon_i^2 \right)^{\frac{3}{2}} &= 2^{\frac{3}{2}} (K-1)^3 M^{\frac{3}{2}}, \\ \sum_i \Psi_{iu} \epsilon_i^2 \leq 2 \sum_i \epsilon_i^2 &= 4(K-1)^2 M, \\ \sum_i \Psi_{iu} \epsilon_i = (\Psi\epsilon)_u \leq |(\Psi\epsilon)_u| \leq \sqrt{(\Psi\epsilon)\cdot(\Psi\epsilon)} &= \sqrt{2K}(K-1)\sqrt{M}. \end{aligned} \tag{35}$$

Note the second property follows from monotonicity of L_p norms, the third uses $\Psi_{iu} \in \{0, 2\}$ for all i, u, and, the last one $\|\psi_u\| = \sqrt{2K}$ for all columns/rows ψ_u of Ψ. This produces the following sufficient condition for a negative derivative:

$$\begin{aligned} 0 \geq &-2K\mu(K-1)^2 M + 2^{\frac{3}{2}} K\mu(K-1)^3 M^{\frac{3}{2}} + 4K\mu(K-1)^3 M^2 \\ &+ \sum_u \left[\sqrt{2K}(K-1)\sqrt{M} + 4(K-1)^2 M \right] \underbrace{\left[\sum_{n=2}^{\infty} \frac{1}{n!} [-\mu(\Psi\epsilon)_u]^n \right]}_{R_u} . \end{aligned}$$

We can bound the quantity R_u using $e^x - 1 - x \leq \frac{1}{2} x^2 e^{|x|}$, giving

$$R_u \leq \frac{1}{2} \mu^2 (\Psi\epsilon)_u^2 \exp(|\mu(\Psi\epsilon)_u|) \leq K\mu^2 (K-1)^2 M \exp\left(\mu\sqrt{2K}(K-1)\sqrt{M} \right) ,$$

where in the second step we used the bound for $(\Psi\epsilon)_u$ from Eq. (35). Hence, it is sufficient to show

$$\begin{aligned} 0 \geq &-2K\mu(K-1)^2 M + 2^{\frac{3}{2}} K\mu(K-1)^3 M^{\frac{3}{2}} + 4K\mu(K-1)^3 M^2 \\ &+ \left[\sqrt{2K}(K-1)^2\sqrt{M} + 4(K-1)^3 M \right] K\mu^2 (K-1)^2 M \exp\left(\mu\sqrt{2K}(K-1)\sqrt{M} \right) , \end{aligned}$$

where we also picked up an extra factor of $(K-1)$ due to the summation over u.

Now, dropping a factor of $2K(K-1)^2\mu M$, this reduces to

$$\begin{aligned} 0 \geq &-1 + 2^{\frac{1}{2}} (K-1) M^{\frac{1}{2}} + 2(K-1)M \\ &+ \left[\sqrt{\frac{K}{2}} (K-1)^2 \sqrt{M} + 2(K-1)^3 M \right] \mu \exp\left(\mu\sqrt{2K}(K-1)\sqrt{M} \right) . \end{aligned}$$

Letting $\tilde{M} = 2KM$ and $\tilde{\mu} = (K-1)\mu$, this further reduces to

$$0 \geq -1 + \frac{K-1}{\sqrt{K}} \tilde{M}^{\frac{1}{2}} + \frac{K-1}{K} \tilde{M} + \left[\frac{1}{2} (K-1) \tilde{M}^{\frac{1}{2}} + \frac{(K-1)^2}{K} \tilde{M} \right] \tilde{\mu} e^{\tilde{\mu}\tilde{M}^{\frac{1}{2}}} ,$$

and finally using that $K - 1 \leq K$ it is sufficient to show

$$1 \geq \sqrt{K} \tilde{M}^{\frac{1}{2}} + \tilde{M} + \left[\frac{1}{2} \tilde{M}^{\frac{1}{2}} + \tilde{M} \right] K\tilde{\mu} e^{\tilde{\mu}\tilde{M}^{\frac{1}{2}}} .$$

Now assume $\tilde{M}^{\frac{1}{2}} \leq \min\left(\frac{1}{2\sqrt{K}}, \frac{1}{3K\tilde{\mu}}\right)$, then we have:

$$\begin{aligned}\sqrt{K}\tilde{M}^{\frac{1}{2}} + \tilde{M} + \left[\frac{1}{2}\tilde{M}^{\frac{1}{2}} + \tilde{M}\right] K\tilde{\mu}e^{\tilde{\mu}\tilde{M}^{\frac{1}{2}}} &\leq \frac{1}{2} + \frac{1}{4K} + \left[\frac{1}{2} + \frac{1}{2\sqrt{K}}\right]\frac{1}{3}e^{\frac{1}{3K}} \\ &\leq \frac{1}{2} + \frac{1}{8} + \left[\frac{1}{2} + \frac{1}{2\sqrt{2}}\right]\frac{1}{3}e^{\frac{1}{6}} \approx 0.961\,,\end{aligned}$$

where we used that $K \geq 2$. This quantity is less than one, and hence $\frac{dM}{dt}$ is negative subject to this constraint on $\tilde{M}$. Stated in terms of M and $\|a\|_1$, this constraint is

$$M \leq \min\left(\frac{1}{8K^2}, \frac{1}{18K^3\|a\|_1^2}\right).$$

Recall that we also used $|(\Psi\epsilon)_u| \leq K$, previously, but this is a weaker bound than the above (by Eq. (35)) and so holds subject to our condition. Hence we have shown that

$$\frac{dM}{dt} \leq 0, \quad \text{when } M \leq \min\left(\frac{1}{8K^2}, \frac{1}{18K^3\|a\|_1^2}\right).$$

G.6 Proof of Theorem 6

We now consider randomly initialized matrices for the UFM model of Eq. (2). Assume that W, H are initialized as $W(0) = \epsilon d^{-\frac{1}{2}}A_1$, $H(0) = \epsilon d^{-\frac{1}{2}}A_0$, so that the logit matrix is given by $Z(0) = \epsilon^2 d^{-1}A_1A_0$.

We define the quantity

$$t^* = \frac{1}{2\sqrt{2Kn}}\log\left(\frac{1}{\epsilon}\right).$$

We first show that on the time interval $[0, t^*]$, the logits must remain $O(\epsilon)$. First define the quantity

$$\Gamma(t) = \|W(t)\|_F + \|H(t)\|_F,$$

then we have

$$\frac{d}{dt}\Gamma \leq \left\|\frac{d}{dt}W\right\|_F + \left\|\frac{d}{dt}H\right\|_F \leq \|Y - P(Z)\|_F\,\Gamma.$$

Since Y and $P(Z)$ are probability matrices, we must have

$$\|Y - P(Z)\|_F \leq \sqrt{2Kn},$$

hence

$$\Gamma(t) \leq \Gamma(0)e^{t\sqrt{2Kn}},$$

and so

$$\|Z(t)\|_F \leq \Gamma^2(0)e^{2t\sqrt{2Kn}}.$$

Now note that $\Gamma(0) = C_0\epsilon$, where $C_0 = d^{-\frac{1}{2}}(\|A_1\|_F + \|A_0\|_F)$, and so for the time interval $t \in [0, t^*]$ we must have

$$\|Z(t)\|_F \leq C_0^2\epsilon,$$

and so on this time domain we have that $Z(t)$ remains $O(\epsilon)$.

Hence for this time window, we have by Taylor expanding that the softmax is given by:

$$P(t) = \frac{1}{K}1_K 1_{Kn}^T + P'(t, \epsilon),$$

where $P'(t, \epsilon) = O(\epsilon)$. From the gradient flow equations for this model, described in Eq. (3), we have for $t \in [0, t^*]$

$$\frac{dW}{dt} = (S \otimes 1_n^T)H^T - P'H^T, \qquad \frac{dH^T}{dt} = (S \otimes 1_n^T)^T W - (P')^T W.$$

We can reduce this to a second order ODE

$$\frac{d^2W}{dt^2} = nSW + \underbrace{\frac{d}{dt}(-P'H^T) - (S \otimes 1_n^T)(P')^T W}_{O(\epsilon^2)},$$

where we have the boundary conditions $W(0) = \epsilon d^{-\frac{1}{2}} A_1$, $\frac{dW}{dt}(0) = \epsilon d^{-\frac{1}{2}}(S \otimes 1_n^T)A_0^T + O(\epsilon^2)$. We also note that the terms that feature P' are $O(\epsilon^2)$ on the considered time interval, so should be viewed as perturbative.

This has solution given by

$$W(t) = \epsilon d^{-\frac{1}{2}} \left[\frac{1}{K} 1_K 1_K^T A_1 + \cosh(t\sqrt{n}) S A_1 + \frac{1}{\sqrt{n}} \sinh(t\sqrt{n})(S \otimes 1_n^T) A_0^T \right] + O(\epsilon^2 t^2),$$

and corresponding solution for H given by

$$H^T(t) = \epsilon d^{-\frac{1}{2}} \Bigg[(I_{Kn} - \frac{1}{n} S \otimes 1_n 1_n^T) A_0^T + \frac{1}{\sqrt{n}} \sinh(t\sqrt{n})(S \otimes 1_n) A_1 + \frac{1}{n} \cosh(t\sqrt{n})(S \otimes 1_n 1_n^T) A_0^T \Bigg] + O(\epsilon^2 t^2).$$

We now want to compute $Z(t^*) = W(t^*)H(t^*)$. In this expansion we recover terms of a variety of different orders in t and ϵ. Recalling our expression for t^*, this diverges in the $\epsilon \to 0$ limit, but (as showed earlier) slowly enough that the logit matrix norm goes to zero, meaning our perturbative terms remain small. By considering the terms in $W(t)$ and $H(t)$, it becomes clear that the leading order term in $Z(t^*) = W(t^*)H(t^*)$ comes from the product of the terms in W, H that are exponential in t^*, giving

$$Z(t^*) = \frac{1}{4d} \epsilon^2 e^{2t^*\sqrt{n}} \left[S A_1 + \frac{1}{\sqrt{n}} (S \otimes 1_n^T) A_0^T \right] \left[\frac{1}{\sqrt{n}} (S \otimes 1_n) A_1 + \frac{1}{n} (S \otimes 1_n 1_n^T) A_0^T \right]^T + \dots,$$

where by $\dots$ we mean terms that vanish faster than leading order. Continuing

$$Z(t^*) = \frac{1}{4d} \epsilon^2 e^{2t^*\sqrt{n}} S \left[A_1 + \frac{1}{\sqrt{n}} (I_K \otimes 1_n^T) A_0^T \right] \left[A_1 + \frac{1}{\sqrt{n}} (I_K \otimes 1_n^T) A_0^T \right]^T (S \otimes \frac{1}{\sqrt{n}} 1_n^T) + \dots.$$

Now define the matrix $X = \frac{1}{\sqrt{2}}[A_1 + (I_K \otimes \frac{1}{\sqrt{n}} 1_n^T) A_0^T]$, this gives

$$Z(t^*) = \frac{1}{2d} \epsilon^2 e^{2t^*\sqrt{n}} S X X^T (S \otimes \frac{1}{\sqrt{n}} 1_n^T) + \dots.$$

Now suppose that the entries of $A_1 \in \mathbb{R}^{K \times d}$ and $A_0^T \in \mathbb{R}^{Kn \times d}$ are initialized with i.i.d. samples from a standard Gaussian. Then the entries of the matrix $\frac{1}{\sqrt{n}}(I_K \otimes 1_n^T) A_0^T \in \mathbb{R}^{K \times d}$ are also i.i.d. samples of a standard Gaussian, and hence the matrix $X \in \mathbb{R}^{K \times d}$ also has entries i.i.d. sampled from a standard Gaussian.

Considering the entries of the matrix $\frac{1}{d} X X^T$ in the large d limit, we find:

$$i = j: \ \left[\frac{1}{d} (X X^T) \right]_{ii} = \frac{1}{d} \sum_u X_{iu} X_{iu} \to \mathbb{E}[N_1^2] = 1,$$

$$i \neq j: \ \left[\frac{1}{d} (X X^T) \right]_{ij} = \frac{1}{d} \sum_u X_{iu} X_{ju} \to \mathbb{E}[N_1 N_2] = \mathbb{E}[N_1]\mathbb{E}[N_2] = 0.$$

Where $N_1, N_2 \sim \mathcal{N}(0, 1)$.

Hence in the large d limit, we have the following convergence almost surely $\frac{1}{d} X X^T \to I_K$, and we find that

$$\frac{1}{d} S X X^T (S \otimes \frac{1}{\sqrt{n}} 1_n^T) \to \frac{1}{\sqrt{n}} (S \otimes 1_n^T),$$

and so $Z(t^*)$ is well approximated as

$$Z(t^*) = \frac{1}{2\sqrt{n}} \epsilon^2 e^{2t^*\sqrt{n}} S \otimes 1_n^T + \dots,$$

where the higher order terms decay faster in the described limits, and so this gives the asymptotic leading term.

G.7 Proof of Theorem 7

Recall our dynamics are given by:

$$\frac{da_i}{dt} = \frac{1}{D} a_i b_i, \qquad \text{where } b_i = \sum_j \Psi_{ij} e^{-(\Psi a)_j}, \quad D = 1 + \sum_j e^{-(\Psi a)_j}.$$

Suppose that $\|a\|_1 = \epsilon \ll 1$. Then expanding in ϵ gives $e^{-(\Psi a)_j} = 1 - (\Psi a)_j + O(\epsilon^2)$. Also recall that $\sum_j \Psi_{ij} = K$. We hence have:

$$\begin{aligned} D &= 1 + \sum_j [1 - (\Psi a)_j] + O(\epsilon^2) \\ &= K - K \sum_j a_j + O(\epsilon^2) \end{aligned}$$

and hence:

$$D = K(1 - \epsilon) + O(\epsilon^2).$$

Similarly, we have

$$\begin{aligned} b_i &= \sum_j \Psi_{ij} (1 - (\Psi a)_j) + O(\epsilon^2) \\ &= K - (\Psi^2 a)_i + O(\epsilon^2). \end{aligned}$$

Using the properties of the underlying Hadamard matrix, we have $\Psi^2 = K(I + 1_{K-1} 1_{K-1}^T)$, giving

$$b_i = K - K a_i - K\epsilon + O(\epsilon^2).$$

Hence, our differential equation for a_i is given by:

$$\begin{aligned} \frac{da_i}{dt} &= a_i \left[K - K a_i - K\epsilon + O(\epsilon^2)\right] \left[K(1 - \epsilon) + O(\epsilon^2)\right]^{-1} \\ &= a_i \left[1 - a_i - \epsilon + O(\epsilon^2)\right] \left[1 - \epsilon + O(\epsilon^2)\right]^{-1} \\ &= a_i \left[1 - a_i - \epsilon + O(\epsilon^2)\right] \left[1 + \epsilon + O(\epsilon^2)\right] \\ &= a_i (1 - a_i) + O(\epsilon^3). \end{aligned}$$

Hence, we see that the linearized-softmax dynamics do decouple. This equation is simple to solve, and gives a sigmoid solution, specifically:

$$a_i(t) = \frac{a_i(0)}{a_i(0) + (1 - a_i(0)) e^{-t}}.$$

This solution is only valid as an approximation in the linearized regime.

G.8 Proof of Theorem 8

When optimizing the logits directly, as detailed in Eq. (12), the gradient flow equations are given by:

$$\frac{dZ}{dt} = Y - P.$$

We initialize our matrix Z as $Z = U D_Z V^T$, where $U = \frac{1}{\sqrt{K}} \Phi$, and $V = U \otimes Q$, where Q is a right singular-vector matrix of 1_n^T, meaning it obeys $1_n^T = \sqrt{n} e_1^{(n)T} Q^T$, where $e_1^{(n)}$ is the first standard basis vector in $\mathbb{R}^n$. Also recall D_Z has its only non-zero singular values given by $a_1, \ldots, a_K$. The exact same arguments used in the proof of Theorem 2 then give that, at initialization, Z can be written as

$$Z = (U D_{\tilde{a}} U^T) \otimes 1_n^T,$$

where $D_{\tilde{a}} = \text{diag}(\tilde{a}_1, \ldots, \tilde{a}_K)$, with $\tilde{a}_i = \frac{1}{\sqrt{n}} a_i$. Then Theorem 1 gives us that

$$P(Z) = (U D_{\tilde{\nu}} U^T) \otimes 1_n^T,$$

where

$$\tilde{\nu}_i = \frac{\sum_j \Phi_{ij} e^{\frac{1}{K}(\Phi \tilde{a})_j}}{\sum_j e^{\frac{1}{K}(\Phi \tilde{a})_j}}.$$

We can now turn to our gradient flow equation, and use the change of variable $Z = U\tilde{Z}V^T$, which evolves as:

$$\frac{d}{dt}\left(\tilde{Z}\right) = U^T(Y - P)V.$$

As in Theorem 2, define $B = U^T(Y-P)V$. Since we have $Y = \sqrt{n}UD_YV^T$ and $P = \sqrt{n}UD_PV^T$, where $D_Y = [I_K, 0_{K\times K(n-1)}] \in \mathbb{R}^{K\times Kn}$ and $D_P = [\text{diag}(\tilde{\nu}), 0_{K\times K(n-1)}] \in \mathbb{R}^{K\times Kn}$, we have:

$$B = \sqrt{n}[\text{diag}(1_K - \tilde{\nu}), 0_{K\times K(n-1)}].$$

Hence the evolution in $\tilde{Z}$ becomes:

$$\frac{d}{dt}\left(\tilde{Z}_{ij}\right) = \sqrt{n}(1 - \tilde{\nu}_i)\delta_{ij}.$$

We see that the only entries of $\tilde{Z}$ that evolve are the entries on its diagonal. Consequently if at initialization it takes the form of a singular value matrix, this property continues to hold for all future times. At initialization we have $\tilde{Z} = [\text{diag}(a), 0_{K\times K(n-1)}]$, and so these singular values evolve as:

$$\frac{d}{dt}(a_i) = \sqrt{n}(1 - \tilde{\nu}_i).$$

Translating to the original variable Z, this tells us that the singular vectors remain fixed, and all time evolution occurs in the singular values.

Just as in the proof of Theorem 2, using $\tilde{a} = \frac{1}{\sqrt{n}}a$ we can write:

$$(1 - \tilde{\nu}_i) = \frac{1}{\sum_j e^{-\frac{1}{K\sqrt{n}}([1_K 1_K^T - \Phi]a)_j}} \left[\sum_j [1_K 1_K^T - \Phi]_{ij} e^{-\frac{1}{K\sqrt{n}}([1_K 1_K^T - \Phi]a)_j}\right].$$

Note that this quantity is zero for $i = 1$, and so a_1 stays at its initial value. It also does not appear in the evolution equations of the other variables, and so we drop it and only consider the non-trivial modes, viewing $a \in \mathbb{R}^{K-1}$ as the vector containing only the non-trivial singular values.

The evolution equations are then:

$$\frac{d}{dt}(a_i) = \frac{\sqrt{n}}{D}b_i, \qquad \text{where } b_i = \sum_{j=1}^{K-1} \Psi_{ij} e^{-\frac{1}{K\sqrt{n}}(\Psi a)_j}, \quad D = 1 + \sum_{j=1}^{K-1} e^{-\frac{1}{K\sqrt{n}}(\Psi a)_j}.$$

Absorbing constants into a and t via $a' = \frac{1}{K\sqrt{n}}a$ and $t' = \frac{1}{K}t$, then dropping primes since these rescalings do not change the phenomenology of the dynamics, reduces us to

$$\frac{d}{dt}(a_i) = \frac{1}{D}b_i, \qquad \text{where } b_i = \sum_{j=1}^{K-1} \Psi_{ij} e^{-(\Psi a)_j}, \quad D = 1 + \sum_{j=1}^{K-1} e^{-(\Psi a)_j}.$$

If we define the variables $\hat{a}_i = a_i/\|a\|_1$, then these satisfy the evolution equations given by:

$$\frac{d}{dt}\left(\hat{a}_i\right) = \frac{1}{D\|a\|_1}\left[b_i - \hat{a}_i \sum_j b_j\right].$$

Again defining the metric M as

$$M = \frac{1}{2}\sum_i \left[\hat{a}_i - \frac{1}{K-1}\right]^2.$$

This has derivative:

$$\frac{dM}{dt} = \frac{1}{D\|a\|_1}\left[\left(\sum_i \hat{a}_i b_i\right) - \left(\sum_i \hat{a}_i^2\right)\left(\sum_j b_j\right)\right].$$

Now note that, as in the proof of Theorem 3, if we define $x_i = (\Psi a)_i$, we have:

$$\sum_i \hat{a}_i b_i = \frac{1}{\|a\|_1} \sum_j x_j e^{-x_j},$$

and as was shown in the proof, this satisfies:

$$\frac{1}{\|a\|_1} \sum_j x_j e^{-x_j} \leq \frac{1}{(K-1)\|a\|_1} \left(\sum_i x_i \right) \left(\sum_j e^{-x_j} \right),$$

with equality only when $x \propto 1_{K-1}$. Hence we have that the derivative of M satisfies:

$$\frac{dM}{dt} \leq \frac{1}{D\|a\|_1} \left[\frac{1}{(K-1)\|a\|_1} \left(\sum_i x_i \right) \left(\sum_j e^{-x_j} \right) - \left(\sum_i \hat{a}_i^2 \right) \left(\sum_j b_j \right) \right].$$

Then use that $\sum_i b_i = K \sum_i e^{-x_i}$, and $\sum_i x_i = K\|a\|_1$, reducing this to

$$\frac{dM}{dt} \leq \frac{K \sum_j e^{-x_j}}{D\|a\|_1} \left[\frac{1}{(K-1)} - \left(\sum_i \hat{a}_i^2 \right) \right].$$

Then note by Cauchy-Schwarz we have $\sum_i \hat{a}_i^2 \geq \frac{1}{K-1}$, with equality only when $\hat{a} \propto 1_{K-1}$, consequently we arrive at

$$\frac{dM}{dt} \leq 0,$$

with equality only when $x \propto 1_{K-1}$ and $\hat{a} \propto 1_{K-1}$. These both only occur when $\hat{a} = \frac{1}{K-1} 1_{K-1}$, as argued in the proof of Theorem 3.

Hence the metric M monotonically decreases in the direct logit optimization case.

H Further numerical experiments

H.1 Further details on figures in main body

H.1.1 Fig. 1

For Fig. 1, we use gradient descent with learning rate 0.01 and $1e7$ training steps, recorded at 1000 logarithmically spaced checkpoints. The initial logit singular values are sampled independently from the uniform distribution on $[0, 2]$. We then compare the evolution of the full matrix dynamics under Hadamard initialization with the reduced vector dynamics in Eq. (9). The two trajectories are initialized with the same singular values. The reduced vector ODE is solved using an adaptive Runge–Kutta method with relative tolerance 10^{-8} and absolute tolerance 10^{-10}; to compare with gradient descent, we evaluate it at continuous times $t_k = \eta k$, where k is the training step and $\eta = 0.01$ is the learning rate.

H.1.2 Fig. 3

For Fig. 3, we use $K = 8$ classes, one sample per class, and 10 random realizations. We train for $1e7$ gradient-descent steps with learning rate 0.01, using 5000 logarithmically spaced checkpoints. The random initialization is $W(0) = e^{-\delta} \frac{A_1}{\sqrt{K}}$, and $H(0) = e^{-\delta} \frac{A_0}{\sqrt{K}}$, where the entries of A_1 and A_0 are sampled independently from a standard Gaussian distribution and we here use $\delta = 12$. We also set $d = 5K$. The Hadamard initialization uses the same initial logit singular values as the corresponding random initialization, but places the singular vectors exactly in the Hadamard eigenspaces. Concretely, we initialize $W = UD_W R^T$ and $H = RD_H V^T$ as described in Theorem 1, $U = V = \frac{1}{\sqrt{K}}\Phi$. Recall from Sec. 4 that the singular value corresponding to the all 1s column of U is not updated by the flow; thus, as in theory, we assume without loss of generality that the respective entry of D_W, D_H is set to zero. To further ensure direct comparison to random initialization, we set the remaining $K - 1$ singular values such that $D_W = D_H$ and they are identical to the largest $K - 1$ singular values of

the logit matrix from its corresponding random initialization pair. This isolates the *initial* difference between the two schemes to only the singular vectors.

The residual energy plotted in Fig. 3(f) measures $\|Z - U(U^T ZV)V^T\|_F / \|Z\|_F$, for $U = V = \Phi(1 : K, 2 : K)$. We use residual threshold 0.005 to identify the end of the alignment phase. Specifically, we first wait until the residual drops below 0.005, and then define the "burn-in time" as the first logged time at which the residual rises above 0.005 again. This choice selects the point at which the random trajectory has already passed through its closest approach to the Hadamard subspace. The late-Hadamard trajectory is then initialized at this burn-in time by matching the singular values of the random trajectory and placing the singular vectors in the Hadamard eigenspaces. The vertical dotted line in Fig. 3 indicates the average burn-in time across the 10 realizations.

H.2 Metrics used in the experiments

We track the following metrics as functions of iteration:

(a) *Cross-Entropy Loss:* Plots the CE loss in Eq. (2).

(b) *Alignment Metrics:* Tracks the Frobenius norm distance between the nuclear-norm-normalized model components ($WW^T, H^T H, Z = WH$) and the ground-truth Simplex ETF S. For example for the logit matrix $Z = WH$ it plots $\left\| \frac{Z}{\|Z\|_*} - \frac{S}{\|S\|_*} \right\|_F$. For Hadamard Init., where Z and S share singular spaces, this is identical to the (square root of the) metric M in Eq. (11).

(c) *KL Divergence:* Plots the Lyapunov function from Theorem 3, measuring the KL divergence of the normalized logit singular values $\hat{a}$ from the uniform distribution $\frac{1}{K-1} 1_{K-1}$.

(d) *Singular Values:* Plots the evolution of the raw singular values $\sigma_i(Z)$ of the logit matrix $Z = WH$. For the *Hadamard Init.* and *ODE* runs, these are equivalent to the main body's logit singular values a_i.

(e) *Normalized Singular Values:* Plots the L_1-normalized singular values $\hat{\sigma}_i(Z) = \sigma_i(Z) / \sum_j \sigma_j(Z)$. For the *Hadamard Init.* and *ODE* runs, these are equivalent to the main body's $\hat{a}_i = a_i / \|a\|_1$.

(f) *Residual Energy:* Measures the fraction of the logit matrix Z that lies outside the signal subspace of the Hadamard Init. Specifically, for $U = V = \Phi(1 : K, 2 : K)$, it measures $\|Z - U(U^T ZV)V^T\|_F / \|Z\|_F$

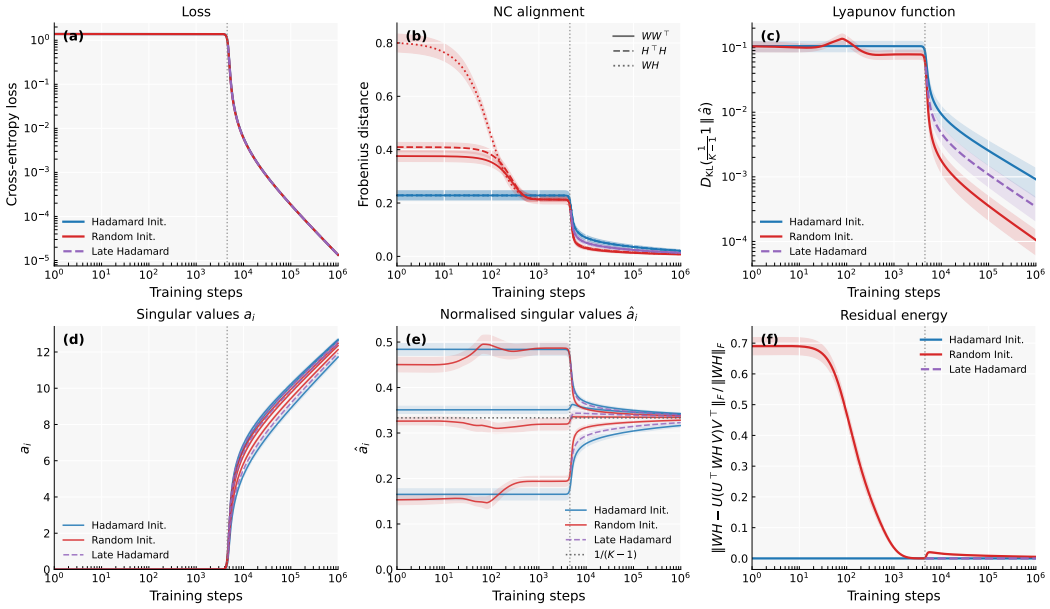
We track and plot these metrics in 5000 logarithmically spaced time instants. All results are averaged over 10 runs, with the shaded regions in our plots representing the standard error of the mean.

H.3 Additional random-versus-Hadamard experiments for $K = 4$ and $K = 16$

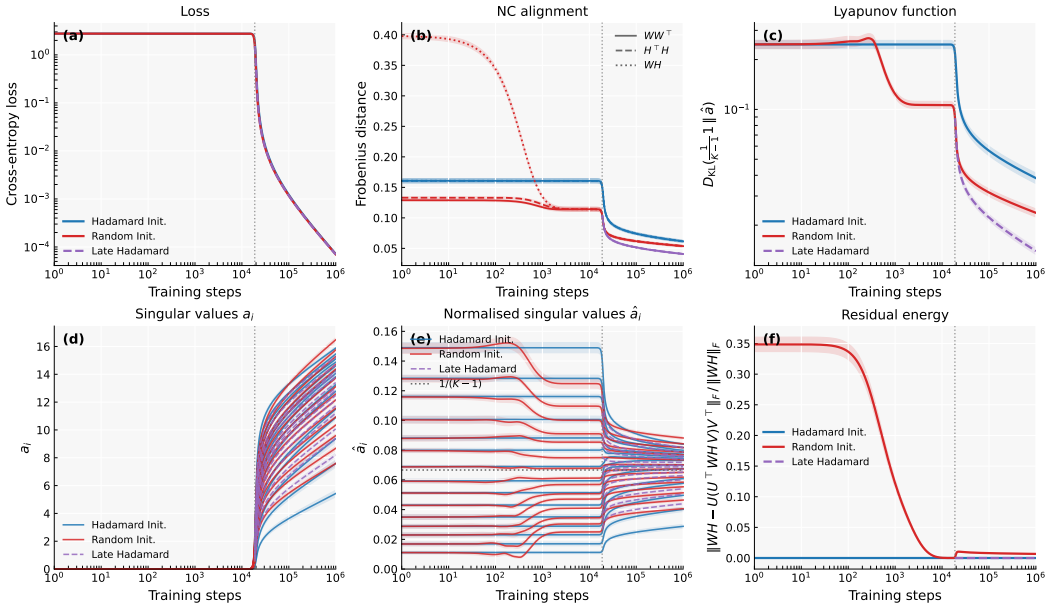
We provide additional experiments comparing small random initialization and Hadamard initialization for $K = 4$ and $K = 16$, with $d = 5K$, in Fig. 5. These experiments complement the $K = 8$ results in Fig. 3 (setup is the same as described in Sec. H.1). In all cases, both random and Hadamard initialization exhibit a similar qualitative two-stage behavior. In the first, small-logit phase, the loss changes slowly and the singular values remain small. For random initialization, this phase also corresponds to alignment of the logit matrix toward the Hadamard subspace; for Hadamard initialization, the trajectory is already in this subspace, and the normalized singular values remain constant. In the second, large-logit phase, the singular values grow, the loss decreases, and the normalized singular values begin moving toward their limiting equilibrium. This behavior is very similar to that observed in Fig. 3 of the main text.

H.4 The effect of narrow width

In the experiments thus far, we have set the UFM hidden dimension to $d = 5K$. To provide a case where the quantitative link between Hadamard initialization and random initialization is weaker (whilst the qualitative link remains unchanged), we consider the extreme narrow network case of $d = K - 1$. This is the minimal width where NC remains feasible.



(a) $K = 4, d = 5K$



(b) $K = 16, d = 5K$

Figure 5: Additional comparison between gradient flow under small random initialization and Hadamard initialization for $K = 4$ and $K = 16$. Plots are the same as Fig. 3 other than updated class number. The two-stage behavior of random initialization is visible in both cases.

For Hadamard initialization, our theory predicts that changing network width d should not affect the dynamics, since the trajectory remains confined to the same invariant spectral subspace. This is consistent with Fig. 6, where the Hadamard trajectories follow the same qualitative behavior as in the previous cases.

By contrast, Fig. 6 shows that, at narrow width, the quality of the Hadamard approximation depends on K. For $K = 4, 8$, the agreement between random and Hadamard initialization is quite strong after the alignment phase. In contrast, for $K = 16$, the qualitative behavior remains similar, but the quantitative agreement deteriorates. This is consistent with the increasingly coupled nature of the CE singular-value dynamics as the number of classes grows: small differences in singular-vector

alignment or in the raw singular values can lead to more visible discrepancies in the normalized singular values. Thus, Hadamard initialization should be interpreted as a tractable representative trajectory that captures the qualitative post-alignment behavior of random initialization, rather than as an exact approximation uniformly over K.

H.5 The effect of initialization scale

Thus far we used small initializations: recall from Sec. H.1 that W and H are initialized with i.i.d. Gaussian entries and then scaled by $\epsilon = e^{-\delta}$ with $\delta = 12$. For MSE, following [62], this vanishing initialization yields dynamics similar to spectral initialization; our previous CE experiments were the analogous setting. For completeness, we also report results under a standard scale, i.e., $\epsilon = 1$ ($\delta = 0$). Figures 7(a) and 7(b) show $K = 4$ and $K = 8$, respectively. The two-stage dynamics seen with infinitesimal initialization persist, but with quantitative changes: e.g., the initial "burn-in" shortens to $t \sim 5 \times 10^2$. Notably, while the alignment of logit singular values to their NC equilibrium (measured by NC metrics and KL divergence) remains comparable to the infinitesimal case, the residual energy is markedly higher: subspace alignment from random Initialization to the simplex ETF subspace proceeds much more slowly.

H.6 Normalized gradient flow

We have seen that gradient flow converges to the NC solution (with all normalized singular values $\hat{a}_i$ equal), but convergence can be very slow. The counterexamples constructed for metric M in Sec. 5.3—specifically Figure 2—illustrate this. For the initialization shown, metric M increases, but Theorem 3 guarantees it must eventually decrease to 0 as the flow converges to equilibrium. However, this occurs extremely slowly. To numerically verify this behavior, we implement an accelerated version of the reduced flow in Eq. (9), the *normalized* gradient flow:

$$\frac{da}{dt} = \frac{F(a)}{\|F(a)\|_2} \qquad \text{where } F(a) = -a \odot \nabla_a E(a) = \frac{a \odot b}{D}. \tag{36}$$

For convex models, L_2-normalized flow accelerates optimization without modifying convergence to the same max-margin solution as standard flow [54]. We leave proving this property in our non-convex setting to future work.

Unfortunately, direct implementation of the updates in Eq. (36) encounters numerical issues: The gradients of the energy functional $E(a)$ involve terms that can overflow when margins become large during training. This is particularly worrisome for normalized flow, which accelerates margin growth. To implement gradient flow stably, we rewrite gradient terms by factoring out dominant exponential terms and subtracting the minimum margin, an idea inspired by [50, App. I].

First note that given the gradient components $F_i(a) = a_i b_i / D$, the denominator D cancels entirely:

$$\frac{da_i}{dt} = \frac{F_i(a)}{\|F(a)\|_2} = \frac{a_i b_i / D}{\sqrt{\sum_k (a_k b_k / D)^2}} = \frac{a_i b_i}{\|a \odot b\|_2}.$$

We now apply a stability trick to the vector b itself. Let $m_j = (\Psi a)_j$ be the margins and $m_{\min} = \min_j m_j$. We define a numerically stable rescaled vector $\hat{b}$ as:

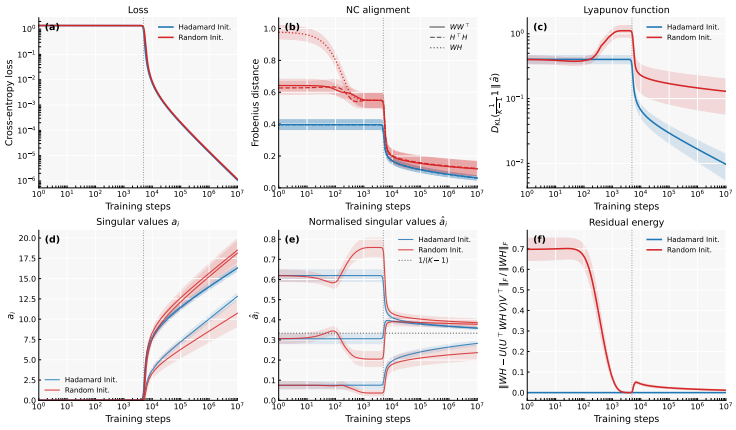
$$\hat{b}_i := b_i \cdot e^{m_{\min}} = \left(\sum_{j=1}^{K-1} \Psi_{ij} e^{-m_j} \right) e^{m_{\min}} = \sum_{j=1}^{K-1} \Psi_{ij} e^{-(m_j - m_{\min})}. \tag{37}$$

Substituting $b_i = \hat{b}_i e^{-m_{\min}}$ back into the normalized dynamics, the large exponential terms cancel:

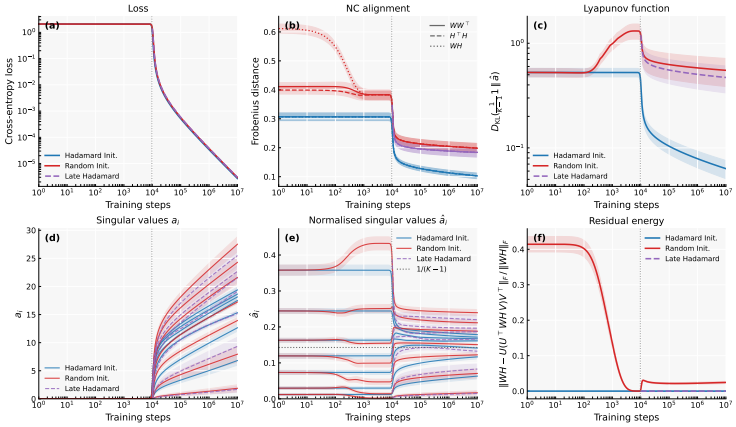
$$\frac{da_i}{dt} = \frac{a_i (\hat{b}_i e^{-m_{\min}})}{\|a \odot (\hat{b} e^{-m_{\min}})\|_2} = \frac{a_i \hat{b}_i}{\|a \odot \hat{b}\|_2}.$$

This gives the final, stable formulation for the normalized dynamics:

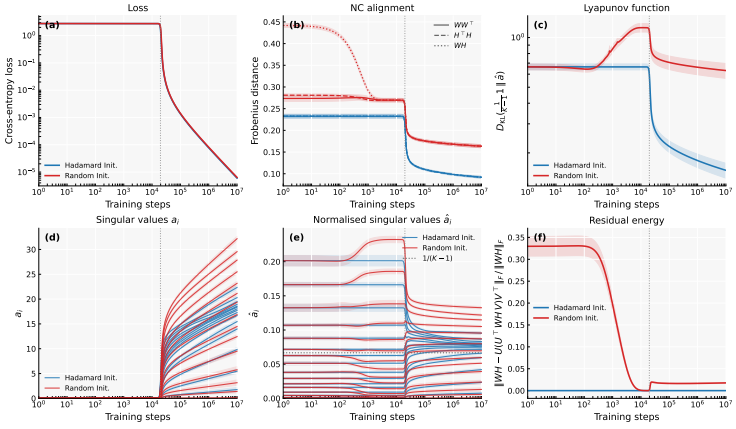
$$\frac{da_i}{dt} = \frac{a_i \sum_{j=1}^{K-1} \Psi_{ij} e^{-(m_j - m_{\min})}}{\sqrt{\sum_k \left(a_k \sum_{j=1}^{K-1} \Psi_{kj} e^{-(m_j - m_{\min})} \right)^2}} \tag{38}$$



(a) $K = 4$, $d = K - 1$

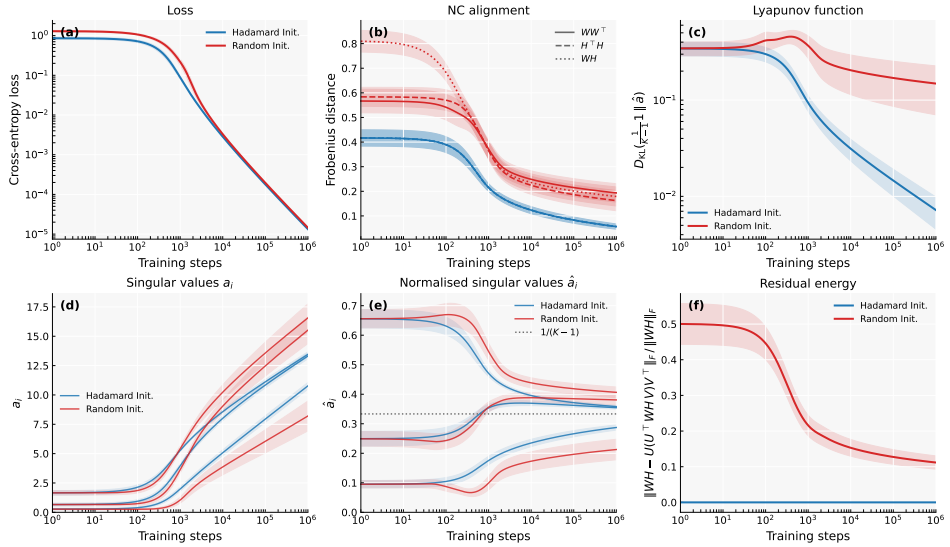


(b) $K = 8$, $d = K - 1$

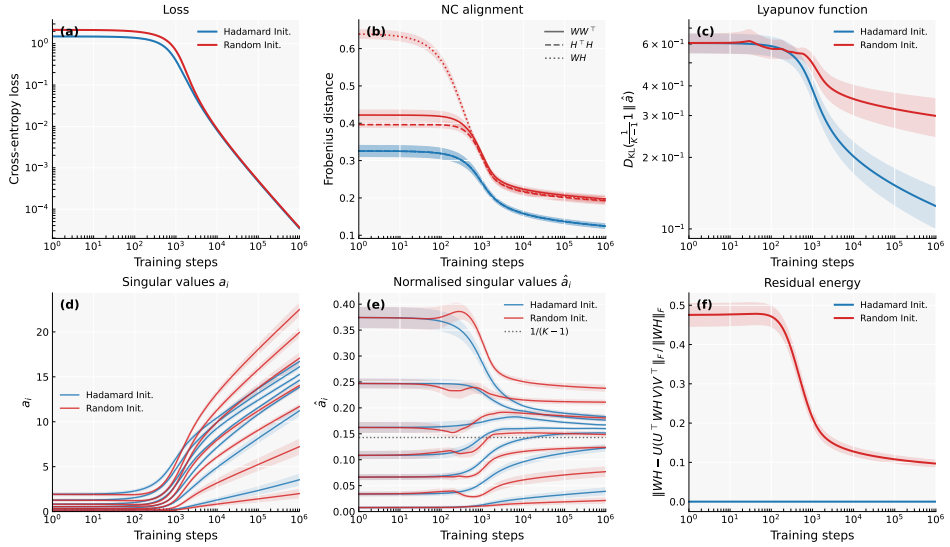


(c) $K = 16$, $d = K - 1$

Figure 6: Effect of narrow width on the comparison between small random initialization, Hadamard initialization, and late-Hadamard dynamics. In all cases, the hidden dimension is set to $d = K - 1$. Whilst qualitatively Hadamard still models random initialization, the quantitative agreement deteriorates when using narrow networks and larger class number K.



(a) $K = 4$, standard initialization scale



(b) $K = 8$, standard initialization scale

Figure 7: Same as Fig. 6, but now with standard initialization scale rather than infinitesimal initialization. Top: $K = 4$. Bottom: $K = 8$.

In this form, all exponents are non-positive, with at least one exponent in the sums being exactly zero. This is numerically stable for large values of a.

Figure 8 compares the normalized flow (solid lines) with regular gradient flow (dashed lines) for the non-monotonic initialization from Figure 2. The acceleration is dramatic: while regular flow requires time scales beyond $t = 10^6$ to approach the NC equilibrium $\hat{a}_i = 1/7$ (top right), normalized flow achieves this by path length $\tau \approx 10^3$. Both the Lyapunov function D_{KL} (bottom left) and metric M (bottom right) converge orders of magnitude faster under normalized flow.

The non-monotonic behavior of metric M observed in Figure 2 remains visible: normalized flow (solid blue) exhibits a similar but compressed trajectory where M first increases before eventually decreasing. Notably, the singular values (top left) become extremely large when the decrease in M begins. Without the stable implementation above, we would have encountered numerical overflow at this stage. This confirms that normalized flow preserves the qualitative dynamics while making convergence tractable on practical time scales.

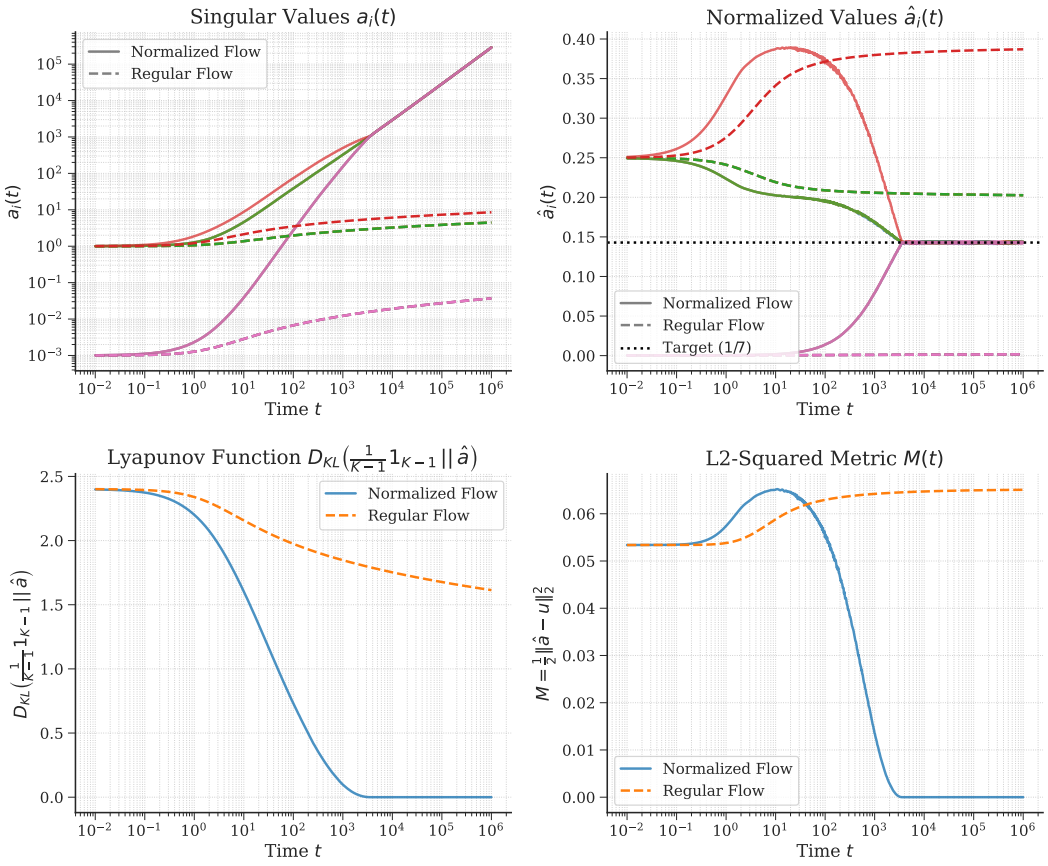


Figure 8: **Normalized Flow vs. Regular ODE Flow ($K = 8$).** A direct comparison of the normalized flow (solid lines) and the regular stable ODE flow (dashed lines) for the highly non-uniform initialization $a(0) = [1, 1, 1, 1, 0.001, 0.001, 0.001]$.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Each claim in the abstract and introduction is evidenced by a theorem or empirical result in Sections 4-7.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A paragraph on limitations is included in the conclusion.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions for results are stated in the theorems, and all proofs appear in the appendices.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper includes enough details in App. H so as to perform the same kind of experiments as the ones presented in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Our contributions are primarily theoretical, using experiments to illustrate the theory. The source code will be provided on request.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: All experimental details are provided in the main text and appendices.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For the majority of our experiments we do report error bars. In some cases we provide isolated examples to illustrate phenomenology, where reporting individual runs is sufficient.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: Our experiments are for our modeling context. These are two layer neural networks, which can be run without specific hardware set ups.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: No societal impact of the work performed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: We focus on our modeling context, and do not use external datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs were only used for grammar checking of the paper.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.