THE ACCURACY WELL: NONMONOTONE ACCURACY DYNAMICS AT VANISHING INITIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We study classification dynamics under small initialization and identify a pronounced non-monotone behavior of accuracy that we call the *accuracy well*: accuracy rises rapidly early in training, subsequently drops, and only much later recovers. We observe this phenomenon in deep linear networks, fully connected ReLU networks on MNIST, deep convolutional networks on CIFAR-10, and controlled synthetic tasks, across different losses and optimizers and in both training and test accuracy. We propose a three-stage explanation in terms of target-driven organization, a degeneracy bottleneck, and subsequent recruitment and recovery. We develop a rigorous theory of these dynamics for shallow and deep linear networks, showing in particular that high accuracy can emerge while the predictor and the change in loss remain asymptotically vanishing, before continued optimization causes accuracy to deteriorate. For a simple synthetic classification task, we prove an extreme instance in which accuracy asymptotically rises to 100%, collapses to 0, and later recovers permanently to 100%. Our results show that classification accuracy can reveal aspects of learning dynamics that are not captured by the training loss alone.

1 INTRODUCTION

Initialization scale can qualitatively change the dynamics of gradient-based learning. In suitable parameterizations and scalings, training can enter a lazy, kernel-like regime in which the nonlinear trajectory remains close to its linearization around initialization (Jacot et al., 2018; Chizat et al., 2019), whereas small or vanishing initialization can expose substantially richer evolution of the learned predictor (Woodworth et al., 2020). This regime has been studied from several complementary perspectives. In linear and homogeneous models it can produce sequential learning of spectral components, incremental low-rank structure, and saddle-to-saddle dynamics (Saxe et al., 2014; Gissin et al., 2020; Gidel et al., 2019; Jacot et al., 2021; Zhang et al., 2026); in nonlinear networks, substantial directional or feature organization can already occur before appreciable loss decay (Atanasov et al., 2022; Boursier et al., 2022; Min et al., 2024). Small initialization is therefore not merely a choice of scale: it can separate qualitatively different stages of learning that are compressed together at larger initialization.

In this paper we describe and analyze a small-initialization phenomenon that we call the *accuracy well*. We consider classification problems in which a model is trained from small initialization and follow its classification accuracy (i.e., the fraction of examples classified correctly) throughout training. At sufficiently small initialization, we repeatedly observe a pronounced non-monotone accuracy trajectory: it first rises rapidly, then significantly drops, and only much later recovers. We call the early rise an *accuracy hill*; together with the subsequent drop and recovery, it forms an *accuracy well*; see Figure 1. The phenomenon appears in deep linear networks, fully connected ReLU networks on MNIST, and deep convolutional networks on CIFAR-10, is visible in both training and test accuracy, persists across different losses and optimization methods, and disappears in several informative control settings.

Related non-monotone predictive phenomena include transient prediction-level effects arising from unequal spectral learning rates in deep linear networks (Saxe et al., 2019), example forgetting and subset-specific accuracy dynamics (Toneva et al., 2019; Schneider, 2025), as well as epoch-wise double descent, where the pronounced non-monotonicity is primarily observed in test or general-

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

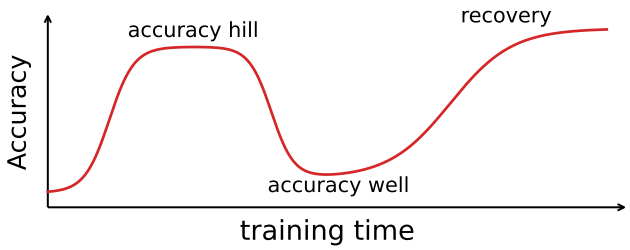

Figure 1: Schematic accuracy well: non-monotone accuracy dynamics at small initialization.

ization performance (Nakkiran et al., 2020; Heckel & Yilmaz, 2021; Olmin & Lindsten, 2024). Existing theoretical accounts of these neighboring phenomena do not directly characterize the pronounced non-monotone evolution of the training accuracy studied here. We instead follow how the induced classifier and its class ordering evolve from small initialization, and develop rigorous results for the resulting sequence of early improvement, intermediate deterioration, and later recovery. To our knowledge, such dynamics have not previously been analyzed directly under full-batch training. Detailed connections to prior work are discussed in Appendix A.

Our contributions.

- We *systematically analyze the phenomenology of the accuracy well* across linear and nonlinear models, including classification on MNIST, CIFAR-10, and controlled synthetic tasks; see Section 2.
- We *propose a general heuristic explanation of the accuracy well* in terms of a target-driven organization, a degeneracy bottleneck, and subsequent recruitment; see Section 3.
- We *develop a rigorous theory of the accuracy well* in shallow and deep linear networks. We obtain general results for classification dynamics under small initialization; see Section 4.

2 Phenomenology

We first analyze the accuracy well empirically across a range of models and training settings. Figures 2 and 3 collect representative examples and controlled sweeps that reveal several recurring features of the phenomenon. Our aim is not to provide an exhaustive taxonomy of when an accuracy well occurs, but to identify the main empirical regularities that motivate our theoretical study.

Experimental settings and breadth of the phenomenon. We consider both standard image-classification tasks and simple synthetic problems, using architectures ranging from fully connected linear and ReLU networks to a convolutional network. For natural data we use MNIST and CIFAR-10. As controlled synthetic tasks, we consider unary operations $f : [p] \to [p]$, where $[p] = \{1, \dots, p\}$, represented using one-hot inputs and outputs. We study random maps, where the values $f(i)$ are chosen independently and uniformly, and permutations, where f is bijective. Across these different settings we repeatedly observe pronounced accuracy wells, while several controls show that the phenomenon is not universal. On MNIST, the accuracy-well pattern is also reproducible across five independent random seeds. On CIFAR-10, we also observe an accuracy well in a five-hidden-layer convolutional network. Experimental details and additional figures are given in Appendix C.

Relation to loss decay. A recurring pattern is that the early accuracy hill forms while the loss does not visibly change. The accuracy then starts to fall, and the low point of the well typically occurs around the time when the loss begins to decrease noticeably. This behavior is visible across several sweeps in Figures 2 and 3. Thus large changes in classification can happen before substantial loss optimization, while the deepest part of the well tends to align with the onset of loss decay.

Initialization scale. The most consistent control parameter is the initialization scale. At relatively large σ, accuracy typically evolves more monotonically, whereas decreasing σ progressively sepa-

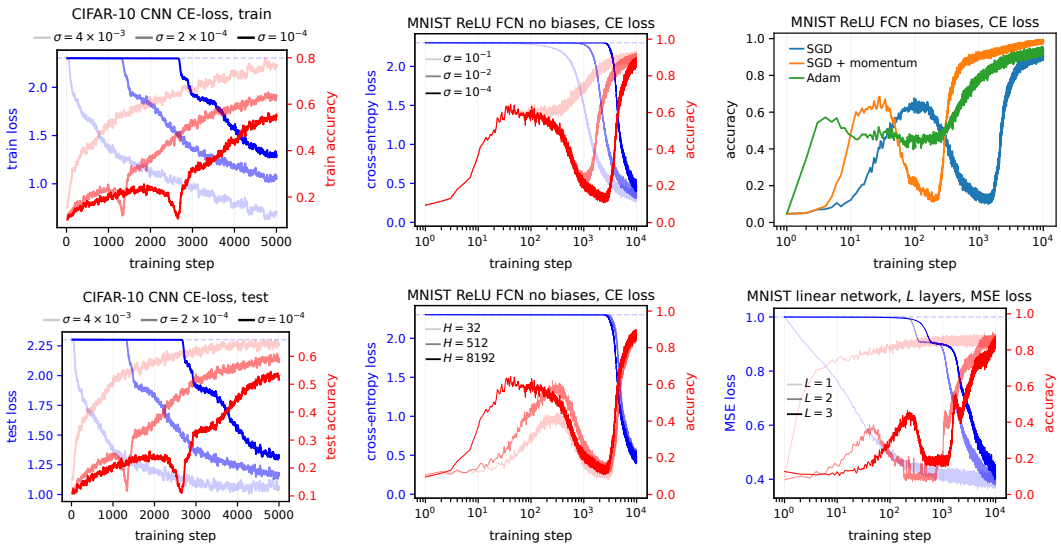

Figure 2: **Accuracy-well phenomenology on CIFAR-10 and MNIST.** *Left*: train and test dynamics of a two-hidden-layer CIFAR-10 CNN under an initialization-scale sweep. *Middle*: MNIST ReLU networks without biases under cross-entropy, varying initialization scale (top) and width (bottom). *Right*: optimizer comparison for the MNIST ReLU network (top) and a comparison of linear networks $f = \mathbf{W}_L \cdots \mathbf{W}_1$ on MNIST with $L = 1, 2, 3$ (bottom).

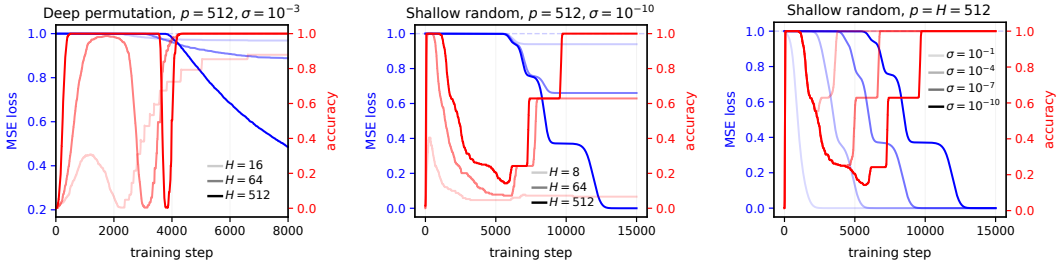

Figure 3: **Representative unary-operation experiments.** *Left*: a deep linear network ($L = 3$ layers) trained on a permutation operation develops a pronounced accuracy well as width increases. *Middle*: a shallow network ($L = 2$ layers) trained on a random unary operation develops a well in which the early classifier can substantially outperform the late-time predictor. *Right*: decreasing the initialization scale σ separates the rise, collapse, and recovery over increasingly distinct time scales.

rates the early hill, the collapse, and the eventual recovery. This is especially clear in the top-middle panel of Figure 2 and the right panel of Figure 3, where the same qualitative phases become increasingly well separated in training time as σ decreases.

Small initialization is nevertheless not sufficient. The shallow permutation model remains free of a well at small initialization, while the directly optimized linear MNIST model (bottom-right panel of Figure 2) also exhibits monotone accuracy dynamics. Thus small initialization makes the well sharply visible in many settings, but does not by itself determine whether one occurs.

Model capacity. Increasing width generally strengthens the early classifier and deepens the resulting well. This is visible for MNIST in the bottom-middle panel of Figure 2 and in the width sweeps in the left and middle panels of Figure 3. We do not claim a universal width-scaling law, since the quantitative effect is visibly task- and architecture-dependent.

Loss, train–test behavior, and optimization. The well appears across different losses and optimizers. The unary experiments use mean-squared error, while CIFAR-10 and most MNIST experi-

ments use cross-entropy. On CIFAR-10, train and test accuracy show closely aligned hills, collapses, and recoveries, so the phenomenon is not specific to either training or test accuracy. On MNIST, the well is present with SGD, SGD with momentum, and Adam, although the drop is smaller for Adam under the hyperparameters used here. A systematic optimizer comparison is beyond our scope.

Depth and target structure. Depth can qualitatively change accuracy dynamics even when the end-to-end function class is unchanged. The bottom-right panel of Figure 2 compares linear networks $f(\mathbf{x}) = \mathbf{W}_L \cdots \mathbf{W}_1 \mathbf{x}$ trained on MNIST with $L = 1, 2, 3$. All represent linear maps $\mathbb{R}^{784} \to \mathbb{R}^{10}$ and start from small initialization, but the deeper models ($L > 1$) develop pronounced accuracy wells while direct optimization ($L = 1$) does not.

Target structure also matters. For a two-layer linear network, a permutation target retains its early high accuracy, whereas a random unary target develops a pronounced well (Figure 3). At greater depth, however, even the permutation target can develop a well. Thus neither small initialization nor target structure alone determines the phenomenon; the depth matters as well. Additional architectural controls, including bias ablations, are reported in Appendix C.

Early versus late classifiers. The middle panel of Figure 3 shows a particularly striking effect: narrow random-unary networks can reach very high accuracy early in training but settle at substantially lower accuracy after much longer optimization of the MSE. Thus continued loss minimization does not necessarily improve classification accuracy: an early predictor can be a better classifier than a much later predictor with substantially smaller loss.

These observations raise the central question: what mechanism produces the accuracy well?

3 A HEURISTIC PICTURE OF THE ACCURACY WELL

We propose a three-stage picture for the accuracy well: *target-driven organization*, a *degeneracy bottleneck*, and *recruitment and recovery*. The first two stages occur on a microscopic scale: the classifier can reorganize substantially while the predictor remains small and the loss is nearly unchanged. The final recovery instead occurs on the macroscopic training scale, together with substantial growth of the predictor and loss decay.

Target-driven organization. The first stage can be understood from a local linearization of the initial dynamics, closely related to the picture underlying lazy learning (Jacot et al., 2018; Chizat et al., 2019). Here, however, we use this only as a short-time description suitable for illustrating the first stage. Subsequent learning may involve dynamics beyond this initial first-order behavior.

Consider a C-class model $f_{\boldsymbol{\theta}} : \mathbb{R}^p \to \mathbb{R}^C$, parametrized by $\boldsymbol{\theta} \in \mathbb{R}^d$, and stack its outputs and the corresponding ground-truth targets over the training set into $\mathbf{F}_{\boldsymbol{\theta}}, \mathbf{Y} \in \mathbb{R}^{NC}$. For squared loss,

$$\dot{\mathbf{F}}_{\boldsymbol{\theta}} = \mathbf{K}_{\boldsymbol{\theta}} (\mathbf{Y} - \mathbf{F}_{\boldsymbol{\theta}}), \qquad \mathbf{K}_{\boldsymbol{\theta}} = \mathbf{J}_{\boldsymbol{\theta}} \mathbf{J}_{\boldsymbol{\theta}}^{\top}, \qquad \mathbf{J}_{\boldsymbol{\theta}} = \frac{\partial \mathbf{F}_{\boldsymbol{\theta}}}{\partial \boldsymbol{\theta}}, \tag{1}$$

where $\mathbf{K}_{\boldsymbol{\theta}}$ is the empirical tangent kernel. If vanishing initialization makes $\mathbf{F}_{\boldsymbol{\theta}(0)}$ small, then over a sufficiently short initial interval the tangent kernel is approximately frozen to first order, and

$$\mathbf{F}_{\boldsymbol{\theta}(\tau)} \approx \underbrace{\mathbf{F}_{\boldsymbol{\theta}(0)}}_{\text{initialization noise}} + \underbrace{\tau \mathbf{K}_{\boldsymbol{\theta}(0)} \mathbf{Y}}_{\text{target-driven signal}}. \tag{2}$$

Equation (2) suggests a simple competition during the first stage. The predictor initially contains a random contribution inherited from initialization, while training begins to build a coherent component aligned with the targets. At very early times the former may still dominate, but the target-driven component grows with τ. Increasing the width favors the target-driven component in this competition: initialization-dependent fluctuations are reduced by self-averaging, whereas the target-driven contribution adds coherently. Thus there can be an early window in which τ is already large enough for the target-driven signal to control the score ordering, but still small enough for the local linearized description above to remain accurate. In this window the predictor itself may remain microscopic—indeed, under vanishing initialization the tangent kernel and hence the target-driven signal may also vanish in absolute magnitude—yet the classifier can already achieve high accuracy while the loss has barely changed. Our linear theory in Section 4 makes this target-driven organization rigorous,

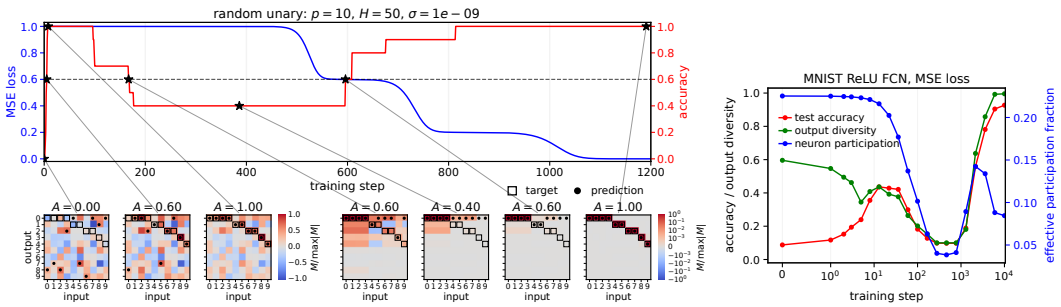

Figure 4: **Diagnostics of the proposed accuracy-well mechanism.** *Left*: evolution of a two-layer linear network $\mathbf{M} = \mathbf{W}_2\mathbf{W}_1$ on a random unary operation. Snapshots of the normalized score matrix show how the model reorganizes along the accuracy-well trajectory. Notably, snapshots with the same accuracy ($A = 0.6$) have qualitatively different score-matrix structure. *Right*: test accuracy, effective hidden-neuron participation fraction, and output diversity in a one-hidden-layer ReLU network on MNIST. The participation fraction measures the effective fraction of hidden neurons contributing to class-dependent scores, while output diversity measures the effective fraction of output classes used by the predictions; see equations (939), (943) in Appendix C.2.

with the precise near-origin dynamics depending on the model and depth. The same qualitative mechanism may also extend beyond squared loss.

Degeneracy bottleneck. We associate the subsequent loss of accuracy with the predictor becoming temporarily *degenerate* while still remaining microscopic. We call this stage a *degeneracy bottleneck*. Its precise form is model-dependent. In two-layer linear networks, the predictor can be viewed as a sum of rank-one singular components whose microscopic growth rates may separate strongly. The dynamics can therefore become dominated by only a small subset of these components, possibly containing several equally dominant modes. If this dominant subset does not preserve the class ordering produced by their earlier combination, accuracy falls. This mechanism is visible in Figure 4 (left) and is made rigorous in Section 4.

For deeper linear networks, our rigorous result uses the stronger nondegeneracy assumption that the leading singular direction of the target is unique. Under generic Gaussian initialization, the first saddle escape is then asymptotically rank one, with the probability guarantees described in Section 4, by the result of Jacot et al. (2021). For nonlinear networks we do not claim an analogous spectral characterization and instead use degeneracy as an empirical diagnostic. In the one-hidden-layer ReLU experiment, both the effective hidden-neuron participation fraction and the output diversity decrease sharply near the bottom of the accuracy well; see Figure 4 (right) and Appendix C.2. Thus, across these settings, the bottleneck corresponds to a transient loss of predictive diversity, although the microscopic mechanism producing it need not be universal.

Recruitment and recovery. The final stage is qualitatively different from the first two. Recovery takes place on the macroscopic training scale, when the predictor grows substantially and the loss begins to decrease. At this stage, predictive structure that was suppressed or absent at the bottleneck can become dynamically relevant. We call this re-entry or emergence of complementary predictive structure *recruitment*. In linear models this corresponds to the progressive incorporation of additional singular modes, closely related to classical mode-learning and incremental-learning dynamics (Saxe et al., 2014). In nonlinear networks, the analogous phenomenon is naturally viewed in terms of feature learning: task-relevant features or directions can emerge progressively over training (Dandi et al., 2024). We use *recruitment* to emphasize specifically this recovery-side increase in predictive diversity.

Both panels of Figure 4 are consistent with this picture. In the linear network example, the score matrix regains a diverse, target-like structure as accuracy recovers. In the ReLU network, effective neuron participation and the output diversity rise again during the recovery stage. Thus the accuracy well separates two microscopic stages, target-driven organization and a degeneracy bottleneck, from a later macroscopic stage of recruitment and recovery.

4 ACCURACY DYNAMICS IN DEEP LINEAR NETWORKS UNDER VANISHING INITIALIZATION [APPENDIX B]

We now develop a rigorous realization of this picture in deep linear networks. Consider a C-class training set $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, with $\mathbf{x}_i \in \mathbb{R}^p$, $\mathbf{y}_i = \mathbf{e}_{c_i} \in \mathbb{R}^C$, $c_i \in \{1, \dots, C\}$, and a linear classifier

$$f(\mathbf{x}) = \mathbf{W}_L \cdots \mathbf{W}_1 \mathbf{x} = \mathbf{M}\mathbf{x}.$$

For $L \geq 2$, all hidden layers have width H: $\mathbf{W}_1 \in \mathbb{R}^{H\times p}$, $\mathbf{W}_L \in \mathbb{R}^{C\times H}$, and $\mathbf{W}_\ell \in \mathbb{R}^{H\times H}, 1 < \ell < L$. For $L = 1$, the end-to-end predictor $\mathbf{M} \in \mathbb{R}^{C\times p}$ is optimized directly. The weights are initialized independently with $\mathcal{N}(0, \sigma^2)$ entries and trained by gradient flow on $\mathcal{L}(\mathbf{M}) = (2N)^{-1} \sum_i \|\mathbf{M}\mathbf{x}_i - \mathbf{y}_i\|_2^2$. Although cross-entropy is more common for classification, we use mean squared loss for analytical tractability.

Our main object is the training accuracy

$$\mathcal{A}(\mathbf{M}) = \frac{1}{N} \sum_{i=1}^N \mathbf{1}\left\{ (\mathbf{M}\mathbf{x}_i)_{c_i} > \max_{c\neq c_i} (\mathbf{M}\mathbf{x}_i)_c \right\}. \tag{3}$$

We focus on small initialization limit, i.e. we first take $\sigma \downarrow 0$, followed by the appropriate large-system limit; precise orders of limits are stated in the corresponding theorems. Throughout this section we assume isotropic input covariance $\frac{1}{N}\sum_{i=1}^N \mathbf{x}_i\mathbf{x}_i^\top = \gamma \mathbf{I}_p$, and define the effective target

$$\mathbf{T} = \frac{1}{N\gamma} \sum_{i=1}^N \mathbf{y}_i \mathbf{x}_i^\top, \tag{4}$$

so the loss can be expressed as $\mathcal{L}(\mathbf{M}) = \frac{\gamma}{2}\|\mathbf{M} - \mathbf{T}\|_{\mathrm{F}}^2 + \mathrm{const}$. We will work with rescaled time $\tau = \gamma t$ throughout the section. We consider SVD and the truncated SVD, respectively:

$$\mathbf{T} = \sum_{j=1}^r s_j \mathbf{p}_j \mathbf{q}_j^\top, \qquad s_1 \geq \cdots \geq s_r > 0, \qquad \mathbf{T}_k := \sum_{j=1}^k s_j \mathbf{p}_j \mathbf{q}_j^\top. \tag{5}$$

Let $k_\star$ be the multiplicity of the largest singular value, $s_1 = \cdots = s_{k_\star} > s_{k_\star+1}$, $s_{r+1} := 0$, so $\mathbf{T}_{k_\star}$ is the complete top singular component and is independent of the choice of singular-vector basis within the degenerate top subspace. We consider two accuracy levels:

$$A_{\mathbf{T}} := \mathcal{A}(\mathbf{T}), \qquad A_{k_\star} := \mathcal{A}(\mathbf{T}_{k_\star}). \tag{6}$$

Note that both cases $A_{\mathbf{T}} > A_{k_\star}$ and $A_{\mathbf{T}} < A_{k_\star}$ are possible (see Appendix B.2), although the former is observed consistently across the standard image datasets in Table 2 in Appendix B.2. For $L = 2$, we show that the microscopic dynamics can pass from the classifier of $\mathbf{T}$ to that of $\mathbf{T}_{k_\star}$. When $k_\star = 1$, this is the rank-one component $\mathbf{T}_1 = s_1\mathbf{p}_1\mathbf{q}_1^\top$. Our $L \geq 3$ saddle-escape result requires this simple-leading-mode case, $k_\star = 1$. Table 1 in Appendix B summarizes the results and their precise regimes.

Direct linear model: $L = 1$ **[B.1].** We begin with direct optimization of the end-to-end predictor $\mathbf{M}$. Its gradient-flow trajectory is explicit:

$$\mathbf{M}(\tau) = e^{-\tau}\mathbf{M}(0) + (1 - e^{-\tau})\mathbf{T}. \tag{7}$$

Thus $\mathbf{M}(\tau)$ moves along the line segment from its initialization to the least-squares solution $\mathbf{T}$. Along this trajectory, agreement with the classifier induced by $\mathbf{T}$ is monotone on every training example: once $\mathbf{M}(\tau)$ and $\mathbf{T}$ assign the same top-scoring class to an example, they continue to agree on that example at all later times (Lemma B.1). This yields a direct bound on any possible nonmonotonicity of the training accuracy.

Theorem 4.1 [B.4] **(Accuracy well bound for $L = 1$).** *Suppose that $\mathbf{T}$ has a unique top-scoring class on every training example and $\mathcal{A}(\mathbf{T}) = 1 - \varepsilon$. Then, for every $0 \leq \tau_1 \leq \tau_2$,*

$$\mathcal{A}(\mathbf{M}(\tau_2)) \geq \mathcal{A}(\mathbf{M}(\tau_1)) - \varepsilon. \tag{8}$$

In particular, if $\mathcal{A}(\mathbf{T}) = 1$, then $\mathcal{A}(\mathbf{M}(\tau))$ is monotone nondecreasing.

Furthermore, under vanishing initialization, any nontrivial classifier dynamics for $L = 1$ are confined to the microscopic regime. For $\mathbf{M}(0) = \sigma\mathbf{Z}$, all classifier changes occur while $\|\mathbf{M}(\tau)\|_{\mathrm{F}} = O(\sigma)$, and by the time the classifier stabilizes only an $O(\sigma)$ fraction of the total loss decrease from initialization to the least-squares solution has occurred (Theorem B.3). After this initial interval, the classifier agrees permanently with that of $\mathbf{T}$. Thus the direct linear model provides a sharply constrained baseline. As we show next, factorized linear networks with $L \geq 2$ can exhibit qualitatively richer microscopic classifier dynamics, including a target-driven rise followed by a degeneracy bottleneck before the later macroscopic recovery.

Two-layer networks: $L = 2$ **[B.2].** We consider $\mathbf{M} = \mathbf{W}_2\mathbf{W}_1$. Gradient flow satisfies

$$\dot{\mathbf{W}}_1 = \mathbf{W}_2^\top \mathbf{T} - \mathbf{W}_2^\top \mathbf{W}_2 \mathbf{W}_1, \qquad \dot{\mathbf{W}}_2 = \mathbf{T}\mathbf{W}_1^\top - \mathbf{W}_2 \mathbf{W}_1 \mathbf{W}_1^\top. \tag{9}$$

The cubic terms vanish as $\sigma \downarrow 0$ on fixed time intervals, leaving a linear limiting dynamics. Integrating the linear limiting dynamics, consider the normalized mean predictor

$$\overline{\mathbf{M}}(\tau) := \lim_{\sigma \downarrow 0} H^{-1}\sigma^{-2}\,\mathbb{E}[\mathbf{M}(\tau)] = \sum_{j=1}^{r} \sinh(2s_j\tau)\,\mathbf{p}_j\mathbf{q}_j^\top, \tag{10}$$

At early times, $\overline{\mathbf{M}}(\tau) = 2\tau\mathbf{T} + O(\tau^3)$, while finite width introduces initialization-dependent fluctuations. Schematically,

$$\mathbf{M}(\tau) = 2H\sigma^2\tau\left[\mathbf{T} + O_{\mathbb{P}}\Big(H^{-1/2}\tau^{-1}\Big) + O(\tau^2)\right]. \tag{11}$$

Thus, for sufficiently large width, there is an early regime in which τ is large enough for the target-driven component to dominate the random fluctuations, yet small enough for the higher-order dynamical correction to remain negligible. We call a time in this regime τ_{org}, the *organization time*; the predictor then induces the classifier of $\mathbf{T}$, with accuracy $A_{\mathbf{T}}$, while remaining microscopic.

At later microscopic times, the factors $\sinh(2s_j\tau)$ increasingly favor the largest singular modes. Hence, at a suitable *bottleneck time* $\tau_{\mathrm{bot}} > \tau_{\mathrm{org}}$,

$$\mathbf{M}(\tau_{\mathrm{bot}}) \approx H\sigma^2 s_1^{-1} \sinh(2s_1\tau_{\mathrm{bot}})\,\mathbf{T}_{k_\star}, \tag{12}$$

with the slower modes and finite-width fluctuations negligible relative to the dominant block. The predictor therefore induces the classifier of $\mathbf{T}_{k_\star}$, with accuracy $A_{k_\star}$. Thus target-driven organization is followed by a rank-$k_\star$ degeneracy bottleneck, while the predictor norm and loss change remain vanishing in the small-initialization limit. At later times the full gradient flow converges to $\mathbf{T}$, restoring the target classifier.

To make it rigorous, we consider a large-system sequence indexed by $H \to \infty$, where p, C, and N may also depend on H and grow sufficiently mildly. We assume bounded input norms and $\|\mathbf{T}\|_{\mathrm{op}}$, sufficient width $H \geq \mathrm{rank}(\mathbf{T})$, a nonvanishing spectral gap between the top singular block and the remaining modes, and decision margins of $\mathbf{T}$ and $\mathbf{T}_{k_\star}$ bounded away from zero. These uniform margin assumptions can be restrictive in growing-dimensional settings, particularly when the rank $k_\star$ of the top singular block remains fixed; the more quantitative margin-dependent conditions underlying the result are stated in Appendix B.2.

Theorem 4.2 [B.11] **(Three-stage accuracy dynamics for $L = 2$).** *There exist deterministic times $\tau_{\mathrm{org}} < \tau_{\mathrm{bot}}$, which may depend on N, p, C, H and the dataset $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ but not on σ or on the random initialization, such that, in the sequential limit $\sigma \downarrow 0$ followed by $H \to \infty$,*

$$\mathbb{P}(\mathcal{A}(\mathbf{M}(\tau_{\mathrm{org}})) = A_{\mathbf{T}},\ \mathcal{A}(\mathbf{M}(\tau_{\mathrm{bot}})) = A_{k_\star}) \longrightarrow 1. \tag{13}$$

Moreover, for each fixed member of the sequence, almost surely as $\sigma \downarrow 0$, the entire trajectory up to the bottleneck time remains microscopic:

$$\sup_{0 \leq \tau \leq \tau_{\mathrm{bot}}} \|\mathbf{M}(\tau)\|_{\mathrm{op}} \longrightarrow 0, \qquad \sup_{0 \leq \tau \leq \tau_{\mathrm{bot}}} |\mathcal{L}(\tau) - \mathcal{L}(0)| \longrightarrow 0. \tag{14}$$

Whenever $A_{\mathbf{T}} > A_{k_\star}$, these stages establish an accuracy well: an early target-level classifier, a degeneracy bottleneck, and eventual permanent recovery.

Unary operations. We now specialize the two-layer model to a unary operation $f : [p] \to [p]$, represented by one-hot inputs and outputs: $N = C = p$, $\mathbf{x}_i = \mathbf{e}_i$, and $\mathbf{y}_i = \mathbf{e}_{f(i)}$. Let $n_c := |\{i : f(i) = c\}|$ denote the occupancy of output c. Then

$$\mathbf{T}\mathbf{e}_i = \mathbf{e}_{f(i)}, \qquad \mathbf{T}\mathbf{T}^\top = \mathrm{diag}(n_1, \ldots, n_p). \tag{15}$$

Thus $\mathbf{T}$ has perfect accuracy, and its nonzero singular values are exactly $\sqrt{n_c}$. The occupancy structure therefore determines the microscopic growth rates exactly.

Consider first a random unary operation, $f(1), \ldots, f(p) \overset{\text{iid}}{\sim} \mathrm{Unif}[p]$. The occupancies are asymptotically Poisson and determine the rates of singular mode growth. This setting does not satisfy the nonvanishing decision-gap assumption for the top singular block in Theorem 4.2: on every submaximal occupancy fiber, $\mathbf{T}_{k_\star}$ vanishes and hence has zero decision margin. Consequently, the collapse is not visible in the mean predictor (10) alone: for every finite $\tau > 0$, its nonzero entry in each column remains on the correct output, so its accuracy is still one. The bottleneck is instead a finite-width effect. Modes associated with larger occupancies grow faster, and on examples associated with slower modes the corresponding mean margin becomes asymptotically negligible relative to the dominant modes, allowing amplified finite-width fluctuations to determine the predicted class. As the problem size grows, this mechanism drives the bottleneck accuracy to zero; the detailed occupancy argument is given in Appendix B.2.

Theorem 4.3 [B.18] **(Random unary operation).** *For every $\alpha > 0$, there exist deterministic times $\tau_{\mathrm{org}} < \tau_{\mathrm{bot}}$, which may depend on p and H, but not on σ, the random operation, or the random initialization, such that, in the sequential limit $\sigma \downarrow 0$ followed by $p, H \to \infty$ with $H/p \to \alpha$,*

$$\mathcal{A}(\mathbf{M}(\tau_{\mathrm{org}})) \longrightarrow 1, \qquad \mathcal{A}(\mathbf{M}(\tau_{\mathrm{bot}})) \longrightarrow 0 \tag{16}$$

in probability. Moreover, for each fixed p, H, almost surely as $\sigma \downarrow 0$, the entire trajectory up to the bottleneck remains microscopic:

$$\sup_{0 \le \tau \le \tau_{\mathrm{bot}}} \|\mathbf{M}(\tau)\|_{\mathrm{op}} \longrightarrow 0, \qquad \sup_{0 \le \tau \le \tau_{\mathrm{bot}}} |\mathcal{L}(\tau) - \mathcal{L}(0)| \longrightarrow 0. \tag{17}$$

If, in addition, $\alpha > 1 - e^{-1}$, then with probability tending to one over the random operation, for every fixed $\sigma > 0$ there almost surely exists a finite, possibly σ- and realization-dependent recovery time $\tau_{\mathrm{rec}} > \tau_{\mathrm{bot}}$, such that

$$\mathcal{A}(\mathbf{M}(\tau)) = 1 \qquad \text{for all } \tau \ge \tau_{\mathrm{rec}}. \tag{18}$$

The random-unary well is driven by heterogeneity of the output occupancies, which produces different singular values and hence different microscopic growth rates. It is therefore natural to ask what happens when this source of spectral separation is removed. A permutation provides the opposite extreme: $n_c = 1$ for every c, so all singular values of $\mathbf{T}$ are equal to one. There is consequently no spectral hierarchy capable of producing the bottleneck above. In fact, the behavior is stronger: the accuracy is nondecreasing, with all correct-versus-incorrect margins increasing until perfect accuracy is reached and then retained permanently.

Theorem 4.4 [B.23] **(Permutation operation: monotone accuracy).** *For every $\alpha > 0$, there exists a deterministic time τ_{up}, which may depend on p and H, but not on σ, the permutation, or the random initialization, such that, in the sequential limit $\sigma \downarrow 0$ followed by $p, H \to \infty$ with $H/p \to \alpha$, with probability tending to one,*

$$\mathcal{A}(\mathbf{M}(\tau)) \quad \text{is nondecreasing for all } \tau \ge 0, \qquad \mathcal{A}(\mathbf{M}(\tau)) = 1 \quad \text{for all } \tau \ge \tau_{\mathrm{up}}. \tag{19}$$

Moreover, for each fixed p, H, almost surely as $\sigma \downarrow 0$,

$$\sup_{0 \le \tau \le \tau_{\mathrm{up}}} \|\mathbf{M}(\tau)\|_{\mathrm{op}} \longrightarrow 0, \qquad \sup_{0 \le \tau \le \tau_{\mathrm{up}}} |\mathcal{L}(\tau) - \mathcal{L}(0)| \longrightarrow 0. \tag{20}$$

Thus heterogeneous occupancies produce a nearly complete accuracy well, whereas the flat spectrum of a permutation eliminates the bottleneck altogether. Remarkably, permanent perfect classification for a permutation holds even for $\alpha < 1$, where $\mathrm{rank}(\mathbf{M}) \le H < p = \mathrm{rank}(\mathbf{T})$ and exact fitting is impossible.

Deeper networks: $L \geq 3$ **[B.3].** For generic Gaussian initialization, we combine the same target-driven organization mechanism as for $L = 2$ with the rank-one saddle-escape result of Jacot et al. (2021). At an early stage, the predictor develops a coherent component aligned with the full target $\mathbf{T}$, while finite-width fluctuations self-average, so the classifier of $\mathbf{T}$ emerges while the predictor norm and the change in loss remain asymptotically vanishing. At a later stage, when $s_1 > s_2$, the trajectory escapes the initial rank-zero saddle through the leading rank-one singular component, producing the classifier of the leading rank-one component $\mathbf{T}_1$.

To make this rigorous, we consider fixed $L \geq 3$ and a large-system sequence indexed by $H \to \infty$. We assume bounded input norms and $\|\mathbf{T}\|_{\mathrm{op}}$, a nonvanishing decision margin for $\mathbf{T}$, unique top-scoring classes for $\mathbf{T}_1$ on each finite problem, an isolated leading singular value $s_1 > s_2$, and sufficiently mild growth of the problem dimensions: $p/H = O(1)$, $C/H = O(1)$, and $\log^{L-1}(2NCH) = o(\sqrt{H})$.

Theorem 4.5 [B.28] **(Target-driven organization and predictive bottleneck for $L \geq 3$).** *Under the assumptions above, there exist times $\tau_{\mathrm{org}} < \tau_{\mathrm{bot}}$, where τ_{bot} may depend on the initialization, such that the following holds. In the sequential limit $\sigma \downarrow 0$ followed by $H \to \infty$, with probability tending to one for $L = 3$, and with asymptotic probability at least $1/2$ for $L > 3$, we have*

$$\mathcal{A}(\mathbf{M}(\tau_{\mathrm{org}})) = A_{\mathbf{T}}, \qquad \mathcal{A}(\mathbf{M}(\tau_{\mathrm{bot}})) = A_1. \tag{21}$$

Moreover, in the same sequential limit, uniformly for $0 \leq \tau \leq \tau_{\mathrm{bot}}$, the predictor norm vanishes and the loss converges to its value at the zero predictor.

Spectral-mode recruitment and recovery under aligned initialization [B.3]. For generic Gaussian initialization, our theory rigorously establishes the first two stages—target-driven organization followed by a rank-one bottleneck—but recovery of the same trajectory remains open, since available convergence results do not exclude lower-rank limiting critical points (Bah et al., 2022; Nguegnang et al., 2024). To isolate the recovery mechanism, we therefore consider the spectrally aligned and balanced dynamics of Saxe et al. (2014), where

$$\mathbf{M}(\tau) = \sum_{k=1}^{r} m_k(\tau)\mathbf{p}_k\mathbf{q}_k^\top, \qquad \dot{m}_k = L(s_k - m_k)m_k^{2-2/L}. \tag{22}$$

The leading mode can dominate transiently, producing the same low-rank bottleneck, but its growth slows as $m_1 \to s_1$, while the remaining modes are progressively recruited. Eventually $\mathbf{M}(\tau) \to \mathbf{T}$, and the target classifier is recovered permanently. Corollary B.29 makes this precise: whenever the aligned trajectory enters the leading-mode bottleneck, its accuracy equals A_1 there and subsequently returns permanently to $A_{\mathbf{T}}$. Hence, if $A_{\mathbf{T}} > A_1$, the aligned dynamics rigorously exhibit the bottleneck-and-recovery part of an accuracy well. Together with the generic-initialization result above, this establishes the three proposed stages in complementary rigorous settings, while a complete three-stage theorem for one generic Gaussian-initialized $L \geq 3$ trajectory remains open.

5 Discussion

Accuracy need not mirror optimization progress. Training loss decreases monotonically under gradient flow, yet classification accuracy can follow a qualitatively different trajectory. We prove that a high-accuracy classifier can emerge while both the predictor and loss change remain microscopic, and can subsequently be destroyed by continued optimization before recovering. In the random-unary model this separation is extreme: accuracy rises asymptotically to 100%, collapses to 0, and later recovers permanently to 100%. Our deep-linear analysis therefore reveals predictive structure that appears before the familiar rank-one saddle escape and before appreciable progress in the optimized objective.

Limitations and future directions. Our rigorous analysis is restricted to linear networks, whereas the experiments show pronounced accuracy wells also in nonlinear models on MNIST and CIFAR-10. Even within linear networks, the pronounced well observed for deep permutation targets is not covered by our $L \geq 3$ theory, which assumes an isolated leading singular value. Understanding these flat-spectrum deep dynamics, identifying the mechanisms responsible for the well in nonlinear networks, and determining whether the organization–bottleneck–recruitment picture extends to broader architectures are natural directions for future work.

AI USE STATEMENT

Generative AI tools were used extensively throughout this work to assist with research brainstorming, the development and refinement of theoretical ideas and hypotheses, mathematical derivations and proof development, the design and analysis of numerical experiments, interpretation of results, and literature search. They were also used to assist with experimental code, figure generation, and the drafting, revision, organization, and presentation of the manuscript.

All AI-assisted mathematical arguments were treated as provisional and were independently checked by the authors, including their assumptions, derivations, theorem statements, and proofs. AI-assisted code and numerical results were likewise inspected and validated, and references identified with the help of AI tools were checked against the original sources before inclusion. The authors reviewed all AI-assisted material and take full responsibility for the mathematical claims, experimental results, citations, and conclusions presented in this work.

REPRODUCIBILITY STATEMENT

We provide the assumptions, definitions, normalization conventions, and proof details needed to reproduce the theoretical results in the main text and appendices. In particular, the model parameterizations, initialization scalings, loss normalizations, orders of limits, and conditions used in the theorems are stated explicitly, and the corresponding proofs are included in the appendix.

For the experimental results, the appendix specifies the model architectures, data preprocessing, initialization procedures, optimization settings, training protocols, evaluation procedures, and other implementation details used in the unary-operation, MNIST, and CIFAR-10 experiments. The Jupyter notebooks used to run the experiments and generate the reported figures will be provided as supplementary material.

REFERENCES

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. In *International Conference on Learning Representations*, 2019.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=1NvflqAdoom.

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. doi: 10.1093/imaiai/iaaa039.

Ioannis Bantzis, James B. Simon, and Arthur Jacot. Saddle-to-saddle dynamics in deep ReLU networks: Low-rank bias in the first saddle escape. In *International Conference on Learning Representations*, 2026.

Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. *Journal of Machine Learning Research*, 26(183):1–75, 2025. URL https://jmlr.org/papers/v26/24-1523.html.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow ReLU networks for square loss and orthogonal inputs. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/7eeb9af3eb1f48e29c05e8dd3342b286-Abstract-Conference.html.

Lénaïc Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/ae614c557843b1df326cb29c57225459-Abstract.html.

Hung-Hsu Chou, Carsten Gieshoff, Johannes Maly, and Holger Rauhut. Gradient descent for deep matrix factorization: Dynamics and implicit bias towards low rank. *Applied and Computational Harmonic Analysis*, 68:101595, 2024. doi: 10.1016/j.acha.2023.101595.

Yatin Dandi, Florent Krzakala, Bruno Loureiro, Luca Pesce, and Ludovic Stephan. How two-layer neural networks learn, one (giant) step at a time. *Journal of Machine Learning Research*, 25(349): 1–65, 2024.

Clémentine C. J. Dominé, Lukas Braun, James E. Fitzgerald, and Andrew M. Saxe. Exact learning dynamics of deep linear networks with prior knowledge. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(11):114004, 2023. doi: 10.1088/1742-5468/ad01b8.

Omer Elkabetz and Nadav Cohen. Continuous vs. discrete optimization of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4947–4960, 2021.

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, 2019. URL https://arxiv.org/abs/1904.13262.

Daniel Gissin, Shai Shalev-Shwartz, and Amit Daniely. The implicit bias of depth: How incremental learning drives generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1lj0nNFwB.

Reinhard Heckel and Fatih Furkan Yilmaz. Early stopping in deep networks: Double descent and how to eliminate it. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tlV90jvZbw.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

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.

Liwei Jiang, Yudong Chen, and Lijun Ding. Algorithmic regularization in model-free overparametrized asymmetric matrix factorization. *SIAM Journal on Mathematics of Data Science*, 5 (3):723–744, 2023. doi: 10.1137/22M1519833.

Dimitris Kalimeris, Gal Kaplun, Preetum Nakkiran, Benjamin Edelman, Tristan Yang, Boaz Barak, and Haofeng Zhang. SGD on neural networks learns functions of increasing complexity. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/b432f34c5a997c8e7c806a895ecc5e25-Abstract.html.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Akshay Kumar and Jarvis Haupt. Directional convergence near small initializations and saddles in two-homogeneous neural networks. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=hfrPag75Y0.

Akshay Kumar and Jarvis Haupt. Towards understanding gradient flow dynamics of homogeneous neural networks beyond the origin. *Journal of Machine Learning Research*, 26(269):1–83, 2025. URL https://jmlr.org/papers/v26/25-1089.html.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

James R. Lucas, Juhan Bae, Michael R. Zhang, Stanislav Fort, Richard Zemel, and Roger B. Grosse. On monotonic linear interpolation of neural network parameters. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 7168–7179. PMLR, 2021. URL https://proceedings.mlr.press/v139/lucas21a.html.

Kaifeng Lyu, Jikai Jin, Zhiyuan Li, Simon S. Du, Jason D. Lee, and Wei Hu. Dichotomy of early and late phase implicit biases can provably induce grokking. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=XsHqr9dEGH.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes ReLU network features. *arXiv preprint arXiv:1803.08367*, 2018. URL https://arxiv.org/abs/1803.08367.

Hancheng Min, Enrique Mallada, and René Vidal. Early neuron alignment in two-layer ReLU networks with small initialization. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QibPzdVrRu.

Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomization and Probabilistic Techniques in Algorithms and Data Analysis*. Cambridge University Press, 2 edition, 2017.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=B1g5sA4twr.

Gabin Maxime Nguegnang, Holger Rauhut, and Ulrich Terstiege. Convergence of gradient descent for learning linear neural networks. *Advances in Continuous and Discrete Models*, 2024(1):23, 2024. doi: 10.1186/s13662-023-03797-x.

Amanda Olmin and Fredrik Lindsten. Towards understanding epoch-wise double descent in two-layer linear neural networks. *arXiv preprint arXiv:2407.09845*, 2024. URL https://arxiv.org/abs/2407.09845.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C. Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1256–1272, 2021.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022. URL https://arxiv.org/abs/2201.02177.

Martin Raab and Angelika Steger. Balls into Bins—a simple and tight analysis. In *Randomization and Approximation Techniques in Computer Science*, volume 1518 of *Lecture Notes in Computer Science*, pp. 159–170. Springer, 1998. doi: 10.1007/3-540-49543-6_13.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *International Conference on Learning Representations*, 2014.

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. doi: 10.1073/pnas.1820226116.

Johannes Schneider. Learning in NN: From fitting most to fitting a few. *Neural Computing and Applications*, 37:23423–23446, 2025. doi: 10.1007/s00521-025-11528-4.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J. Gordon. An empirical study of example forgetting during deep neural network learning. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1812.05159.

Xiang Wang, Annie N. Wang, Mo Zhou, and Rong Ge. Plateau in monotonic linear interpolation—a "biased" view of loss landscape for deep networks. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=z289SIQOQna.

Blake Woodworth, Suriya Gunasekar, Jason D. Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Proceedings of the Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 3635–3673, 2020. URL https://proceedings.mlr.press/v125/woodworth20a.html.

Yedi Zhang, Andrew Saxe, and Peter E. Latham. Saddle-to-saddle dynamics explains a simplicity bias across neural network architectures. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Vit5M0G5Gb.

CONTENTS

1 Introduction **1**

2 Phenomenology **2**

3 A heuristic picture of the accuracy well **4**

4 Accuracy Dynamics in Deep Linear Networks under Vanishing Initialization **6**

5 Discussion **9**

A Related Work **15**

A.1 Small initialization and stagewise dynamics in linear networks 15

A.2 Vanishing initialization and early feature dynamics beyond linear models 16

A.3 Accuracy dynamics and neighboring non-monotone phenomena 17

B Accuracy Dynamics in Deep Linear Networks under Vanishing Initialization **18**

B.1 Direct linear optimization: $L = 1$ 19

B.2 Two-layer linear networks: $L = 2$ 25

B.3 Deeper linear networks: $L \geq 3$ 73

C Phenomenology and Experimental Details **92**

C.1 Controlled unary-operation experiments 92

C.2 Accuracy wells in fully connected networks on MNIST 95

C.3 Accuracy wells in a convolutional network on CIFAR-10 103

A Related Work

A.1 Small initialization and stagewise dynamics in linear networks

Deep linear networks provide a classical setting in which gradient-based learning separates into distinct dynamical stages. In the spectrally aligned setting, Saxe et al. (2014) derive exact gradient-flow solutions in which different singular modes are learned on different time scales, producing plateaus and rapid transitions. Building on these dynamics, Saxe et al. (2019) show that unequal mode-learning time scales can also produce non-monotone prediction-level behavior, including transient "illusory correlations" in which individual item–property predictions first strengthen and later reverse. This provides a particularly close precedent for the predictive effects studied here. More general exact analyses extend the deep-linear picture beyond specially aligned initializations and characterize the accompanying evolution of representations and alignment (Dominé et al., 2023). These works establish spectral organization as a central description of deep-linear learning. Our focus is specifically on the induced multiclass classifier and its aggregate top-one accuracy: we study how an accurate class ordering can emerge at microscopic predictor scale and subsequently be destroyed by further spectral concentration before later recovery.

Small or infinitesimal initialization makes this stagewise structure particularly sharp. Gidel et al. (2019) show that vanishing-initialization linear-network dynamics can successively approximate reduced-rank regression solutions of increasing rank, while Gissin et al. (2020) formalize incremental learning in deep linear networks and show that depth substantially relaxes the initialization-scale requirements needed for successive learning of increasingly complex solutions. In matrix factorization, Li et al. (2021) connect infinitesimal-initialization gradient flow to a greedy low-rank learning procedure. Related results show sequential recovery of principal components under small random initialization (Jiang et al., 2023), and Chou et al. (2024) rigorously identify intermediate time intervals in deep matrix factorization where the iterate has effective rank close to that of low-rank projections of the target. Thus preferential growth of spectral components and passage through effectively low-rank predictors are well-established features of small-initialization dynamics.

Our results reveal a classifier-level stage that precedes this familiar spectral concentration. For the two-factor model, at sufficiently early microscopic times all target singular modes grow, to leading order, with the relative weights needed to reconstruct the full target direction. After finite-width fluctuations self-average, this *target-driven organization* can already produce the classifier of the full target while both the predictor norm and the change in loss remain asymptotically vanishing. Continued microscopic evolution then amplifies different singular modes at unequal rates, eventually concentrating the normalized predictor onto its fastest-growing singular subspace. Consequently, an intermediate low-rank stage need not represent improvement from the perspective of classification: it can destroy class ordering that had already formed during the earlier target-driven stage. We refer to this transient loss of predictive diversity as a *degeneracy bottleneck*. In this sense, our results complement the incremental-learning literature by resolving these stages at the level of aggregate multiclass accuracy and by showing that spectral concentration can destroy a class ordering that had already emerged during an earlier microscopic stage.

The saddle-to-saddle viewpoint provides another closely related description of small-initialization dynamics. For deep linear networks, Jacot et al. (2021) characterize the first escape from the rank-zero saddle and conjecture a subsequent sequence of increasing-rank saddles. More recent work develops this picture across a broader class of neural architectures (Zhang et al., 2026). Our deep-linear result fits naturally into this framework but identifies an additional, earlier predictive stage. Before the leading rank-one mode governs the saddle escape, target-driven organization can already establish the classifier of the full target at microscopic scale. The subsequent rank-one escape can therefore act as a degeneracy bottleneck, replacing an already accurate classifier by the classifier of the leading singular component. Thus, from the viewpoint of classification, escape from the origin need not mark the first emergence of useful predictive structure.

The final recovery stage raises a separate convergence question. In the two-factor case, available convergence theory implies that for almost every initialization gradient flow reaches a global minimizer (Bah et al., 2022). We use this endpoint result to turn the microscopic target-driven-organization–bottleneck transition into a complete rise–collapse–recovery trajectory: after spectral concentration degrades the classifier, the remaining target directions are eventually recovered and the full target classifier is restored permanently. At greater depth, available convergence results do not

in general determine whether generic Gaussian initialization converges to a predictor with the rank and classification properties required for such a recovery (Nguegnang et al., 2024). Accordingly, our generic $L \geq 3$ theory establishes target-driven organization and the subsequent degeneracy bottleneck, while the later recruitment-and-recovery mechanism is proved separately under spectrally aligned dynamics.

A.2 Vanishing initialization and early feature dynamics beyond linear models

Small initialization also plays a fundamental role in nonlinear networks. The lazy-training analysis of Chizat et al. (2019) and the kernel–rich regimes of Woodworth et al. (2020) show that initialization scale and parameterization can determine whether training remains close to a fixed linearization or instead develops substantial feature learning. Our use of vanishing initialization has a different emphasis. A small predictor norm does not by itself imply lazy dynamics; rather, we exploit the early small-output regime to resolve changes in class ordering that can occur while the predictor and the loss remain nearly unchanged.

A growing literature shows that substantial task-dependent organization can already occur in this early regime. Maennel et al. (2018) describe the tendency of ReLU neurons at small initialization to concentrate along a small set of data-dependent directions. Atanasov et al. (2022) identify *silent alignment*, in which the directional structure of the neural tangent kernel becomes increasingly adapted to the task before appreciable loss decay, and demonstrate related behavior in nonlinear networks. Rigorous analyses of shallow ReLU networks likewise establish early directional alignment and staged dynamics under small initialization and structured data assumptions (Boursier et al., 2022; Min et al., 2024). These results show that substantial task-dependent organization can emerge before ordinary loss convergence.

This behavior is closely related to what we call *target-driven organization*. Our emphasis, however, is on the classifier induced by this early structure. In our linear theory, target-driven organization can already establish the classifier of the full target while both the predictor and the loss change remain asymptotically vanishing. Subsequent dynamics can then temporarily destroy this class ordering rather than merely refine it. This distinction is also conceptually related to Boursier & Flammarion (2025), who show that early alignment can impose a strong simplicity bias and, in suitable settings, steer an overparameterized two-layer ReLU network toward a suboptimal stationary point rather than a global minimizer. Relatedly, Pezeshki et al. (2021) identify *gradient starvation*, in which optimization preferentially captures a subset of predictive features while other useful features remain underlearned. This feature imbalance is close in spirit to our degeneracy bottleneck, although the mechanisms and dynamical claims are distinct.

Recent work further shows that stagewise dynamics beyond the initial alignment phase are not specific to linear factorization. Kumar & Haupt (2024) establish directional convergence near small initialization and near certain saddles for two-homogeneous neural networks, while Kumar & Haupt (2025) follow homogeneous gradient flow beyond its escape from the origin and characterize, under suitable assumptions, the subsequent saddle reached by the trajectory. For deep ReLU networks, Bantzis et al. (2026) characterize optimal escape directions from the origin and show that they exhibit a pronounced low-rank bias in the deeper layers, motivating a corresponding saddle-to-saddle picture. Zhang et al. (2026) provide a broader account of progressively more complex saddle-to-saddle dynamics across several neural architectures.

Together, these works establish that directional selection, low-rank structural bias, and saddle-to-saddle progression extend well beyond deep linear networks. Our organization–bottleneck–recruitment picture concerns a complementary classifier-level question: whether early target-driven organization produces a useful class ordering, whether subsequent concentration of the dynamics onto a smaller effective set of predictive directions can destroy that ordering, and whether later recruitment of additional directions can restore it. Our linear theory proves concrete instances of this mechanism. For the nonlinear networks studied experimentally here, we regard this broader interpretation as a hypothesis rather than a consequence of existing feature-alignment or saddle-to-saddle theory.

A.3 ACCURACY DYNAMICS AND NEIGHBORING NON-MONOTONE PHENOMENA

Non-monotone classification behavior during training has several important precedents, although the observables and mechanisms considered in prior work differ from the aggregate training-accuracy dynamics studied here.

At the level of individual examples, Toneva et al. (2019) show that training examples can undergo *forgetting events*, changing from correct to incorrect classification during ordinary single-task training. This establishes that correctness need not evolve monotonically even when the training set and objective are fixed. Our object is different: rather than tracking individual examples, we study a coherent non-monotone trajectory of the aggregate classifier, in which a large fraction of the training set can become correctly classified, subsequently lose that classification, and later recover.

A closer analogue is provided by Schneider (2025). They show empirically that geometrically selected atypical subsets can lose accuracy during an intermediate learning phase and later recover, and analyze the associated competition between typical and atypical examples through their losses and learning dynamics. Their theoretical results do not establish a non-monotone trajectory of the aggregate top-one training accuracy itself. Our mechanism is also different: in our linear theory, target-driven organization can first establish the classifier of the full target at microscopic predictor scale, after which unequal growth of predictive directions can concentrate the predictor onto a smaller effective subspace and produce a *degeneracy bottleneck*. Thus the deterioration occurs through changes in the spectral composition of the predictor rather than through competition between different subsets of examples.

Another neighboring phenomenon is epoch-wise double descent, in which test or generalization error decreases, subsequently rises, and then decreases again as training proceeds (Nakkiran et al., 2020). Analytical accounts relate this behavior to components learned on different time scales (Heckel & Yilmaz, 2021). Particularly close mathematically is Olmin & Lindsten (2024), who analyze epoch-wise double descent in two-layer linear neural networks and derive conditions for non-monotone generalization MSE from spectrally decoupled mode dynamics. These mechanisms are related to the unequal spectral growth that appears in our linear analysis, but the observables are different: epoch-wise double descent concerns generalization loss or error, whereas we follow top-one class ordering directly. In our unary models, the complete training population can undergo the microscopic transition $0 \to 1 \to 0$ before appreciable loss decay, so neither a train–test split nor a generalization effect is required.

A complementary line of work studies *monotonic linear interpolation* (MLI), namely the behavior along the line segment connecting an initialization to a trained solution. Lucas et al. (2021) provide broad empirical evidence for monotonic loss interpolation and derive sufficient conditions under mean-squared error, while Wang et al. (2023) study both loss and accuracy along interpolation paths and emphasize the role of biases and extended plateaus. Our $L = 1$ analysis intersects this literature in a particularly rigid setting: under isotropic inputs, the actual gradient-flow trajectory in predictor space is itself exactly the line segment from the initialization to the least-squares predictor $\mathbf{T}$, rather than a post-hoc interpolation between two parameter vectors. This exact structure allows us to resolve the classifier along the optimization trajectory itself. Agreement with the terminal least-squares classifier evolves monotonically, which sharply constrains any possible accuracy well; when the terminal classifier is perfect, non-monotone accuracy is ruled out altogether. Under vanishing initialization, all remaining classifier changes are confined to the early microscopic regime in which the predictor and the loss have barely changed.

Related work on stagewise prediction provides another useful contrast. Kalimeris et al. (2019) find that an early simple classifier learned by SGD is largely retained as training subsequently acquires more complex structure. Our results establish the complementary possibility: an early classifier can already be highly accurate, yet continued optimization can temporarily destroy its class ordering before later recovery.

Grokking provides a particularly relevant neighboring example of widely separated predictive time scales. In grokking, training performance typically improves early while generalization emerges only much later (Power et al., 2022), and theoretical work has related this delayed transition to distinct implicit biases operating at different stages of training (Lyu et al., 2024). The accuracy well differs in an important way: it contains an intermediate deterioration of the training classifier itself, and its early rise and collapse can both occur before substantial progress in the optimized loss. Thus

Table 1: Summary of the deep-linear accuracy theory. "Microscopic" means that the predictor norm and loss change vanish uniformly up to the stated time. Recovery statements are for fixed $\sigma > 0$.

Depth	Setting	Result
$L = 1$	General	$\exists\, \tau_* = O(\sigma)$ such that $\mathcal{A}(\mathbf{M}(\tau)) = A_{\mathbf{T}}$ for all $\tau \geq \tau_*$. Any accuracy drawdown is at most $1 - A_{\mathbf{T}}$. All classifier dynamics occur in a microscopic $O(\sigma)$ boundary layer.
$L = 2$	General	As $\sigma \downarrow 0$, then $H \to \infty$, $\exists\, \tau_{\mathrm{org}} < \tau_{\mathrm{bot}}$ such that, with probability tending to one, $\mathcal{A}(\mathbf{M}(\tau_{\mathrm{org}})) = A_{\mathbf{T}}$ and $\mathcal{A}(\mathbf{M}(\tau_{\mathrm{bot}})) = A_{k_\star}$. The trajectory through τ_{bot} is microscopic; for fixed $\sigma > 0$, it eventually recovers permanently to $A_{\mathbf{T}}$.
	Random unary, $H/p \to \alpha$	As $\sigma \downarrow 0$, then $p, H \to \infty$, $\exists\, \tau_{\mathrm{org}} < \tau_{\mathrm{bot}}$ such that $\mathcal{A}(\mathbf{M}(\tau_{\mathrm{org}})) \to 1$ and $\mathcal{A}(\mathbf{M}(\tau_{\mathrm{bot}})) \to 0$ in probability. The trajectory is microscopic through τ_{bot}. If $\alpha > 1 - e^{-1}$, fixed-σ recovery to 1 is permanent with probability tending to one.
	Permutation, $H/p \to \alpha$	As $\sigma \downarrow 0$, then $p, H \to \infty$, accuracy is nondecreasing with probability tending to one and reaches 1 during the microscopic regime, remaining 1 thereafter.
$L \geq 3$	Gaussian, $s_1 > s_2$	As $\sigma \downarrow 0$, then $H \to \infty$, $\exists\, \tau_{\mathrm{org}} < \tau_{\mathrm{bot}}$ such that $\mathcal{A}(\mathbf{M}(\tau_{\mathrm{org}})) = A_{\mathbf{T}}$ and $\mathcal{A}(\mathbf{M}(\tau_{\mathrm{bot}})) = A_1$. The trajectory through τ_{bot} is microscopic. The probability tends to 1 for $L = 3$, and is asymptotically at least $1/2$ for $L > 3$. Generic recovery remains open.
	Spectrally aligned	If a leading-mode bottleneck occurs, its accuracy is A_1. Subsequent mode recruitment gives $\mathbf{M}(\tau) \to \mathbf{T}$, hence permanent recovery to $A_{\mathbf{T}}$. Thus $A_{\mathbf{T}} > A_1$ yields a bottleneck followed by recovery.

the separation of time scales concerns the evolution of class ordering within the training dynamics, rather than only delayed generalization.

Thus our contribution is not the general observation that predictive quantities can evolve non-monotonically; such behavior already appears, for example, in the illusory-correlation dynamics of Saxe et al. (2019), as well as in the neighboring phenomena discussed above. We identify a regime in which the *aggregate training classifier* undergoes a pronounced rise–collapse–recovery along the optimization trajectory on a fixed training set, with the first two stages occurring at microscopic predictor scale and essentially constant loss. Our linear theory further identifies a concrete mechanism for these stages: target-driven organization can first establish the class ordering of the full target, after which unequal spectral growth can produce a degeneracy bottleneck that destroys this ordering, followed eventually by recruitment and recovery. The direct $L = 1$ dynamics provide a complementary control: agreement with the terminal least-squares classifier is monotone, any accuracy drawdown is bounded by its terminal error, and all classifier changes under vanishing initialization are confined to an $O(\sigma)$ boundary layer. Factorization therefore does not merely change the rate of approach to the same linear function class; it can enable qualitatively richer aggregate classifier trajectories that are impossible, or sharply restricted, under direct linear optimization.

B Accuracy Dynamics in Deep Linear Networks under Vanishing Initialization [Main Section 4]

This appendix develops the linear-network theory underlying the results in Section 4. Our main object is not only the evolution of the mean-squared training loss, but the accompanying evolution of classification accuracy. In particular, we ask how the parameterization depth of a linear predictor can change its accuracy dynamics, even though the realized function class remains linear.

We consider a supervised classification dataset

$$\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N, \qquad \mathbf{x}_i \in \mathbb{R}^p, \qquad \mathbf{y}_i = \mathbf{e}_{c_i} \in \mathbb{R}^C, \qquad c_i \in [C],$$

where $\mathbf{e}_{c_i}$ is the one-hot vector corresponding to the correct class c_i. A depth-L linear network has the form

$$f(\mathbf{x}) = \mathbf{W}_L \mathbf{W}_{L-1} \cdots \mathbf{W}_1 \mathbf{x} = \mathbf{M}\mathbf{x}, \tag{23}$$

where

$$\mathbf{M} = \mathbf{W}_L \mathbf{W}_{L-1} \cdots \mathbf{W}_1 \in \mathbb{R}^{C \times p} \tag{24}$$

is the end-to-end predictor. The factors are trained by gradient flow on the mean-squared error

$$\mathcal{L} = \frac{1}{2N} \sum_{i=1}^{N} \|\mathbf{M}\mathbf{x}_i - \mathbf{y}_i\|_2^2. \tag{25}$$

Classification depends only on the ordering of the C output scores. We use the strict training accuracy

$$\mathcal{A}(\mathbf{M}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left\{ (\mathbf{M}\mathbf{x}_i)_{c_i} > \max_{c \neq c_i} (\mathbf{M}\mathbf{x}_i)_c \right\}, \tag{26}$$

so ties are counted as errors. Thus the loss and the accuracy probe very different aspects of the same trajectory: the loss depends on the magnitudes of the prediction errors, whereas the accuracy depends only on whether the correct score exceeds every competing score. At small initialization, these two observables can evolve on parametrically different scales. A predictor whose norm is still vanishingly small can already have nontrivial or even perfect accuracy, and subsequent changes in the relative strengths of its predictive components can alter the class ordering while producing little visible change in the loss.

Unless stated otherwise, the trainable factors are initialized independently with centered Gaussian entries of variance σ^2, and we are interested in the small-initialization regime $\sigma \downarrow 0$. Although the end-to-end map $\mathbf{x} \mapsto \mathbf{M}\mathbf{x}$ belongs to the same linear function class for every depth L, its gradient-flow dynamics depends strongly on how M is parameterized. We therefore separate three regimes. For $L = 1$, the end-to-end matrix is optimized directly and the dynamics can be solved explicitly. For $L = 2$, factorization already creates nontrivial microscopic spectral dynamics, which can be analyzed in considerable detail. For $L > 2$, the dynamics is generally less tractable, but depth introduces an additional mechanism for strongly separating the growth of different predictive modes.

The two-factor case will be particularly important. Under isotropic inputs, its microscopic dynamics can be diagonalized in the singular coordinates of an effective target matrix. This separates two roles that will recur throughout the analysis: singular values determine the rates at which different modes are amplified, while singular vectors determine how those modes affect the classification margins. Unary-operation learning provides an especially tractable instance of this general theory: its target matrix has explicit singular vectors and singular values, allowing the spectral dynamics to be converted into precise statements about classification accuracy. This leads both to rigorous accuracy wells for heterogeneous unary targets and to a complementary permanence result for flat-spectrum permutation targets.

B.1 Direct linear optimization: $L = 1$

We first consider depth $L = 1$, where the end-to-end matrix $\mathbf{M}$ is optimized directly. Assume that the empirical input second-moment matrix is isotropic,

$$\mathbf{G} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top = \gamma \mathbf{I}_p, \qquad \gamma > 0, \tag{27}$$

and define the effective target matrix

$$\mathbf{T} = \frac{1}{N\gamma} \sum_{i=1}^{N} \mathbf{y}_i \mathbf{x}_i^\top. \tag{28}$$

Expanding the squared loss gives

$$\begin{aligned} \mathcal{L}(\mathbf{M}) &= \frac{1}{2N} \sum_{i=1}^{N} \|\mathbf{M}\mathbf{x}_i - \mathbf{y}_i\|_2^2 \\ &= \frac{\gamma}{2} \|\mathbf{M}\|_{\mathrm{F}}^2 - \gamma \langle \mathbf{M}, \mathbf{T} \rangle_{\mathrm{F}} + \mathcal{L}_{\mathrm{const}} \\ &= \frac{\gamma}{2} \|\mathbf{M} - \mathbf{T}\|_{\mathrm{F}}^2 + \mathcal{L}(\mathbf{T}), \end{aligned} \tag{29}$$

where the final term is independent of $\mathbf{M}$. Thus $\mathbf{T}$ is the unique global minimizer of the squared loss over unconstrained linear predictors.

After the deterministic time rescaling

$$\tau = \gamma t, \tag{30}$$

gradient flow satisfies

$$\dot{\mathbf{M}} = \mathbf{T} - \mathbf{M}, \qquad \mathbf{M}(\tau) = e^{-\tau}\mathbf{M}(0) + (1 - e^{-\tau})\mathbf{T}. \tag{31}$$

Thus direct linear optimization moves the predictor along the line segment joining its initialization to the least-squares solution $\mathbf{T}$.

This places the $L = 1$ dynamics in direct connection with the literature on monotonic linear interpolation (MLI) (Lucas et al., 2021; Wang et al., 2023). Here, however, the linear interpolation is not a post-hoc path between initialization and a trained model: equation (31) shows that it is the exact gradient-flow trajectory in predictor space. This additional structure allows us to obtain pointwise guarantees for the classifier itself, including quantitative control of both the depth and the time scale of any possible nonmonotone accuracy transient.

The convergence $\mathbf{M}(\tau) \to \mathbf{T}$ alone is not sufficient to exclude nonmonotone accuracy dynamics. In particular, under vanishing initialization, a nonmonotone transient could in principle be confined to a correspondingly vanishing time scale. We therefore first identify a quantity whose dynamics are monotone at all times.

Suppose throughout this subsection that $\mathbf{T}$ has a unique top-scoring class on every training example, and define

$$\widehat{c}_i = \arg\max_{c \in [C]} (\mathbf{T}\mathbf{x}_i)_c. \tag{32}$$

We measure agreement with the classifier induced by $\mathbf{T}$ by

$$\mathrm{Agr}_{\mathbf{T}}(\mathbf{M}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left\{ (\mathbf{M}\mathbf{x}_i)_{\widehat{c}_i} > \max_{c \neq \widehat{c}_i} (\mathbf{M}\mathbf{x}_i)_c \right\}. \tag{33}$$

As in the definition of training accuracy, ties are counted as disagreement.

Lemma B.1 (Monotone agreement with the least-squares classifier)**.** *For every training example i, once the classifier induced by $\mathbf{M}(\tau)$ agrees with the classifier induced by $\mathbf{T}$, this agreement is never subsequently lost. Consequently, for every $0 \le \tau_1 \le \tau_2$,*

$$\mathrm{Agr}_{\mathbf{T}}(\mathbf{M}(\tau_2)) \ge \mathrm{Agr}_{\mathbf{T}}(\mathbf{M}(\tau_1)), \tag{34}$$

and

$$\lim_{\tau \to \infty} \mathrm{Agr}_{\mathbf{T}}(\mathbf{M}(\tau)) = 1. \tag{35}$$

Proof. Fix a training example i, and suppose that at some time τ_1 the classifier induced by $\mathbf{M}(\tau_1)$ agrees with that induced by $\mathbf{T}$. Then, for every $c \neq \widehat{c}_i$,

$$(\mathbf{M}(\tau_1)\mathbf{x}_i)_{\widehat{c}_i} - (\mathbf{M}(\tau_1)\mathbf{x}_i)_c > 0. \tag{36}$$

By the definition of $\widehat{c}_i$,

$$(\mathbf{T}\mathbf{x}_i)_{\widehat{c}_i} - (\mathbf{T}\mathbf{x}_i)_c > 0. \tag{37}$$

For every $\tau_2 \ge \tau_1$, the semigroup property of (31) gives

$$\mathbf{M}(\tau_2) = e^{-(\tau_2 - \tau_1)}\mathbf{M}(\tau_1) + \left(1 - e^{-(\tau_2 - \tau_1)}\right)\mathbf{T}. \tag{38}$$

Hence, for each $c \neq \widehat{c}_i$,

$$\begin{aligned} &(\mathbf{M}(\tau_2)\mathbf{x}_i)_{\widehat{c}_i} - (\mathbf{M}(\tau_2)\mathbf{x}_i)_c \\ &\quad = e^{-(\tau_2 - \tau_1)} \left[(\mathbf{M}(\tau_1)\mathbf{x}_i)_{\widehat{c}_i} - (\mathbf{M}(\tau_1)\mathbf{x}_i)_c \right] \\ &\qquad + \left(1 - e^{-(\tau_2 - \tau_1)}\right) \left[(\mathbf{T}\mathbf{x}_i)_{\widehat{c}_i} - (\mathbf{T}\mathbf{x}_i)_c \right]. \end{aligned} \tag{39}$$

The right-hand side is a convex combination of two strictly positive quantities and is therefore strictly positive. Thus, once agreement with $\mathbf{T}$ is achieved on example i, it persists for all later times. Summing the corresponding pointwise indicators proves (34).

Finally, $\mathbf{M}(\tau) \to \mathbf{T}$ as $\tau \to \infty$. Since $\mathbf{T}$ has a unique top-scoring class on every training example and the training set is finite, every example eventually has the same predicted class under $\mathbf{M}(\tau)$ and $\mathbf{T}$. This proves (35). □

The lemma shows that agreement with the terminal least-squares classifier evolves in only one direction: individual training examples may enter the agreement set, but can never leave it. In particular, full agreement is an absorbing state of the classifier dynamics.

Corollary B.2 (Permanent stabilization after full classifier agreement)**.** *Suppose that, at some time* $\tau_0 \geq 0$*,*

$$\mathrm{Agr}_{\mathbf{T}}(\mathbf{M}(\tau_0)) = 1. \tag{40}$$

Then, for every $\tau \geq \tau_0$*,*

$$\mathrm{Agr}_{\mathbf{T}}(\mathbf{M}(\tau)) = 1, \qquad \mathcal{A}(\mathbf{M}(\tau)) = \mathcal{A}(\mathbf{T}). \tag{41}$$

Thus, once the current classifier agrees with the least-squares classifier on every training example, its training accuracy has reached its terminal value and can never subsequently change.

Proof. By Lemma B.1, agreement with $\mathbf{T}$ on each individual training example, once achieved, is never lost. Hence $\mathrm{Agr}_{\mathbf{T}}(\mathbf{M}(\tau)) = 1$ for every $\tau \geq \tau_0$. The two classifiers therefore make the same prediction on every training example, which implies

$$\mathcal{A}(\mathbf{M}(\tau)) = \mathcal{A}(\mathbf{T}) \tag{42}$$

for all $\tau \geq \tau_0$. □

The relevant monotone quantity under direct linear optimization is therefore agreement with the terminal least-squares classifier, rather than training accuracy itself. Under the coupled Gaussian initialization

$$\mathbf{M}^{(\sigma)}(0) = \sigma\mathbf{Z}, \tag{43}$$

where the entries of $\mathbf{Z}$ are independent standard Gaussian random variables, the C initial scores on every fixed training example are exchangeable. Hence

$$\mathbb{E}\left[\mathrm{Agr}_{\mathbf{T}}(\mathbf{M}^{(\sigma)}(0))\right] = \frac{1}{C}, \tag{44}$$

independently of σ, whereas the agreement increases monotonically to one along every realization of the trajectory.

The next result quantifies the time scale on which this convergence of the classifier takes place. It shows that, in the vanishing-initialization limit, all nontrivial classifier dynamics are compressed into an initial boundary layer during which both the predictor itself and the change in loss remain vanishingly small.

Theorem B.3 (Boundary-layer stabilization under vanishing initialization)**.** *Consider the coupled initialization*

$$\mathbf{M}^{(\sigma)}(0) = \sigma\mathbf{Z}, \qquad \sigma > 0, \tag{45}$$

with fixed $\mathbf{Z}$*. For every training example* i *and every* $c \neq \widehat{c}_i$*, define the target margin*

$$\Delta_{ic} = (\mathbf{T}\mathbf{x}_i)_{\widehat{c}_i} - (\mathbf{T}\mathbf{x}_i)_c > 0 \tag{46}$$

and the corresponding initialization margin

$$\zeta_{ic} = (\mathbf{Z}\mathbf{x}_i)_{\widehat{c}_i} - (\mathbf{Z}\mathbf{x}_i)_c. \tag{47}$$

Define the finite constant

$$s_* = 1 + \max_{1 \leq i \leq N} \max_{c \neq \widehat{c}_i} \left(-\frac{\zeta_{ic}}{\Delta_{ic}}\right)_+, \tag{48}$$

where $(a)_+ = \max\{a, 0\}$*, and let*

$$\tau_*^{(\sigma)} = \log(1 + \sigma s_*). \tag{49}$$

Then

$$\mathrm{Agr}_{\mathbf{T}}\left(\mathbf{M}^{(\sigma)}(\tau)\right) = 1, \qquad \mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau)\right) = \mathcal{A}(\mathbf{T}) \tag{50}$$

for every

$$\tau \geq \tau_*^{(\sigma)}. \tag{51}$$

Moreover,

$$\tau_*^{(\sigma)} = \sigma s_* + O(\sigma^2) \qquad \text{as } \sigma \downarrow 0. \tag{52}$$

Throughout the entire boundary layer $0 \leq \tau \leq \tau_*^{(\sigma)}$, the predictor remains at the initialization scale:

$$\sup_{0\leq\tau\leq\tau_*^{(\sigma)}} \left\|\mathbf{M}^{(\sigma)}(\tau)\right\|_{\mathrm{F}} = O(\sigma), \tag{53}$$

and its displacement from initialization also satisfies

$$\sup_{0\leq\tau\leq\tau_*^{(\sigma)}} \left\|\mathbf{M}^{(\sigma)}(\tau) - \mathbf{M}^{(\sigma)}(0)\right\|_{\mathrm{F}} = O(\sigma). \tag{54}$$

Finally, only a vanishing fraction of the total available loss decrease has occurred by the stabilization time:

$$\frac{\mathcal{L}(\mathbf{M}^{(\sigma)}(0)) - \mathcal{L}(\mathbf{M}^{(\sigma)}(\tau_*^{(\sigma)}))}{\mathcal{L}(\mathbf{M}^{(\sigma)}(0)) - \mathcal{L}(\mathbf{T})} = 2\sigma s_* + O(\sigma^2). \tag{55}$$

In particular,

$$\frac{\mathcal{L}(\mathbf{M}^{(\sigma)}(0)) - \mathcal{L}(\mathbf{M}^{(\sigma)}(\tau_*^{(\sigma)}))}{\mathcal{L}(\mathbf{M}^{(\sigma)}(0))} = O(\sigma). \tag{56}$$

Thus all possible nontrivial accuracy dynamics are confined to a time interval of width $O(\sigma)$ in the rescaled time τ, or $O(\sigma/\gamma)$ in the original gradient-flow time t, while both the predictor and the relative loss change remain vanishingly small.

Proof. Using (31) and (45), the score difference between the class selected by $\mathbf{T}$ and a competing class is

$$\begin{aligned}\left(\mathbf{M}^{(\sigma)}(\tau)\mathbf{x}_i\right)_{\widehat{c}_i} - \left(\mathbf{M}^{(\sigma)}(\tau)\mathbf{x}_i\right)_c \\ = \sigma e^{-\tau}\zeta_{ic} + (1-e^{-\tau})\Delta_{ic} \\ = \sigma e^{-\tau}\left[\zeta_{ic} + \frac{e^{\tau}-1}{\sigma}\Delta_{ic}\right].\end{aligned} \tag{57}$$

Introduce the fast time variable

$$s(\tau) = \frac{e^{\tau}-1}{\sigma}. \tag{58}$$

By construction of s_*,

$$\zeta_{ic} + s_*\Delta_{ic} > 0 \tag{59}$$

for every i and every $c \neq \widehat{c}_i$. Since $s(\tau)$ is strictly increasing and

$$s(\tau_*^{(\sigma)}) = s_*, \tag{60}$$

all pairwise margins relative to the class $\widehat{c}_i$ are strictly positive at $\tau_*^{(\sigma)}$. Therefore

$$\mathrm{Agr}_{\mathbf{T}}\left(\mathbf{M}^{(\sigma)}(\tau_*^{(\sigma)})\right) = 1. \tag{61}$$

Corollary B.2 then implies that full agreement, and hence the terminal accuracy, persists for all later times. Expanding (49) gives

$$\tau_*^{(\sigma)} = \sigma s_* + O(\sigma^2). \tag{62}$$

The fast-time parametrization also gives the exact representation

$$\mathbf{M}^{(\sigma)}(\tau) = \frac{\sigma}{1+\sigma s(\tau)}\left[\mathbf{Z} + s(\tau)\mathbf{T}\right]. \tag{63}$$

For $0 \le \tau \le \tau_*^{(\sigma)}$, one has

$$0 \le s(\tau) \le s_*. \tag{64}$$

Therefore

$$\left\|\mathbf{M}^{(\sigma)}(\tau)\right\|_{\mathrm{F}} \le \sigma\left(\|\mathbf{Z}\|_{\mathrm{F}} + s_*\|\mathbf{T}\|_{\mathrm{F}}\right), \tag{65}$$

which proves (53). Moreover,

$$\mathbf{M}^{(\sigma)}(\tau) - \sigma\mathbf{Z} = \frac{\sigma s(\tau)}{1+\sigma s(\tau)}\left(\mathbf{T} - \sigma\mathbf{Z}\right), \tag{66}$$

which proves (54).

It remains to quantify the loss change. From (31),

$$\mathbf{M}^{(\sigma)}(\tau) - \mathbf{T} = e^{-\tau}\left(\mathbf{M}^{(\sigma)}(0) - \mathbf{T}\right). \tag{67}$$

Using the quadratic form of the loss in (29), we obtain the exact identity

$$\mathcal{L}(\mathbf{M}^{(\sigma)}(\tau)) - \mathcal{L}(\mathbf{T}) = e^{-2\tau}\left[\mathcal{L}(\mathbf{M}^{(\sigma)}(0)) - \mathcal{L}(\mathbf{T})\right]. \tag{68}$$

Consequently,

$$\frac{\mathcal{L}(\mathbf{M}^{(\sigma)}(0)) - \mathcal{L}(\mathbf{M}^{(\sigma)}(\tau_*^{(\sigma)}))}{\mathcal{L}(\mathbf{M}^{(\sigma)}(0)) - \mathcal{L}(\mathbf{T})} = 1 - e^{-2\tau_*^{(\sigma)}} = 1 - \frac{1}{(1+\sigma s_*)^2}. \tag{69}$$

Expanding the right-hand side yields

$$1 - \frac{1}{(1+\sigma s_*)^2} = 2\sigma s_* + O(\sigma^2), \tag{70}$$

which proves (55).

Finally, because the labels are one-hot,

$$\mathcal{L}(\mathbf{0}) = \frac{1}{2}, \tag{71}$$

and

$$\mathcal{L}(\mathbf{M}^{(\sigma)}(0)) = \frac{1}{2} + O(\sigma). \tag{72}$$

The denominator in (56) is therefore bounded away from zero for sufficiently small σ, while the loss decrease in the numerator is $O(\sigma)$ by (55). This proves (56). □

Theorem B.3 shows that vanishing initialization does not merely make classifier stabilization occur early in absolute time. Rather, the complete nontrivial classifier dynamics are squeezed into a boundary layer of width $O(\sigma)$ during which the end-to-end predictor itself remains only $O(\sigma)$ and the loss has completed only an $O(\sigma)$ fraction of its eventual decrease. Thus any accuracy well that may occur for $L = 1$ is necessarily confined to this vanishing initial layer.

Although agreement with $\mathbf{T}$ is monotone, training accuracy itself need not be: moving toward the prediction of $\mathbf{T}$ improves classification on examples that $\mathbf{T}$ classifies correctly, but may worsen it on examples that $\mathbf{T}$ misclassifies. The monotone agreement property nevertheless places a sharp bound on how large such a decrease can be.

Theorem B.4 [4.1] **(Accuracy well bound under direct linear optimization).**

Suppose that the least-squares predictor satisfies

$$\mathcal{A}(\mathbf{T}) = 1 - \varepsilon, \qquad 0 \le \varepsilon \le 1. \tag{73}$$

Then, for every pair of times $0 \le \tau_1 \le \tau_2$,

$$\mathcal{A}(\mathbf{M}(\tau_2)) \ge \mathcal{A}(\mathbf{M}(\tau_1)) - \varepsilon. \tag{74}$$

Equivalently,

$$\sup_{0\le\tau_1\le\tau_2}\left[\mathcal{A}(\mathbf{M}(\tau_1)) - \mathcal{A}(\mathbf{M}(\tau_2))\right]_+ \le \varepsilon. \tag{75}$$

Thus any accuracy well along the $L = 1$ trajectory has depth at most the classification error of the terminal least-squares predictor.

Proof. Partition the training examples according to whether $\mathbf{T}$ classifies them correctly:

$$\mathcal{G}_{\mathbf{T}} = \{i : \widehat{c}_i = c_i\}, \qquad \mathcal{B}_{\mathbf{T}} = \{i : \widehat{c}_i \neq c_i\}. \tag{76}$$

By assumption,

$$\frac{|\mathcal{G}_{\mathbf{T}}|}{N} = 1 - \varepsilon, \qquad \frac{|\mathcal{B}_{\mathbf{T}}|}{N} = \varepsilon. \tag{77}$$

For an example $i \in \mathcal{G}_{\mathbf{T}}$, agreement with the classifier of $\mathbf{T}$ is equivalent to correct classification, because $\widehat{c}_i = c_i$. By Lemma B.1, if such an example is correct at time τ_1, it remains correct at every later time $\tau_2 \geq \tau_1$. Therefore every correct-to-incorrect transition between τ_1 and τ_2 must occur on an example belonging to $\mathcal{B}_{\mathbf{T}}$.

There are only

$$|\mathcal{B}_{\mathbf{T}}| = \varepsilon N \tag{78}$$

such examples. Hence at most εN correct predictions can be lost between τ_1 and τ_2, and therefore

$$N\mathcal{A}(\mathbf{M}(\tau_2)) \geq N\mathcal{A}(\mathbf{M}(\tau_1)) - \varepsilon N. \tag{79}$$

This proves (74). Taking the supremum over $0 \leq \tau_1 \leq \tau_2$ yields (75). □

Combining Theorems B.3 and B.4 gives a quantitative description of every possible accuracy well under direct linear optimization. If $\mathcal{A}(\mathbf{T}) = 1 - \varepsilon$, then its vertical depth is at most ε, while under initialization $\mathbf{M}^{(\sigma)}(0) = \sigma\mathbf{Z}$ all nontrivial classifier dynamics, including any such well, are confined to a horizontal time scale $O(\sigma)$. Moreover, this transient occurs while the predictor remains at scale $O(\sigma)$ and while only an $O(\sigma)$ fraction of the total loss decrease has taken place.

The perfect-classification case gives an especially strong consequence.

Corollary B.5 (Perfect least-squares classification excludes accuracy wells)**.** *If*

$$\mathcal{A}(\mathbf{T}) = 1, \tag{80}$$

then the training accuracy is monotone nondecreasing:

$$0 \leq \tau_1 \leq \tau_2 \qquad \Longrightarrow \qquad \mathcal{A}(\mathbf{M}(\tau_2)) \geq \mathcal{A}(\mathbf{M}(\tau_1)). \tag{81}$$

Consequently, direct $L = 1$ gradient flow cannot exhibit an accuracy well at any time scale.

Proof. The claim follows immediately from Theorem B.4 by setting $\varepsilon = 0$. □

The condition $\mathcal{A}(\mathbf{T}) = 1$ should not be identified with ordinary linear separability. Linear separability guarantees only the existence of some linear predictor whose predicted classes are correct on all training examples, whereas $\mathbf{T}$ is the particular least-squares minimizer selected by the $L = 1$ dynamics. A linearly separable dataset need not satisfy $\mathcal{A}(\mathbf{T}) = 1$. A stronger sufficient condition is exact linear realizability of the labels: if there exists a matrix $\mathbf{M}_*$ such that

$$\mathbf{M}_*\mathbf{x}_i = \mathbf{y}_i \qquad \text{for every } i, \tag{82}$$

then the minimum squared loss is zero, so the least-squares solution also interpolates all labels and therefore satisfies $\mathcal{A}(\mathbf{T}) = 1$.

Thus the $L = 1$ dynamics are strongly constrained. Agreement with the terminal least-squares classifier evolves monotonically; all classifier changes are compressed into an $O(\sigma)$ initial boundary layer under vanishing initialization; and the maximal downward excursion of training accuracy is bounded by the terminal classification error. In particular, when the least-squares classifier is perfect, accuracy wells are impossible altogether. This contrasts with factorized linear networks, where the end-to-end predictor no longer follows a direct convex interpolation toward its terminal value, allowing a high-accuracy classifier to form and subsequently be destroyed during training.

B.2 TWO-LAYER LINEAR NETWORKS: $L = 2$

For $L = 2$, we write the network as[1]

$$f_{\mathbf{U},\mathbf{V}}(\mathbf{x}) = \mathbf{V}^\top \mathbf{U}\mathbf{x}, \qquad \mathbf{U} \in \mathbb{R}^{H\times p}, \qquad \mathbf{V} \in \mathbb{R}^{H\times C}, \tag{83}$$

with end-to-end predictor

$$\mathbf{M} = \mathbf{V}^\top \mathbf{U} \in \mathbb{R}^{C\times p}. \tag{84}$$

The empirical squared loss is

$$\mathcal{L}(\mathbf{U},\mathbf{V}) = \frac{1}{2N}\sum_{i=1}^{N} \left\|\mathbf{V}^\top \mathbf{U}\mathbf{x}_i - \mathbf{y}_i\right\|_2^2. \tag{85}$$

Introduce the empirical input second-moment matrix and input–target association matrix

$$\mathbf{G} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i\mathbf{x}_i^\top, \qquad \mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{y}_i\mathbf{x}_i^\top. \tag{86}$$

Throughout this subsection we assume isotropic inputs,

$$\mathbf{G} = \gamma \mathbf{I}_p, \qquad \gamma > 0, \tag{87}$$

and define the effective target

$$\mathbf{T} = \frac{\mathbf{S}}{\gamma}. \tag{88}$$

Then, up to a term independent of $\mathbf{U}, \mathbf{V}$,

$$\mathcal{L}(\mathbf{U},\mathbf{V}) = \frac{\gamma}{2}\left\|\mathbf{V}^\top \mathbf{U} - \mathbf{T}\right\|_{\mathrm{F}}^2 + \mathcal{L}_{\mathrm{const}}. \tag{89}$$

Writing t for the original gradient-flow time, we have

$$\frac{d\mathbf{U}}{dt} = \gamma \mathbf{V}\left(\mathbf{T} - \mathbf{V}^\top \mathbf{U}\right), \qquad \frac{d\mathbf{V}}{dt} = \gamma \mathbf{U}\left(\mathbf{T}^\top - \mathbf{U}^\top \mathbf{V}\right). \tag{90}$$

We therefore introduce the rescaled time

$$\tau = \gamma t. \tag{91}$$

Throughout the remainder of this subsection, all times are expressed in τ, and a dot denotes differentiation with respect to τ. In this time variable the gradient flow becomes

$$\dot{\mathbf{U}} = \mathbf{V}\left(\mathbf{T} - \mathbf{V}^\top \mathbf{U}\right), \qquad \dot{\mathbf{V}} = \mathbf{U}\left(\mathbf{T}^\top - \mathbf{U}^\top \mathbf{V}\right). \tag{92}$$

Let

$$\mathbf{T} = \mathbf{P}\mathbf{\Lambda}\mathbf{Q}^\top = \sum_{j=1}^{r} s_j \mathbf{p}_j \mathbf{q}_j^\top, \qquad s_1 \geq \cdots \geq s_r > 0, \qquad r = \mathrm{rank}(\mathbf{T}), \tag{93}$$

be a compact singular-value decomposition. Let $k_\star$ be the multiplicity of the largest singular value,

$$s_1 = \cdots = s_{k_\star} > s_{k_\star+1}, \qquad s_{r+1} := 0, \tag{94}$$

and define the complete top singular block

$$\mathbf{T}_{k_\star} := \sum_{j=1}^{k_\star} s_j \mathbf{p}_j \mathbf{q}_j^\top. \tag{95}$$

Because the entire singular subspace associated with s_1 is retained, $\mathbf{T}_{k_\star}$ is independent of the choice of singular-vector basis within that subspace.

We will repeatedly use the two accuracy levels

$$A_{\mathbf{T}} := \mathcal{A}(\mathbf{T}), \qquad A_{k_\star} := \mathcal{A}(\mathbf{T}_{k_\star}). \tag{96}$$

They describe the classifiers associated with the full effective target and with its complete top singular block. Below we prove that, under suitable large-width assumptions, there is an early time τ_{org} at which the network classifier agrees with $\mathbf{T}$, and a later bottleneck time $\tau_{\mathrm{bot}} > \tau_{\mathrm{org}}$ at which it agrees with $\mathbf{T}_{k_\star}$. In the realizable case, the full gradient flow eventually converges back to $\mathbf{T}$. Hence the relation between $A_{\mathbf{T}}$ and $A_{k_\star}$ determines whether the transitions established below constitute an accuracy well.

[1] In the general notation introduced above, the end-to-end predictor is $\mathbf{M} = \mathbf{W}_2\mathbf{W}_1$. Throughout this subsection we use the equivalent parametrization $\mathbf{W}_1 = \mathbf{U}$ and $\mathbf{W}_2 = \mathbf{V}^\top$, which simplifies the notation.

Ordering of the two accuracy levels. There is no universal ordering between $A_{\mathbf{T}}$ and $A_{k_\star}$. The truncated singular-value decomposition is optimal for matrix approximation in Frobenius or operator norm, but classification accuracy is determined by the ordering of class scores on individual examples. Consequently, subleading singular modes may either correct or spoil top-one decisions.

We give an explicit example in which $A_{\mathbf{T}} < A_{k_\star}$. Let $p = 2$, $C = 3$, and $N = 8$, and take the isotropic inputs

$$\mathbf{x}_j = \sqrt{2}\begin{pmatrix}\cos((j-1)\pi/4)\\ \sin((j-1)\pi/4)\end{pmatrix}, \qquad j = 1, \dots, 8. \tag{97}$$

The eight vectors are the equally spaced points on a circle of radius $\sqrt{2}$, and therefore

$$\frac{1}{8}\sum_{j=1}^{8}\mathbf{x}_j\mathbf{x}_j^\top = \mathbf{I}_2. \tag{98}$$

Assign the labels

$$(c_1, \dots, c_8) = (1, 1, 2, 1, 2, 2, 3, 2), \qquad \mathbf{y}_j = \mathbf{e}_{c_j}. \tag{99}$$

Using $\mathbf{T} = N^{-1}\sum_j \mathbf{y}_j\mathbf{x}_j^\top$, we obtain

$$\mathbf{T} = \frac{1}{8}\begin{pmatrix}\sqrt{2} & 2\\ -\sqrt{2} & \sqrt{2}-2\\ 0 & -\sqrt{2}\end{pmatrix}. \tag{100}$$

Indeed, the three rows are obtained by summing the inputs belonging to each class:

$$\sum_{j:c_j=1}\mathbf{x}_j = \begin{pmatrix}\sqrt{2}\\ 2\end{pmatrix}, \qquad \sum_{j:c_j=2}\mathbf{x}_j = \begin{pmatrix}-\sqrt{2}\\ \sqrt{2}-2\end{pmatrix}, \qquad \sum_{j:c_j=3}\mathbf{x}_j = \begin{pmatrix}0\\ -\sqrt{2}\end{pmatrix}. \tag{101}$$

For this matrix,

$$\mathbf{T}^\top\mathbf{T} = \frac{1}{64}\begin{pmatrix}4 & 4\sqrt{2}-2\\ 4\sqrt{2}-2 & 12-4\sqrt{2}\end{pmatrix}, \tag{102}$$

whose two eigenvalues are

$$s_{1,2}^2 = \frac{8-2\sqrt{2}\pm 2\sqrt{15-8\sqrt{2}}}{64}. \tag{103}$$

Thus

$$s_1 \approx 0.37524, \qquad s_2 \approx 0.14425, \tag{104}$$

so the largest singular value is simple and $k_\star = 1$.

Evaluating $\mathbf{T}\mathbf{x}_j$ on the eight training inputs gives the predicted classes

$$(1, 1, 1, 2, 2, 2, 3, 3). \tag{105}$$

Comparing these predictions with $(1, 1, 2, 1, 2, 2, 3, 2)$, the correctly classified examples are $j = 1, 2, 5, 6, 7$. Hence

$$A_{\mathbf{T}} = \frac{5}{8}. \tag{106}$$

The leading rank-one truncation is

$$\mathbf{T}_1 = s_1\mathbf{p}_1\mathbf{q}_1^\top \approx \begin{pmatrix}0.18046 & 0.24731\\ -0.09629 & -0.13196\\ -0.08417 & -0.11536\end{pmatrix}. \tag{107}$$

Its predicted classes on the same eight inputs are

$$(1, 1, 1, 1, 2, 2, 2, 2). \tag{108}$$

These are correct for $j = 1, 2, 4, 5, 6, 8$, and therefore

$$A_{k_\star} = A_1 = \frac{6}{8} = \frac{3}{4} > A_{\mathbf{T}}. \tag{109}$$

Table 2: Accuracy of the full effective target and its complete top singular block on centered and whitened training data. In all cases $k_\star = 1$.

Dataset	$A_{\mathbf{T}}$	$A_{k_\star}$	$A_{\mathbf{T}} > A_{k_\star}$
MNIST	0.859	0.209	Yes
Fashion-MNIST	0.831	0.200	Yes
KMNIST	0.770	0.196	Yes
CIFAR-10	0.487	0.180	Yes
CIFAR-100	0.442	0.019	Yes

Thus, in this example, adding the subleading singular component decreases classification accuracy even though it improves the matrix approximation to the full target.

The opposite ordering,

$$A_{\mathbf{T}} > A_{k_\star}, \tag{110}$$

is the one that produces the accuracy wells studied in this work. To verify that this ordering is not specific to the synthetic constructions below, we evaluate it on several standard image datasets after centering and whitening the training inputs, so that the empirical covariance satisfies the isotropy assumption of our theory. The results are shown in Table 2. In all five datasets the largest singular value is simple, $k_\star = 1$, and we find $A_{\mathbf{T}} > A_{k_\star}$.

The same ordering also holds for the heterogeneous random unary operations analyzed below: the full target has perfect accuracy, whereas the top singular block retains only the outputs with maximal occupancy and therefore loses the examples associated with lower occupancies. Flat-spectrum unary operations provide an important equality case. In particular, for a permutation all positive singular values are equal, so $k_\star = r$, $\mathbf{T}_{k_\star} = \mathbf{T}$, and

$$A_{k_\star} = A_{\mathbf{T}} = 1. \tag{111}$$

Microscopic linearization. We now analyze the vanishing-initialization dynamics generated by (92). The factor equations contain a linear term determined by $\mathbf{T}$ and a cubic term generated by the factorization. Since the factors are initialized at scale σ, the cubic term is suppressed by σ^2 after normalization and therefore vanishes on every fixed time interval as $\sigma \downarrow 0$.

To formulate this limit on a common probability space, sample independent matrices

$$\mathbf{A} \in \mathbb{R}^{H\times p}, \qquad \mathbf{B} \in \mathbb{R}^{H\times C}, \tag{112}$$

with independent standard Gaussian entries, and for every $\sigma > 0$ initialize the exact gradient flow by

$$\mathbf{U}^{(\sigma)}(0) = \sigma\mathbf{A}, \qquad \mathbf{V}^{(\sigma)}(0) = \sigma\mathbf{B}. \tag{113}$$

For every fixed σ, this is precisely independent Gaussian initialization with entrywise variance σ^2.

Define the normalized factors

$$\widehat{\mathbf{U}}^{(\sigma)} = \frac{\mathbf{U}^{(\sigma)}}{\sigma}, \qquad \widehat{\mathbf{V}}^{(\sigma)} = \frac{\mathbf{V}^{(\sigma)}}{\sigma}. \tag{114}$$

Substituting into (92) gives

$$\begin{aligned} \dot{\widehat{\mathbf{U}}}^{(\sigma)} &= \widehat{\mathbf{V}}^{(\sigma)}\mathbf{T} - \sigma^2\widehat{\mathbf{V}}^{(\sigma)}\widehat{\mathbf{V}}^{(\sigma)\top}\widehat{\mathbf{U}}^{(\sigma)}, \\ \dot{\widehat{\mathbf{V}}}^{(\sigma)} &= \widehat{\mathbf{U}}^{(\sigma)}\mathbf{T}^\top - \sigma^2\widehat{\mathbf{U}}^{(\sigma)}\widehat{\mathbf{U}}^{(\sigma)\top}\widehat{\mathbf{V}}^{(\sigma)}. \end{aligned} \tag{115}$$

Thus the nonlinear terms are explicitly multiplied by σ^2. The limiting system obtained as $\sigma \downarrow 0$ is

$$\dot{\mathbf{U}}_0 = \mathbf{V}_0\mathbf{T}, \qquad \dot{\mathbf{V}}_0 = \mathbf{U}_0\mathbf{T}^\top, \qquad \mathbf{U}_0(0) = \mathbf{A}, \qquad \mathbf{V}_0(0) = \mathbf{B}. \tag{116}$$

Using the singular-value decomposition (93), define the active coordinates

$$\mathbf{X}_0(\tau) = \mathbf{U}_0(\tau)\mathbf{Q}, \qquad \mathbf{Y}_0(\tau) = \mathbf{V}_0(\tau)\mathbf{P}. \tag{117}$$

They satisfy

$$\dot{\mathbf{X}}_0 = \mathbf{Y}_0\mathbf{\Lambda}, \qquad \dot{\mathbf{Y}}_0 = \mathbf{X}_0\mathbf{\Lambda}. \tag{118}$$

The components orthogonal to the positive right and left singular subspaces are constant:

$$\begin{aligned} \mathbf{U}_0(\tau)\left(\mathbf{I}_p - \mathbf{Q}\mathbf{Q}^\top\right) &= \mathbf{A}\left(\mathbf{I}_p - \mathbf{Q}\mathbf{Q}^\top\right), \\ \mathbf{V}_0(\tau)\left(\mathbf{I}_C - \mathbf{P}\mathbf{P}^\top\right) &= \mathbf{B}\left(\mathbf{I}_C - \mathbf{P}\mathbf{P}^\top\right). \end{aligned} \tag{119}$$

For the active components, define

$$\mathbf{\Gamma} = \frac{\mathbf{A}\mathbf{Q} + \mathbf{B}\mathbf{P}}{\sqrt{2}}, \qquad \mathbf{\Xi} = \frac{\mathbf{A}\mathbf{Q} - \mathbf{B}\mathbf{P}}{\sqrt{2}}. \tag{120}$$

Then

$$\begin{aligned} \mathbf{X}_0(\tau) &= \frac{1}{\sqrt{2}}\left(\mathbf{\Gamma}e^{\mathbf{\Lambda}\tau} + \mathbf{\Xi}e^{-\mathbf{\Lambda}\tau}\right), \\ \mathbf{Y}_0(\tau) &= \frac{1}{\sqrt{2}}\left(\mathbf{\Gamma}e^{\mathbf{\Lambda}\tau} - \mathbf{\Xi}e^{-\mathbf{\Lambda}\tau}\right). \end{aligned} \tag{121}$$

Equivalently,

$$\begin{aligned} \mathbf{U}_0(\tau) &= \frac{1}{\sqrt{2}}\left(\mathbf{\Gamma}e^{\mathbf{\Lambda}\tau} + \mathbf{\Xi}e^{-\mathbf{\Lambda}\tau}\right)\mathbf{Q}^\top + \mathbf{A}\left(\mathbf{I}_p - \mathbf{Q}\mathbf{Q}^\top\right), \\ \mathbf{V}_0(\tau) &= \frac{1}{\sqrt{2}}\left(\mathbf{\Gamma}e^{\mathbf{\Lambda}\tau} - \mathbf{\Xi}e^{-\mathbf{\Lambda}\tau}\right)\mathbf{P}^\top + \mathbf{B}\left(\mathbf{I}_C - \mathbf{P}\mathbf{P}^\top\right). \end{aligned} \tag{122}$$

Because multiplication by a matrix with orthonormal columns preserves the standard Gaussian law, $\mathbf{AQ}$ and $\mathbf{BP}$ are independent standard Gaussian matrices in $\mathbb{R}^{H\times r}$. The transformation in (120) is orthogonal, so $\mathbf{\Gamma}$ and $\mathbf{\Xi}$ are also independent standard Gaussian matrices. Equation (121) therefore shows explicitly that each singular mode j contains a growing component proportional to $e^{s_j\tau}$ and a decaying component proportional to $e^{-s_j\tau}$. Distinct singular values consequently generate different microscopic growth rates.

We next justify replacing the exact normalized dynamics by this linear system.

Theorem B.6 (Uniform microscopic linearization for two-factor linear networks)**.** *Fix p, C, H, an effective target $\mathbf{T} \in \mathbb{R}^{C\times p}$, and a finite time horizon $\tau_0 < \infty$. Let $\mathbf{A} \in \mathbb{R}^{H\times p}$ and $\mathbf{B} \in \mathbb{R}^{H\times C}$ be fixed, and initialize the exact gradient flow by*

$$\mathbf{U}^{(\sigma)}(0) = \sigma\mathbf{A}, \qquad \mathbf{V}^{(\sigma)}(0) = \sigma\mathbf{B}. \tag{123}$$

Let $\widehat{\mathbf{U}}^{(\sigma)}$ and $\widehat{\mathbf{V}}^{(\sigma)}$ be defined by (114)*, and let $\mathbf{U}_0, \mathbf{V}_0$ solve* (116)*. Then, for all sufficiently small $\sigma > 0$, the exact normalized solution is defined on $[0, \tau_0]$, and*

$$\sup_{0\le\tau\le\tau_0}\left(\left\|\widehat{\mathbf{U}}^{(\sigma)}(\tau) - \mathbf{U}_0(\tau)\right\|_{\mathrm{F}} + \left\|\widehat{\mathbf{V}}^{(\sigma)}(\tau) - \mathbf{V}_0(\tau)\right\|_{\mathrm{F}}\right) \longrightarrow 0 \tag{124}$$

as $\sigma \downarrow 0$.

Define

$$\mathbf{M}_0(\tau) = \mathbf{V}_0(\tau)^\top\mathbf{U}_0(\tau). \tag{125}$$

Then

$$\sup_{0\le\tau\le\tau_0}\left\|\frac{\mathbf{M}^{(\sigma)}(\tau)}{\sigma^2} - \mathbf{M}_0(\tau)\right\|_{\mathrm{F}} \longrightarrow 0, \tag{126}$$

where $\mathbf{M}^{(\sigma)} = \mathbf{V}^{(\sigma)\top}\mathbf{U}^{(\sigma)}$. Moreover,

$$\sup_{0\le\tau\le\tau_0}\left\|\mathbf{M}^{(\sigma)}(\tau)\right\|_{\mathrm{op}} \longrightarrow 0, \tag{127}$$

and

$$\sup_{0\le\tau\le\tau_0}\left|\mathcal{L}\left(\mathbf{U}^{(\sigma)}(\tau), \mathbf{V}^{(\sigma)}(\tau)\right) - \mathcal{L}(\mathbf{0}, \mathbf{0})\right| \longrightarrow 0. \tag{128}$$

When $\mathbf{A}, \mathbf{B}$ *are sampled according to* (112)*, these conclusions hold almost surely with respect to the initialization. Moreover, under this Gaussian initialization,*

$$\mathbb{E}\left[\sup_{0\le\tau\le\tau_0}\left\|\frac{\mathbf{M}^{(\sigma)}(\tau)}{\sigma^2}-\mathbf{M}_0(\tau)\right\|_{\mathrm{F}}\right]\longrightarrow 0. \tag{129}$$

In particular, for every $0\le\tau\le\tau_0$*,*

$$\lim_{\sigma\downarrow 0}\sigma^{-2}\,\mathbb{E}\Big[\mathbf{M}^{(\sigma)}(\tau)\Big]=\mathbb{E}[\mathbf{M}_0(\tau)]\,. \tag{130}$$

Proof. Write

$$\mathbf{Z}^{(\sigma)}=\left(\widehat{\mathbf{U}}^{(\sigma)},\widehat{\mathbf{V}}^{(\sigma)}\right),\qquad \mathbf{Z}_0=(\mathbf{U}_0,\mathbf{V}_0)\,, \tag{131}$$

and equip pairs of matrices with the norm

$$\|(\mathbf{X},\mathbf{Y})\|^2=\|\mathbf{X}\|_{\mathrm{F}}^2+\|\mathbf{Y}\|_{\mathrm{F}}^2. \tag{132}$$

Define the linear operator

$$\mathcal{A}(\mathbf{X},\mathbf{Y})=\left(\mathbf{Y}\mathbf{T},\mathbf{X}\mathbf{T}^\top\right) \tag{133}$$

and the cubic map

$$\mathcal{C}(\mathbf{X},\mathbf{Y})=\left(-\mathbf{Y}\mathbf{Y}^\top\mathbf{X},-\mathbf{X}\mathbf{X}^\top\mathbf{Y}\right). \tag{134}$$

Then (115) becomes

$$\dot{\mathbf{Z}}^{(\sigma)}=\mathcal{A}\mathbf{Z}^{(\sigma)}+\sigma^2\mathcal{C}\left(\mathbf{Z}^{(\sigma)}\right),\qquad \dot{\mathbf{Z}}_0=\mathcal{A}\mathbf{Z}_0, \tag{135}$$

with the same initial condition

$$\mathbf{Z}^{(\sigma)}(0)=\mathbf{Z}_0(0)=(\mathbf{A},\mathbf{B}). \tag{136}$$

The vector field in (135) is polynomial and hence locally Lipschitz. Thus, for every $\sigma>0$, there exists a unique maximal solution $\mathbf{Z}^{(\sigma)}$ on an interval $[0,\tau_{\max}^{(\sigma)})$. We first obtain a bound that prevents $\tau_{\max}^{(\sigma)}$ from falling below τ_0 when σ is sufficiently small.

Set

$$s_{\max}=\|\mathbf{T}\|_{\mathrm{op}}. \tag{137}$$

For every pair $(\mathbf{X},\mathbf{Y})$,

$$\begin{aligned}\|\mathcal{A}(\mathbf{X},\mathbf{Y})\|^2&=\|\mathbf{Y}\mathbf{T}\|_{\mathrm{F}}^2+\|\mathbf{X}\mathbf{T}^\top\|_{\mathrm{F}}^2\\&\le s_{\max}^2\left(\|\mathbf{X}\|_{\mathrm{F}}^2+\|\mathbf{Y}\|_{\mathrm{F}}^2\right).\end{aligned} \tag{138}$$

Hence

$$\|\mathcal{A}\|_{\mathrm{op}}\le s_{\max},\qquad \left\|e^{\tau\mathcal{A}}\right\|_{\mathrm{op}}\le e^{s_{\max}\tau},\qquad \tau\ge 0. \tag{139}$$

For the cubic term,

$$\begin{aligned}\left\|\mathbf{Y}\mathbf{Y}^\top\mathbf{X}\right\|_{\mathrm{F}}&\le\|\mathbf{Y}\|_{\mathrm{F}}^2\|\mathbf{X}\|_{\mathrm{F}},\\\left\|\mathbf{X}\mathbf{X}^\top\mathbf{Y}\right\|_{\mathrm{F}}&\le\|\mathbf{X}\|_{\mathrm{F}}^2\|\mathbf{Y}\|_{\mathrm{F}}.\end{aligned} \tag{140}$$

Therefore

$$\|\mathcal{C}(\mathbf{Z})\|\le\|\mathbf{Z}\|^3. \tag{141}$$

On the maximal existence interval, variation of constants gives

$$\mathbf{Z}^{(\sigma)}(\tau)=e^{\tau\mathcal{A}}\mathbf{Z}^{(\sigma)}(0)+\sigma^2\int_0^\tau e^{(\tau-u)\mathcal{A}}\mathcal{C}\left(\mathbf{Z}^{(\sigma)}(u)\right)\,du. \tag{142}$$

Since $\mathbf{Z}_0(\tau)=e^{\tau\mathcal{A}}\mathbf{Z}_0(0)$,

$$\mathbf{Z}^{(\sigma)}(\tau)-\mathbf{Z}_0(\tau)=\sigma^2\int_0^\tau e^{(\tau-u)\mathcal{A}}\mathcal{C}\left(\mathbf{Z}^{(\sigma)}(u)\right)\,du. \tag{143}$$

We next establish a uniform bound on the normalized exact solution. Set

$$K=1+e^{s_{\max}\tau_0}\|(\mathbf{A},\mathbf{B})\|. \tag{144}$$

Suppose that

$$\|\mathbf{Z}^{(\sigma)}(u)\| \le 2K \qquad \text{for } 0 \le u \le \tau, \tag{145}$$

where $\tau < \min\{\tau_0, \tau_{\max}^{(\sigma)}\}$. Then (139), (141), and (142) imply

$$\begin{aligned}\|\mathbf{Z}^{(\sigma)}(\tau)\| &\le e^{s_{\max}\tau}\|\mathbf{Z}^{(\sigma)}(0)\| + \sigma^2 \int_0^\tau e^{s_{\max}(\tau-u)}\|\mathbf{Z}^{(\sigma)}(u)\|^3\, du \\ &\le K + 8\sigma^2 K^3 \tau_0 e^{s_{\max}\tau_0}.\end{aligned} \tag{146}$$

For sufficiently small σ,

$$8\sigma^2 K^3 \tau_0 e^{s_{\max}\tau_0} < \frac{K}{2}, \tag{147}$$

and hence

$$\|\mathbf{Z}^{(\sigma)}(\tau)\| < \frac{3K}{2} < 2K. \tag{148}$$

A standard first-exit argument therefore shows that

$$\|\mathbf{Z}^{(\sigma)}(\tau)\| < 2K \qquad \text{for every } 0 \le \tau < \min\{\tau_0, \tau_{\max}^{(\sigma)}\}. \tag{149}$$

Indeed, if the norm reached $2K$ before either τ_0 or $\tau_{\max}^{(\sigma)}$, continuity would give a first such time, contradicting (148).

Suppose now, for contradiction, that $\tau_{\max}^{(\sigma)} \le \tau_0$. Then (149) keeps $\mathbf{Z}^{(\sigma)}(\tau)$ in the bounded set $\{\|\mathbf{Z}\| \le 2K\}$ throughout its maximal interval. Since the vector field in (135) is locally Lipschitz, the standard continuation theorem for finite-dimensional ODEs then extends the solution beyond $\tau_{\max}^{(\sigma)}$, contradicting maximality. Hence

$$\tau_{\max}^{(\sigma)} > \tau_0 \tag{150}$$

for all sufficiently small σ, and

$$\sup_{0\le\tau\le\tau_0} \|\mathbf{Z}^{(\sigma)}(\tau)\| \le 2K. \tag{151}$$

Substituting this bound into (143) yields

$$\begin{aligned}\sup_{0\le\tau\le\tau_0} \left\|\mathbf{Z}^{(\sigma)}(\tau) - \mathbf{Z}_0(\tau)\right\| &\le 8\sigma^2 K^3 \tau_0 e^{s_{\max}\tau_0} \\ &\longrightarrow 0.\end{aligned} \tag{152}$$

This proves (124).

Next,

$$\frac{\mathbf{M}^{(\sigma)}(\tau)}{\sigma^2} = \widehat{\mathbf{V}}^{(\sigma)}(\tau)^\top \widehat{\mathbf{U}}^{(\sigma)}(\tau). \tag{153}$$

For every $\tau \in [0, \tau_0]$,

$$\begin{aligned}\left\|\frac{\mathbf{M}^{(\sigma)}(\tau)}{\sigma^2} - \mathbf{M}_0(\tau)\right\|_{\mathrm{F}} &\le \left\|\widehat{\mathbf{V}}^{(\sigma)}(\tau) - \mathbf{V}_0(\tau)\right\|_{\mathrm{F}} \left\|\widehat{\mathbf{U}}^{(\sigma)}(\tau)\right\|_{\mathrm{F}} \\ &\quad + \|\mathbf{V}_0(\tau)\|_{\mathrm{F}} \left\|\widehat{\mathbf{U}}^{(\sigma)}(\tau) - \mathbf{U}_0(\tau)\right\|_{\mathrm{F}}.\end{aligned} \tag{154}$$

The normalized exact factors are uniformly bounded by (151), while the limiting factors are uniformly bounded on the compact interval $[0, \tau_0]$. Together with (124), this proves (126).

Moreover, (151) gives

$$\begin{aligned}\sup_{0\le\tau\le\tau_0} \|\mathbf{M}^{(\sigma)}(\tau)\|_{\mathrm{op}} &\le \sigma^2 \sup_{0\le\tau\le\tau_0} \|\widehat{\mathbf{V}}^{(\sigma)}(\tau)\|_{\mathrm{F}} \|\widehat{\mathbf{U}}^{(\sigma)}(\tau)\|_{\mathrm{F}} \\ &= O(\sigma^2),\end{aligned} \tag{155}$$

which proves (127).

Finally, by (89),

$$\mathcal{L}(\mathbf{U},\mathbf{V})-\mathcal{L}(\mathbf{0},\mathbf{0})=\gamma\left(\frac{1}{2}\|\mathbf{M}\|_{\mathrm{F}}^2-\langle\mathbf{M},\mathbf{T}\rangle_{\mathrm{F}}\right). \tag{156}$$

Hence

$$|\mathcal{L}(\mathbf{U},\mathbf{V})-\mathcal{L}(\mathbf{0},\mathbf{0})|\le\gamma\left(\frac{1}{2}\|\mathbf{M}\|_{\mathrm{F}}^2+\|\mathbf{M}\|_{\mathrm{F}}\|\mathbf{T}\|_{\mathrm{F}}\right). \tag{157}$$

From (151),

$$\begin{aligned}\sup_{0\le\tau\le\tau_0}\|\mathbf{M}^{(\sigma)}(\tau)\|_{\mathrm{F}}&\le\sigma^2\sup_{0\le\tau\le\tau_0}\|\widehat{\mathbf{V}}^{(\sigma)}(\tau)\|_{\mathrm{F}}\|\widehat{\mathbf{U}}^{(\sigma)}(\tau)\|_{\mathrm{F}}\\&=O(\sigma^2).\end{aligned} \tag{158}$$

Since $\mathbf{T}$ is fixed in the present theorem, the right-hand side of (157) therefore converges uniformly to zero on $[0,\tau_0]$. This proves (128).

The argument above is deterministic for every fixed finite pair $(\mathbf{A},\mathbf{B})$. If $\mathbf{A},\mathbf{B}$ are sampled from the finite-dimensional standard Gaussian distributions in (112), their entries are finite almost surely. Hence the preceding conclusions hold almost surely with respect to the initialization.

It remains to justify convergence in expectation. Define

$$S^{(\sigma)}(\tau)=\left\|\widehat{\mathbf{U}}^{(\sigma)}(\tau)\right\|_{\mathrm{F}}^2+\left\|\widehat{\mathbf{V}}^{(\sigma)}(\tau)\right\|_{\mathrm{F}}^2. \tag{159}$$

Using the exact normalized flow,

$$\begin{aligned}\frac{d}{d\tau}S^{(\sigma)}(\tau)&=4\left\langle\mathbf{T},\widehat{\mathbf{V}}^{(\sigma)}(\tau)^\top\widehat{\mathbf{U}}^{(\sigma)}(\tau)\right\rangle_{\mathrm{F}}\\&\quad-4\sigma^2\left\|\widehat{\mathbf{V}}^{(\sigma)}(\tau)^\top\widehat{\mathbf{U}}^{(\sigma)}(\tau)\right\|_{\mathrm{F}}^2\\&\le 2s_{\max}S^{(\sigma)}(\tau).\end{aligned} \tag{160}$$

Therefore

$$S^{(\sigma)}(\tau)\le e^{2s_{\max}\tau}\left(\|\mathbf{A}\|_{\mathrm{F}}^2+\|\mathbf{B}\|_{\mathrm{F}}^2\right),\qquad 0\le\tau\le\tau_0. \tag{161}$$

Consequently,

$$\sup_{0\le\tau\le\tau_0}\left\|\frac{\mathbf{M}^{(\sigma)}(\tau)}{\sigma^2}\right\|_{\mathrm{F}}\le\frac{1}{2}e^{2s_{\max}\tau_0}\left(\|\mathbf{A}\|_{\mathrm{F}}^2+\|\mathbf{B}\|_{\mathrm{F}}^2\right). \tag{162}$$

The same bound holds for $\sup_{0\le\tau\le\tau_0}\|\mathbf{M}_0(\tau)\|_{\mathrm{F}}$, by applying the same argument to the limiting linear flow. Hence

$$\sup_{0\le\tau\le\tau_0}\left\|\frac{\mathbf{M}^{(\sigma)}(\tau)}{\sigma^2}-\mathbf{M}_0(\tau)\right\|_{\mathrm{F}}\le e^{2s_{\max}\tau_0}\left(\|\mathbf{A}\|_{\mathrm{F}}^2+\|\mathbf{B}\|_{\mathrm{F}}^2\right). \tag{163}$$

Under Gaussian initialization, the right-hand side is integrable. Combining this domination with the almost-sure uniform convergence (126) and applying dominated convergence proves (129), and hence (130). □

For later use, we record the restriction of the limiting predictor to the positive singular subspaces. Define

$$\mathbf{N}_0(\tau)=\mathbf{P}^\top\mathbf{M}_0(\tau)\mathbf{Q}=\mathbf{Y}_0(\tau)^\top\mathbf{X}_0(\tau). \tag{164}$$

Writing

$$\mathbf{E}(\tau)=e^{\mathbf{\Lambda}\tau},\qquad\mathbf{F}(\tau)=e^{-\mathbf{\Lambda}\tau}, \tag{165}$$

equation (121) gives

$$\begin{aligned}\mathbf{N}_0(\tau)=\frac{1}{2}\Big(&\mathbf{E}(\tau)\mathbf{\Gamma}^\top\mathbf{\Gamma}\mathbf{E}(\tau)+\mathbf{E}(\tau)\mathbf{\Gamma}^\top\mathbf{\Xi}\mathbf{F}(\tau)\\&-\mathbf{F}(\tau)\mathbf{\Xi}^\top\mathbf{\Gamma}\mathbf{E}(\tau)-\mathbf{F}(\tau)\mathbf{\Xi}^\top\mathbf{\Xi}\mathbf{F}(\tau)\Big).\end{aligned} \tag{166}$$

This explicit random matrix process will be useful below when we analyze finite-width spectral interference.

The order of limits in Theorem B.6 is important. The theorem first fixes p, C, H and a finite time horizon τ_0, and only then sends $\sigma \downarrow 0$. Therefore, in a sequential small-initialization and large-system limit, τ_0 may depend on p, C, H, for example through $\tau_0 = \zeta \log H$, because it remains fixed during the inner limit $\sigma \downarrow 0$. The theorem does not apply to time scales that diverge together with σ^{-1}, such as $\tau \asymp \log(1/\sigma)$, where the cubic terms need no longer be negligible.

Early target-driven organization and target accuracy. The microscopic linearization describes the predictor while its norm and the change in loss are still vanishing, but it does not by itself determine the induced classifier. We now show that, at a sufficiently early time, the linearized gradient-flow dynamics produces a mean predictor aligned with the full effective target $\mathbf{T}$, while averaging over the H hidden units suppresses the initialization-dependent fluctuations around this mean. Together, these two effects make the predictor agree with the classifier of $\mathbf{T}$.

For each training example, let

$$\widehat{c}_i = \arg\max_{c \in [C]} (\mathbf{T}\mathbf{x}_i)_c \tag{167}$$

be the class predicted by $\mathbf{T}$. We assume that this maximizer is unique for every i, and define the minimum decision gap

$$\delta_{\mathbf{T}} = \min_{1 \le i \le N} \left[(\mathbf{T}\mathbf{x}_i)_{\widehat{c}_i} - \max_{c \neq \widehat{c}_i} (\mathbf{T}\mathbf{x}_i)_c \right] > 0. \tag{168}$$

Recall that

$$A_{\mathbf{T}} = \mathcal{A}(\mathbf{T}) = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\{\widehat{c}_i = c_i\}. \tag{169}$$

Thus the result below concerns agreement with the classifier induced by $\mathbf{T}$, whether or not that classifier has perfect accuracy.

We consider a sequence of problems indexed by $H \to \infty$, allowing $p, C, N, \mathbf{T}$, and the dataset to vary with H. Define

$$R = \max_{1 \le i \le N} \|\mathbf{x}_i\|_2, \qquad s = \|\mathbf{T}\|_{\mathrm{op}}, \qquad \ell = \log(2NCH), \qquad \kappa = \frac{Rs}{\delta_{\mathbf{T}}}. \tag{170}$$

The condition below is a convenient sufficient condition for uniform classification agreement and is not intended to be optimal.

Theorem B.7 (Early agreement with the target classifier)**.** *Suppose that*

$$\kappa^3 \frac{\ell}{H} \longrightarrow 0. \tag{171}$$

Then there exists a universal constant $K_0 > 0$ such that, for every fixed $K > K_0$, at the time

$$\tau_{\mathrm{org}} = K \frac{R}{\delta_{\mathbf{T}}} \sqrt{\frac{\ell}{H}}, \tag{172}$$

the limiting predictor $\mathbf{M}_0$ from Theorem B.6 satisfies

$$\mathbb{P}\left(\arg\max_{c \in [C]} (\mathbf{M}_0(\tau_{\mathrm{org}})\mathbf{x}_i)_c = \widehat{c}_i \quad \textit{for every } i \right) \longrightarrow 1. \tag{173}$$

Consequently, for the exact Gaussian-initialized gradient flow,

$$\lim_{H \to \infty} \liminf_{\sigma \downarrow 0} \mathbb{P}\left(\arg\max_{c \in [C]} (\mathbf{M}^{(\sigma)}(\tau_{\mathrm{org}})\mathbf{x}_i)_c = \widehat{c}_i \quad \textit{for every } i \right) = 1. \tag{174}$$

In particular,

$$\lim_{H \to \infty} \liminf_{\sigma \downarrow 0} \mathbb{P}\left(\mathcal{A}\big(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{org}})\big) = A_{\mathbf{T}} \right) = 1. \tag{175}$$

Moreover, for every fixed member of the large-system sequence, almost surely as $\sigma \downarrow 0$,

$$\left\| \mathbf{M}^{(\sigma)}(\tau_{\mathrm{org}}) \right\|_{\mathrm{op}} \longrightarrow 0, \qquad \mathcal{L}\left(\mathbf{U}^{(\sigma)}(\tau_{\mathrm{org}}), \mathbf{V}^{(\sigma)}(\tau_{\mathrm{org}}) \right) - \mathcal{L}(\mathbf{0}, \mathbf{0}) \longrightarrow 0. \tag{176}$$

Proof. We first determine the mean predictor of the limiting linearized dynamics.

Step 1: mean microscopic predictor. Write the rows of the Gaussian initialization as

$$\mathbf{A} = \begin{pmatrix} \mathbf{a}_1^\top \\ \vdots \\ \mathbf{a}_H^\top \end{pmatrix}, \qquad \mathbf{B} = \begin{pmatrix} \mathbf{b}_1^\top \\ \vdots \\ \mathbf{b}_H^\top \end{pmatrix}, \qquad \mathbf{a}_h \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_p), \qquad \mathbf{b}_h \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_C), \tag{177}$$

independently over h and between the two families. Likewise, write the rows of $\mathbf{U}_0(\tau)$ and $\mathbf{V}_0(\tau)$ as $\mathbf{u}_h(\tau)^\top$ and $\mathbf{v}_h(\tau)^\top$. The limiting equations decouple across hidden units:

$$\dot{\mathbf{u}}_h = \mathbf{T}^\top \mathbf{v}_h, \qquad \dot{\mathbf{v}}_h = \mathbf{T}\mathbf{u}_h, \qquad \mathbf{u}_h(0) = \mathbf{a}_h, \qquad \mathbf{v}_h(0) = \mathbf{b}_h. \tag{178}$$

Hence

$$\mathbf{M}_0(\tau) = \sum_{h=1}^{H} \mathbf{v}_h(\tau)\mathbf{u}_h(\tau)^\top. \tag{179}$$

Introduce the symmetric block operator

$$\mathcal{K} = \begin{pmatrix} \mathbf{0} & \mathbf{T}^\top \\ \mathbf{T} & \mathbf{0} \end{pmatrix}. \tag{180}$$

Then

$$\begin{pmatrix} \mathbf{u}_h(\tau) \\ \mathbf{v}_h(\tau) \end{pmatrix} = e^{\tau\mathcal{K}} \begin{pmatrix} \mathbf{a}_h \\ \mathbf{b}_h \end{pmatrix}. \tag{181}$$

The initial vector is standard Gaussian in $\mathbb{R}^{p+C}$, and $\mathcal{K}$ is symmetric. Its covariance at time τ is therefore $e^{2\tau\mathcal{K}}$. Using (93), the lower-left block of this covariance is

$$\mathbb{E}\left[\mathbf{v}_h(\tau)\mathbf{u}_h(\tau)^\top\right] = \mathbf{P}\sinh(2\mathbf{\Lambda}\tau)\mathbf{Q}^\top. \tag{182}$$

Define the normalized mean predictor

$$\overline{\mathbf{M}}(\tau) = \mathbf{P}\sinh(2\mathbf{\Lambda}\tau)\mathbf{Q}^\top = \sum_{j=1}^{r} \sinh(2s_j\tau)\mathbf{p}_j\mathbf{q}_j^\top. \tag{183}$$

Then

$$\mathbb{E}\mathbf{M}_0(\tau) = H\overline{\mathbf{M}}(\tau). \tag{184}$$

Step 2: early alignment of the mean predictor with $\mathbf{T}$. For $z \geq 0$,

$$0 \leq \sinh z - z \leq \frac{z^3}{6}e^z. \tag{185}$$

Since every singular value satisfies $s_j \leq s$, (183) gives

$$\left\|\overline{\mathbf{M}}(\tau) - 2\tau\mathbf{T}\right\|_{\mathrm{op}} \leq \frac{4}{3}s^3\tau^3 e^{2s\tau}. \tag{186}$$

For each training example i and each competitor $c \neq \widehat{c}_i$, define

$$\mathbf{d}_{ic} = \mathbf{e}_{\widehat{c}_i} - \mathbf{e}_c, \qquad m_{ic} = \mathbf{d}_{ic}^\top \mathbf{T}\mathbf{x}_i. \tag{187}$$

By the definition of $\delta_{\mathbf{T}}$,

$$m_{ic} \geq \delta_{\mathbf{T}}, \qquad \|\mathbf{d}_{ic}\|_2 = \sqrt{2}. \tag{188}$$

Therefore

$$\begin{aligned} \mathbf{d}_{ic}^\top \overline{\mathbf{M}}(\tau)\mathbf{x}_i &= 2\tau m_{ic} + \mathbf{d}_{ic}^\top \left(\overline{\mathbf{M}}(\tau) - 2\tau\mathbf{T}\right)\mathbf{x}_i \\ &\geq 2\tau\delta_{\mathbf{T}} - \frac{4\sqrt{2}}{3}Rs^3\tau^3 e^{2s\tau}. \end{aligned} \tag{189}$$

We now set $\tau = \tau_{\mathrm{org}}$. Since

$$\delta_{\mathbf{T}} \leq \mathbf{d}_{ic}^\top \mathbf{T}\mathbf{x}_i \leq \sqrt{2}\, Rs, \tag{190}$$

we have

$$\kappa = \frac{Rs}{\delta_{\mathbf{T}}} \geq \frac{1}{\sqrt{2}}. \tag{191}$$

Condition (171) therefore implies

$$\frac{\ell}{H} \longrightarrow 0. \tag{192}$$

Moreover,

$$s\tau_{\mathrm{org}} = K\kappa\sqrt{\frac{\ell}{H}} \longrightarrow 0, \tag{193}$$

because

$$\left(\kappa\sqrt{\frac{\ell}{H}}\right)^3 = \left(\kappa^3\frac{\ell}{H}\right)\sqrt{\frac{\ell}{H}} \longrightarrow 0. \tag{194}$$

The deterministic error in (189), relative to $\tau_{\mathrm{org}}\delta_{\mathbf{T}}$, satisfies

$$\frac{Rs^3\tau_{\mathrm{org}}^3 e^{2s\tau_{\mathrm{org}}}}{\tau_{\mathrm{org}}\delta_{\mathbf{T}}} = K^2\kappa^3\frac{\ell}{H}e^{2s\tau_{\mathrm{org}}} \longrightarrow 0. \tag{195}$$

Hence, for all sufficiently large H,

$$\mathbf{d}_{ic}^{\top}\overline{\mathbf{M}}(\tau_{\mathrm{org}})\mathbf{x}_i \geq \frac{3}{2}\tau_{\mathrm{org}}\delta_{\mathbf{T}} \qquad \text{for every } i,\ c \neq \widehat{c}_i. \tag{196}$$

Step 3: uniform concentration around the mean. Fix i and $c \neq \widehat{c}_i$, and abbreviate $\mathbf{d} = \mathbf{d}_{ic}$ and $\mathbf{x} = \mathbf{x}_i$. By (179),

$$\mathbf{d}^{\top}\left(\mathbf{M}_0(\tau) - H\overline{\mathbf{M}}(\tau)\right)\mathbf{x} = \sum_{h=1}^{H} Z_h, \tag{197}$$

where

$$Z_h = \left(\mathbf{d}^{\top}\mathbf{v}_h(\tau)\right)\left(\mathbf{u}_h(\tau)^{\top}\mathbf{x}\right) - \mathbb{E}\left[\left(\mathbf{d}^{\top}\mathbf{v}_h(\tau)\right)\left(\mathbf{u}_h(\tau)^{\top}\mathbf{x}\right)\right]. \tag{198}$$

The Z_h are independent and centered.

Since

$$\|\mathcal{K}\|_{\mathrm{op}} = s, \qquad \|e^{\tau\mathcal{K}}\|_{\mathrm{op}} \leq e^{s\tau}, \tag{199}$$

the centered Gaussian variables $\mathbf{d}^{\top}\mathbf{v}_h(\tau)$ and $\mathbf{u}_h(\tau)^{\top}\mathbf{x}$ have standard deviations at most $e^{s\tau}\|\mathbf{d}\|_2$ and $e^{s\tau}\|\mathbf{x}\|_2$, respectively. Their product is sub-exponential, and after centering there exists a universal constant $C_0 > 0$ such that

$$\|Z_h\|_{\psi_1} \leq C_0 e^{2s\tau}\|\mathbf{d}\|_2\|\mathbf{x}\|_2 \leq C_0\sqrt{2}\, e^{2s\tau}R. \tag{200}$$

Bernstein's inequality therefore gives a universal $C_1 > 0$ such that, for every $t \geq 1$,

$$\mathbb{P}\left(\left|\mathbf{d}^{\top}\left(\mathbf{M}_0(\tau) - H\overline{\mathbf{M}}(\tau)\right)\mathbf{x}\right| > C_1 e^{2s\tau}R\left(\sqrt{Ht} + t\right)\right) \leq 2e^{-t}. \tag{201}$$

Taking $t = \ell = \log(2NCH)$ and applying a union bound over the at most $N(C-1)$ training comparisons gives an event $\mathcal{E}_H$ with

$$\mathbb{P}(\mathcal{E}_H^c) \leq 2N(C-1)e^{-\ell} \leq \frac{1}{H}, \tag{202}$$

such that, on $\mathcal{E}_H$,

$$\max_{\substack{1\leq i\leq N\\ c\neq\widehat{c}_i}} \left|\mathbf{d}_{ic}^{\top}\left(\mathbf{M}_0(\tau_{\mathrm{org}}) - H\overline{\mathbf{M}}(\tau_{\mathrm{org}})\right)\mathbf{x}_i\right| \leq C_1 e^{2s\tau_{\mathrm{org}}}R\left(\sqrt{H\ell} + \ell\right). \tag{203}$$

The deterministic target-aligned margin has scale

$$H\tau_{\mathrm{org}}\delta_{\mathbf{T}} = KR\sqrt{H\ell}. \tag{204}$$

Hence the ratio of the fluctuation bound in (203) to this scale is at most

$$\frac{C_1}{K} e^{2s\tau_{\mathrm{org}}} \left(1 + \sqrt{\frac{\ell}{H}}\right) = \frac{C_1}{K} + o(1). \tag{205}$$

Choose $K_0 > 2C_1$. Then, for every fixed $K > K_0$ and all sufficiently large H, on $\mathcal{E}_H$,

$$\max_{\substack{1 \le i \le N \\ c \neq \widehat{c}_i}} \left| \mathbf{d}_{ic}^\top \left(\mathbf{M}_0(\tau_{\mathrm{org}}) - H\overline{\mathbf{M}}(\tau_{\mathrm{org}}) \right) \mathbf{x}_i \right| \le \frac{1}{2} H \tau_{\mathrm{org}} \delta_{\mathbf{T}}. \tag{206}$$

Combining this with (196) yields

$$\mathbf{d}_{ic}^\top \mathbf{M}_0(\tau_{\mathrm{org}}) \mathbf{x}_i \ge H \tau_{\mathrm{org}} \delta_{\mathbf{T}} > 0 \tag{207}$$

simultaneously for every training example and every competing class. Therefore $\mathbf{M}_0(\tau_{\mathrm{org}})$ agrees with the classifier of $\mathbf{T}$ on the full training set with probability tending to one.

Step 4: transfer to the exact gradient flow. For each fixed member of the outer large-system sequence, τ_{org} is a fixed finite time during the inner limit $\sigma \downarrow 0$. Hence Theorem B.6 gives

$$\frac{\mathbf{M}^{(\sigma)}(\tau_{\mathrm{org}})}{\sigma^2} \longrightarrow \mathbf{M}_0(\tau_{\mathrm{org}}) \qquad \text{in Frobenius norm} \tag{208}$$

almost surely as $\sigma \downarrow 0$.

On the event established above, every relevant pairwise decision gap of $\mathbf{M}_0(\tau_{\mathrm{org}})$ is strictly positive. Since multiplying all scores by the positive scalar σ^2 does not change their ordering, convergence in (208) implies that, for all sufficiently small σ, the exact predictor makes the same prediction $\widehat{c}_i$ on every training example. Taking the inner $\sigma \downarrow 0$ limit and then $H \to \infty$ proves (174) and (175).

Finally, the norm and loss conclusions in (176) follow directly from Theorem B.6, applied at the fixed time τ_{org}. □

The proof identifies the two scales responsible for the early organization. At τ_{org}, the coherent target-aligned pairwise margins have size

$$H \tau_{\mathrm{org}} \delta_{\mathbf{T}} \asymp R\sqrt{H \log(NCH)}, \tag{209}$$

whereas the initialization fluctuations have only the same $\sqrt{H \log(NCH)}$ scaling with a smaller constant once K is chosen sufficiently large. At the same time, $s\tau_{\mathrm{org}} \to 0$, so the mean predictor is still asymptotically proportional to the full target $\mathbf{T}$, before substantial spectral reweighting has occurred. This is the target-driven organization stage.

For example, if

$$H \asymp p \asymp C \asymp N, \qquad R = O(1), \qquad \|\mathbf{T}\|_{\mathrm{op}} = O(1), \qquad \delta_{\mathbf{T}}^{-1} = O(1), \tag{210}$$

then condition (171) holds automatically and

$$\tau_{\mathrm{org}} \asymp \sqrt{\frac{\log H}{H}}. \tag{211}$$

Thus width proportional to the problem dimensions is sufficient for the network to reproduce the classifier of $\mathbf{T}$ during the initial small-output, asymptotically constant-loss regime.

Spectral concentration and the predictive bottleneck. The preceding theorem shows that the early linearized dynamics can reproduce the classifier of the full effective target $\mathbf{T}$ while both the end-to-end predictor norm and the change in loss vanish as $\sigma \downarrow 0$. We now show that this target-driven organization need not persist. Since singular modes with larger singular values grow faster, the microscopic mean predictor progressively concentrates onto the complete top singular block $\mathbf{T}_{k_\star}$. Under suitable width and margin conditions, the finite-width predictor inherits the classifier of this block, producing the predictive bottleneck.

Recall from (183) that

$$\overline{\mathbf{M}}(\tau) = \sum_{j=1}^{r} \sinh(2s_j\tau)\, \mathbf{p}_j \mathbf{q}_j^\top. \tag{212}$$

We first record the deterministic concentration of this mean predictor.

Lemma B.8 (Spectral concentration onto the top singular block)**.** *Let $k_\star$ be defined by* (94)*, and set*

$$\rho_\star(\tau) := \frac{\sinh(2s_{k_\star+1}\tau)}{\sinh(2s_1\tau)}, \qquad s_{r+1} := 0. \tag{213}$$

Then, for every $\tau > 0$,

$$\left\| \frac{\overline{\mathbf{M}}(\tau)}{\sinh(2s_1\tau)} - \frac{\mathbf{T}_{k_\star}}{s_1} \right\|_{\mathrm{op}} = \rho_\star(\tau). \tag{214}$$

Equivalently,

$$\left\| \frac{s_1\overline{\mathbf{M}}(\tau)}{\sinh(2s_1\tau)} - \mathbf{T}_{k_\star} \right\|_{\mathrm{op}} = s_1\rho_\star(\tau). \tag{215}$$

In particular,

$$\rho_\star(\tau) \longrightarrow 0 \qquad \text{as } \tau \to \infty, \tag{216}$$

and hence

$$\frac{s_1\overline{\mathbf{M}}(\tau)}{\sinh(2s_1\tau)} \longrightarrow \mathbf{T}_{k_\star} \qquad \text{in operator norm.} \tag{217}$$

Proof. Since $s_1 = \cdots = s_{k_\star}$, dividing (212) by $\sinh(2s_1\tau)$ gives

$$\frac{\overline{\mathbf{M}}(\tau)}{\sinh(2s_1\tau)} - \frac{\mathbf{T}_{k_\star}}{s_1} = \sum_{j=k_\star+1}^{r} \frac{\sinh(2s_j\tau)}{\sinh(2s_1\tau)} \mathbf{p}_j \mathbf{q}_j^\top. \tag{218}$$

The left and right singular vectors in the sum form orthonormal families, so the operator norm of the remainder equals its largest coefficient. Since $s_{k_\star+1} \geq s_j$ for every $j > k_\star$, this coefficient is $\rho_\star(\tau)$, proving (214). Multiplying by s_1 gives (215).

By definition of $k_\star$, $s_{k_\star+1} < s_1$ whenever $k_\star < r$, so $\rho_\star(\tau) \to 0$. If $k_\star = r$, then $s_{k_\star+1} = 0$ and $\rho_\star(\tau) = 0$ identically. This proves the remaining statements. □

The lemma concerns the mean predictor. We next show that the finite-width predictor has the same classifier as $\mathbf{T}_{k_\star}$ at a suitable later time. For each training example, define

$$\widehat{c}_i^{(\star)} = \arg\max_{c\in[C]} (\mathbf{T}_{k_\star}\mathbf{x}_i)_c, \tag{219}$$

and assume that the maximizer is unique for every i. Define the minimum decision gap

$$\delta_{k_\star} = \min_{1\leq i\leq N} \left[(\mathbf{T}_{k_\star}\mathbf{x}_i)_{\widehat{c}_i^{(\star)}} - \max_{c\neq\widehat{c}_i^{(\star)}} (\mathbf{T}_{k_\star}\mathbf{x}_i)_c \right] > 0. \tag{220}$$

Theorem B.9 (Microscopic top-block predictive bottleneck)**.** *Let τ_{org} be the early organization time from Theorem B.7. Suppose that there is a sequence of bottleneck times τ_{bot} such that, for all sufficiently large H,*

$$\tau_{\mathrm{bot}} > \tau_{\mathrm{org}}, \qquad s_1\tau_{\mathrm{bot}} \geq 1, \tag{221}$$

and

$$\frac{Rs_1}{\delta_{k_\star}}\rho_\star(\tau_{\mathrm{bot}}) \longrightarrow 0, \qquad \frac{Rs_1}{\delta_{k_\star}}\sqrt{\frac{\ell}{H}} \longrightarrow 0, \tag{222}$$

where

$$R = \max_i \|\mathbf{x}_i\|_2, \qquad \ell = \log(2NCH). \tag{223}$$

Then the limiting predictor satisfies

$$\mathbb{P}\left(\arg\max_{c\in[C]} (\mathbf{M}_0(\tau_{\mathrm{bot}})\mathbf{x}_i)_c = \widehat{c}_i^{(\star)} \quad \text{for every } i \right) \longrightarrow 1. \tag{224}$$

Consequently, for the exact Gaussian-initialized gradient flow,

$$\lim_{H\to\infty} \liminf_{\sigma\downarrow 0} \mathbb{P}\left(\arg\max_{c\in[C]} (\mathbf{M}^{(\sigma)}(\tau_{\mathrm{bot}})\mathbf{x}_i)_c = \widehat{c}_i^{(\star)} \quad \text{for every } i \right) = 1. \tag{225}$$

In particular,

$$\lim_{H\to\infty} \liminf_{\sigma\downarrow 0} \mathbb{P}\left(\mathcal{A}\big(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{bot}})\big) = A_{k_\star}\right) = 1. \tag{226}$$

Moreover, for each fixed member of the large-system sequence, almost surely as $\sigma \downarrow 0$,

$$\left\|\mathbf{M}^{(\sigma)}(\tau_{\mathrm{bot}})\right\|_{\mathrm{op}} \longrightarrow 0, \qquad \mathcal{L}\left(\mathbf{U}^{(\sigma)}(\tau_{\mathrm{bot}}), \mathbf{V}^{(\sigma)}(\tau_{\mathrm{bot}})\right) - \mathcal{L}(\mathbf{0},\mathbf{0}) \longrightarrow 0. \tag{227}$$

Proof. We first show that the mean predictor has the same classifier as $\mathbf{T}_{k_\star}$ at τ_{bot}.

Step 1: classification by the spectrally concentrated mean. For each example i and each competitor $c \neq \widehat{c}_i^{(\star)}$, define

$$\mathbf{d}_{ic}^{(\star)} = \mathbf{e}_{\widehat{c}_i^{(\star)}} - \mathbf{e}_c. \tag{228}$$

By definition of $\delta_{k_\star}$,

$$(\mathbf{d}_{ic}^{(\star)})^\top \mathbf{T}_{k_\star}\mathbf{x}_i \geq \delta_{k_\star}, \qquad \|\mathbf{d}_{ic}^{(\star)}\|_2 = \sqrt{2}. \tag{229}$$

Set

$$a_\star(\tau) = \sinh(2s_1\tau). \tag{230}$$

By Lemma B.8, we may write

$$\overline{\mathbf{M}}(\tau) = a_\star(\tau)\left(\frac{\mathbf{T}_{k_\star}}{s_1} + \mathbf{E}_\star(\tau)\right), \qquad \|\mathbf{E}_\star(\tau)\|_{\mathrm{op}} = \rho_\star(\tau). \tag{231}$$

Therefore

$$(\mathbf{d}_{ic}^{(\star)})^\top \overline{\mathbf{M}}(\tau)\mathbf{x}_i \geq a_\star(\tau)\left(\frac{\delta_{k_\star}}{s_1} - \sqrt{2}\,R\rho_\star(\tau)\right). \tag{232}$$

The first condition in (222) implies that, for all sufficiently large H,

$$\sqrt{2}\,R\rho_\star(\tau_{\mathrm{bot}}) \leq \frac{\delta_{k_\star}}{4s_1}. \tag{233}$$

Hence

$$(\mathbf{d}_{ic}^{(\star)})^\top \overline{\mathbf{M}}(\tau_{\mathrm{bot}})\mathbf{x}_i \geq \frac{3}{4} a_\star(\tau_{\mathrm{bot}})\frac{\delta_{k_\star}}{s_1} \tag{234}$$

uniformly over all training comparisons.

Step 2: uniform concentration around the mean. As in the proof of Theorem B.7, for every fixed comparison $\mathbf{d} = \mathbf{d}_{ic}^{(\star)}$ and $\mathbf{x} = \mathbf{x}_i$,

$$\mathbf{d}^\top\left(\mathbf{M}_0(\tau) - H\overline{\mathbf{M}}(\tau)\right)\mathbf{x} = \sum_{h=1}^{H} Z_h, \tag{235}$$

where the Z_h are independent and centered.

Since $\|\mathcal{K}\|_{\mathrm{op}} = s_1$, the same sub-exponential estimate gives

$$\|Z_h\|_{\psi_1} \leq C_0\sqrt{2}\,e^{2s_1\tau}R \tag{236}$$

for a universal constant $C_0 > 0$. Bernstein's inequality and a union bound over all $N(C-1)$ comparisons therefore give an event $\mathcal{F}_H$ satisfying

$$\mathbb{P}(\mathcal{F}_H^c) \leq \frac{1}{H}, \tag{237}$$

such that, on $\mathcal{F}_H$,

$$\max_{\substack{1\leq i\leq N\\ c\neq\widehat{c}_i^{(\star)}}} \left|(\mathbf{d}_{ic}^{(\star)})^\top\left(\mathbf{M}_0(\tau_{\mathrm{bot}}) - H\overline{\mathbf{M}}(\tau_{\mathrm{bot}})\right)\mathbf{x}_i\right| \leq C_1 e^{2s_1\tau_{\mathrm{bot}}}R\left(\sqrt{H\ell} + \ell\right) \tag{238}$$

for a universal constant $C_1 > 0$.

We compare this with the deterministic top-block margin

$$Ha_\star(\tau_{\mathrm{bot}})\frac{\delta_{k_\star}}{s_1}. \tag{239}$$

Since $s_1\tau_{\mathrm{bot}} \geq 1$,

$$\frac{e^{2s_1\tau_{\mathrm{bot}}}}{\sinh(2s_1\tau_{\mathrm{bot}})} \leq \frac{2}{1-e^{-4}} =: C_2. \tag{240}$$

Thus the ratio of the fluctuation bound in (238) to the scale in (239) is bounded by

$$C_1C_2\frac{Rs_1}{\delta_{k_\star}}\left(\sqrt{\frac{\ell}{H}}+\frac{\ell}{H}\right). \tag{241}$$

Because $\|\mathbf{T}_{k_\star}\|_{\mathrm{op}} = s_1$, the decision gap satisfies

$$\delta_{k_\star} \leq \sqrt{2}\,Rs_1, \tag{242}$$

and therefore

$$\frac{Rs_1}{\delta_{k_\star}} \geq \frac{1}{\sqrt{2}}. \tag{243}$$

The second condition in (222) consequently implies $\ell/H \to 0$, and hence (241) tends to zero. For all sufficiently large H, on $\mathcal{F}_H$,

$$\max_{\substack{1\leq i\leq N\\ c\neq \widehat{c}_i^{(\star)}}} \left|(\mathbf{d}_{ic}^{(\star)})^\top\left(\mathbf{M}_0(\tau_{\mathrm{bot}}) - H\overline{\mathbf{M}}(\tau_{\mathrm{bot}})\right)\mathbf{x}_i\right| \leq \frac{1}{4}Ha_\star(\tau_{\mathrm{bot}})\frac{\delta_{k_\star}}{s_1}. \tag{244}$$

Combining (234) and (244) yields

$$(\mathbf{d}_{ic}^{(\star)})^\top\mathbf{M}_0(\tau_{\mathrm{bot}})\mathbf{x}_i \geq \frac{1}{2}Ha_\star(\tau_{\mathrm{bot}})\frac{\delta_{k_\star}}{s_1} > 0 \tag{245}$$

simultaneously for every training example and every competing class. Since $\mathbb{P}(\mathcal{F}_H) \to 1$, this proves (224).

Step 3: transfer to the exact gradient flow. Although τ_{bot} may grow along the outer large-system sequence, it is a fixed finite time for every fixed member of that sequence during the inner limit $\sigma \downarrow 0$. Therefore Theorem B.6, applied on $[0, \tau_{\mathrm{bot}}]$, gives

$$\frac{\mathbf{M}^{(\sigma)}(\tau_{\mathrm{bot}})}{\sigma^2} \longrightarrow \mathbf{M}_0(\tau_{\mathrm{bot}}) \qquad \text{in Frobenius norm} \tag{246}$$

almost surely as $\sigma \downarrow 0$.

On the event established in (224), all relevant pairwise decision gaps are strictly positive. Hence, for all sufficiently small σ, the exact predictor makes the same predictions $\widehat{c}_i^{(\star)}$ on every training example. Taking first $\sigma \downarrow 0$ and then $H \to \infty$ proves (225) and (226).

Finally, the norm and loss limits in (227) follow directly from Theorem B.6, applied at τ_{bot}. □

The required temporal separation is particularly transparent when the problem geometry and the spectral gap below the top block remain controlled. Suppose, for example, that

$$H \asymp p \asymp C \asymp N, \qquad R = O(1), \qquad \|\mathbf{T}\|_{\mathrm{op}} = O(1), \qquad \delta_{\mathbf{T}}^{-1} = O(1), \qquad \delta_{k_\star}^{-1} = O(1), \tag{247}$$

and that

$$s_1 - s_{k_\star+1} \geq \Delta > 0 \tag{248}$$

uniformly along the large-system sequence. Then Theorem B.7 gives

$$\tau_{\mathrm{org}} \asymp \sqrt{\frac{\log H}{H}}, \tag{249}$$

whereas for any fixed $\zeta > 0$ we may choose

$$\tau_{\mathrm{bot}} = \zeta \log H. \tag{250}$$

Indeed,

$$\rho_\star(\tau_{\mathrm{bot}}) = O\left(e^{-2(s_1 - s_{k_\star+1})\tau_{\mathrm{bot}}}\right) = O\left(H^{-2\zeta(s_1 - s_{k_\star+1})}\right) \longrightarrow 0, \tag{251}$$

while

$$\sqrt{\frac{\ell}{H}} = O\left(\sqrt{\frac{\log H}{H}}\right) \longrightarrow 0. \tag{252}$$

Consequently,

$$\tau_{\mathrm{org}} < \tau_{\mathrm{bot}} \tag{253}$$

for all sufficiently large H. Thus the classifier of the full target $\mathbf{T}$ is established first, and only later does spectral concentration produce the classifier of the complete top singular block $\mathbf{T}_{k_\star}$.

Together with the definitions of $A_{\mathbf{T}}$ and $A_{k_\star}$, the preceding results imply that, with probability tending to one in the stated sequential limit,

$$\mathcal{A}\Big(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{org}})\Big) = A_{\mathbf{T}}, \qquad \mathcal{A}\Big(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{bot}})\Big) = A_{k_\star}. \tag{254}$$

Hence an accuracy decrease between these two stages occurs precisely when

$$A_{\mathbf{T}} > A_{k_\star}. \tag{255}$$

As discussed at the beginning of this subsection, this ordering is a classification property of the dataset and target rather than a consequence of the singular-value decomposition alone.

Recovery to the target classifier. The preceding results concern the microscopic small-initialization regime: the classifier first agrees with the full effective target $\mathbf{T}$, and at a later microscopic time spectral concentration makes it agree with the top singular block $\mathbf{T}_{k_\star}$. We now leave the microscopic approximation and consider the full gradient flow at arbitrarily late times. In the realizable two-factor setting, the end-to-end predictor converges to $\mathbf{T}$. The positive decision gap of $\mathbf{T}$ then implies permanent recovery of its classifier.

Proposition B.10 (Eventual recovery of the target classifier)**.** *Under the standing isotropy assumption $\mathbf{G} = \gamma \mathbf{I}_p$, suppose that*

$$H \geq \mathrm{rank}(\mathbf{T}), \tag{256}$$

and that the effective target $\mathbf{T}$ has the strictly positive decision gap $\delta_{\mathbf{T}}$ defined in (168)*. Then, for every fixed $\sigma > 0$, the Gaussian-initialized gradient flow satisfies*

$$\mathbf{M}^{(\sigma)}(\tau) = \mathbf{V}^{(\sigma)}(\tau)^\top \mathbf{U}^{(\sigma)}(\tau) \longrightarrow \mathbf{T} \qquad \text{as } \tau \to \infty \tag{257}$$

almost surely.

Consequently, almost surely there exists a finite random time $\tau_{\mathrm{rec}}^{(\sigma)}$ such that

$$\arg\max_{c \in [C]} \left(\mathbf{M}^{(\sigma)}(\tau)\mathbf{x}_i\right)_c = \widehat{c}_i \qquad \text{for every } i \text{ and every } \tau \geq \tau_{\mathrm{rec}}^{(\sigma)}. \tag{258}$$

In particular,

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau)\right) = A_{\mathbf{T}} \qquad \text{for every } \tau \geq \tau_{\mathrm{rec}}^{(\sigma)}. \tag{259}$$

Proof. Let

$$\mathbf{X} = (\mathbf{x}_1 \quad \cdots \quad \mathbf{x}_N) \in \mathbb{R}^{p \times N}, \qquad \mathbf{Y} = (\mathbf{y}_1 \quad \cdots \quad \mathbf{y}_N) \in \mathbb{R}^{C \times N}. \tag{260}$$

By the definition of the empirical second-moment matrix and the standing isotropy assumption,

$$\frac{1}{N}\mathbf{X}\mathbf{X}^\top = \gamma \mathbf{I}_p, \qquad \mathbf{X}\mathbf{X}^\top = N\gamma \mathbf{I}_p. \tag{261}$$

Hence $\mathbf{X}\mathbf{X}^\top$ is full rank.

We may therefore apply the two-layer global convergence result of (Bah et al., 2022, Theorem 39(b)). In our notation, the two factors are $\mathbf{U}$ and $\mathbf{V}^\top$. It follows that, for almost every initialization, the factorized gradient flow converges to a global minimizer of $\mathcal{L}(\mathbf{U}, \mathbf{V})$.

It remains to identify the corresponding end-to-end predictor. By (89),

$$\mathcal{L}(\mathbf{U}, \mathbf{V}) = \frac{\gamma}{2} \left\| \mathbf{V}^\top \mathbf{U} - \mathbf{T} \right\|_{\mathrm{F}}^2 + \mathcal{L}_{\mathrm{const}}. \tag{262}$$

Condition (256) guarantees that $\mathbf{T}$ is representable by the factorization. Indeed, using the compact singular-value decomposition

$$\mathbf{T} = \mathbf{P} \mathbf{\Lambda} \mathbf{Q}^\top, \qquad r = \mathrm{rank}(\mathbf{T}) \le H, \tag{263}$$

choose any $\mathbf{E} \in \mathbb{R}^{H \times r}$ satisfying

$$\mathbf{E}^\top \mathbf{E} = \mathbf{I}_r, \tag{264}$$

and define

$$\mathbf{U}_\star = \mathbf{E} \mathbf{\Lambda}^{1/2} \mathbf{Q}^\top, \qquad \mathbf{V}_\star = \mathbf{E} \mathbf{\Lambda}^{1/2} \mathbf{P}^\top. \tag{265}$$

Then

$$\mathbf{V}_\star^\top \mathbf{U}_\star = \mathbf{P} \mathbf{\Lambda} \mathbf{Q}^\top = \mathbf{T}. \tag{266}$$

Thus the global minimum of (262) is $\mathcal{L}_{\mathrm{const}}$, and every global minimizer satisfies

$$\mathbf{V}^\top \mathbf{U} = \mathbf{T}. \tag{267}$$

The global convergence theorem therefore implies (257) for almost every initialization.

For every fixed $\sigma > 0$, the Gaussian initialization

$$\mathbf{U}^{(\sigma)}(0) = \sigma \mathbf{A}, \qquad \mathbf{V}^{(\sigma)}(0) = \sigma \mathbf{B} \tag{268}$$

has a density with respect to Lebesgue measure on the full parameter space. Hence the exceptional set of initializations of Lebesgue measure zero has probability zero, and (257) holds almost surely.

It remains to translate convergence of the end-to-end predictor into classification recovery. For every $c \neq \widehat{c}_i$, recall

$$\mathbf{d}_{ic} = \mathbf{e}_{\widehat{c}_i} - \mathbf{e}_c, \qquad \mathbf{d}_{ic}^\top \mathbf{T} \mathbf{x}_i \ge \delta_{\mathbf{T}}, \qquad \|\mathbf{d}_{ic}\|_2 = \sqrt{2}. \tag{269}$$

For any predictor $\mathbf{M}$,

$$\begin{aligned} \mathbf{d}_{ic}^\top \mathbf{M} \mathbf{x}_i &= \mathbf{d}_{ic}^\top \mathbf{T} \mathbf{x}_i + \mathbf{d}_{ic}^\top (\mathbf{M} - \mathbf{T}) \mathbf{x}_i \\ &\ge \delta_{\mathbf{T}} - \sqrt{2}\, R \|\mathbf{M} - \mathbf{T}\|_{\mathrm{op}}. \end{aligned} \tag{270}$$

By (257), almost surely there exists a finite time $\tau_{\mathrm{rec}}^{(\sigma)}$ such that

$$\left\| \mathbf{M}^{(\sigma)}(\tau) - \mathbf{T} \right\|_{\mathrm{op}} < \frac{\delta_{\mathbf{T}}}{2\sqrt{2}\, R} \qquad \text{for every } \tau \ge \tau_{\mathrm{rec}}^{(\sigma)}. \tag{271}$$

Substituting this into (270) gives

$$\mathbf{d}_{ic}^\top \mathbf{M}^{(\sigma)}(\tau) \mathbf{x}_i > \frac{\delta_{\mathbf{T}}}{2} > 0 \tag{272}$$

simultaneously for every training example and every competing class. Therefore the exact predictor agrees with the classifier of $\mathbf{T}$ on the full training set for every $\tau \ge \tau_{\mathrm{rec}}^{(\sigma)}$, proving (258) and (259). □

Unlike the target-driven organization and the predictive bottleneck, recovery is not a statement about the microscopic limiting dynamics. For every fixed $\sigma > 0$, Proposition B.10 concerns the full gradient flow as $\tau \to \infty$. In particular, the recovery time $\tau_{\mathrm{rec}}^{(\sigma)}$ may depend on σ and may diverge as $\sigma \downarrow 0$.

Theorem B.11 [4.2] **(Three-stage classifier dynamics for $L = 2$).**

Consider a sequence of two-factor problems with $H \to \infty$, allowing N, p, C, the dataset, and $\mathbf{T}$ to vary with H. Assume

$$R := \max_i \|\mathbf{x}_i\|_2 = O(1), \qquad \|\mathbf{T}\|_{\mathrm{op}} = O(1), \qquad \log(2NCH) = o(H), \qquad H \ge \mathrm{rank}(\mathbf{T}). \tag{273}$$

Suppose further that there exists a constant $\Delta > 0$ such that

$$s_1 - s_{k_\star+1} \geq \Delta \tag{274}$$

uniformly along the sequence, where $s_{r+1} := 0$, and that the decision gaps of both $\mathbf{T}$ and $\mathbf{T}_{k_\star}$ are uniformly bounded away from zero:

$$\inf_H \delta_{\mathbf{T}} > 0, \qquad \inf_H \delta_{k_\star} > 0. \tag{275}$$

Then, for all sufficiently large H, there exist deterministic times

$$\tau_{\mathrm{org}} < \tau_{\mathrm{bot}} \tag{276}$$

such that

$$\lim_{H\to\infty} \liminf_{\sigma\downarrow 0} \mathbb{P}\left(\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{org}})\right) = A_{\mathbf{T}}, \quad \mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{bot}})\right) = A_{k_\star}\right) = 1. \tag{277}$$

At both times, for every fixed member of the large-system sequence, almost surely as $\sigma \downarrow 0$,

$$\left\|\mathbf{M}^{(\sigma)}(\tau)\right\|_{\mathrm{op}} \longrightarrow 0, \qquad \mathcal{L}\left(\mathbf{U}^{(\sigma)}(\tau), \mathbf{V}^{(\sigma)}(\tau)\right) - \mathcal{L}(\mathbf{0},\mathbf{0}) \longrightarrow 0, \qquad \tau \in \{\tau_{\mathrm{org}}, \tau_{\mathrm{bot}}\}. \tag{278}$$

Moreover, for every fixed member of the sequence and every fixed $\sigma > 0$, almost surely there exists a finite random time

$$\tau_{\mathrm{rec}}^{(\sigma)} > \tau_{\mathrm{bot}} \tag{279}$$

such that

$$\arg\max_{c\in[C]} \left(\mathbf{M}^{(\sigma)}(\tau)\mathbf{x}_i\right)_c = \widehat{c}_i \qquad \textit{for every } i \textit{ and every } \tau \geq \tau_{\mathrm{rec}}^{(\sigma)}. \tag{280}$$

In particular,

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau)\right) = A_{\mathbf{T}} \qquad \textit{for every } \tau \geq \tau_{\mathrm{rec}}^{(\sigma)}. \tag{281}$$

Consequently, if

$$A_{\mathbf{T}} > A_{k_\star}, \tag{282}$$

then

$$\lim_{H\to\infty} \liminf_{\sigma\downarrow 0} \mathbb{P}\left(\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{org}})\right) > \mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{bot}})\right)\right) = 1, \tag{283}$$

and for every fixed $\sigma > 0$ the classifier subsequently recovers permanently to accuracy $A_{\mathbf{T}}$.

Proof. Set

$$\ell = \log(2NCH), \qquad \kappa = \frac{R\|\mathbf{T}\|_{\mathrm{op}}}{\delta_{\mathbf{T}}}. \tag{284}$$

By the assumed bounds on R, $\|\mathbf{T}\|_{\mathrm{op}}$, and $\delta_{\mathbf{T}}$,

$$\kappa = O(1). \tag{285}$$

Since $\ell = o(H)$,

$$\kappa^3 \frac{\ell}{H} \longrightarrow 0. \tag{286}$$

Therefore Theorem B.7 applies. For any fixed $K > K_0$, choose

$$\tau_{\mathrm{org}} = K\frac{R}{\delta_{\mathbf{T}}}\sqrt{\frac{\ell}{H}}. \tag{287}$$

The standing assumptions imply

$$\tau_{\mathrm{org}} \longrightarrow 0, \tag{288}$$

and, in the sequential limit $\sigma \downarrow 0$ followed by $H \to \infty$, the classifier at τ_{org} agrees with that of $\mathbf{T}$ with probability tending to one.

For the bottleneck, fix any constant $\zeta > 0$ and set

$$\tau_{\mathrm{bot}} = \zeta \log H. \tag{289}$$

Since

$$s_1 - s_{k_\star+1} \geq \Delta > 0, \tag{290}$$

we have $s_1 \geq \Delta$, and therefore

$$s_1 \tau_{\mathrm{bot}} \geq 1 \tag{291}$$

for all sufficiently large H.

Furthermore, using (213),

$$\begin{aligned} \rho_\star(\tau_{\mathrm{bot}}) &= \frac{\sinh(2s_{k_\star+1}\tau_{\mathrm{bot}})}{\sinh(2s_1\tau_{\mathrm{bot}})} \\ &\leq \frac{1}{1-e^{-4}}\, e^{-2(s_1 - s_{k_\star+1})\tau_{\mathrm{bot}}} \\ &\leq \frac{1}{1-e^{-4}}\, H^{-2\zeta\Delta} \longrightarrow 0. \end{aligned} \tag{292}$$

Because

$$Rs_1 = O(1), \qquad \delta_{k_\star}^{-1} = O(1), \tag{293}$$

we obtain

$$\frac{Rs_1}{\delta_{k_\star}} \rho_\star(\tau_{\mathrm{bot}}) \longrightarrow 0. \tag{294}$$

Likewise, since $\ell = o(H)$,

$$\frac{Rs_1}{\delta_{k_\star}} \sqrt{\frac{\ell}{H}} \longrightarrow 0. \tag{295}$$

Thus all conditions of Theorem B.9 are satisfied.

Finally,

$$\tau_{\mathrm{org}} \longrightarrow 0, \qquad \tau_{\mathrm{bot}} \longrightarrow \infty, \tag{296}$$

so $\tau_{\mathrm{org}} < \tau_{\mathrm{bot}}$ for all sufficiently large H. The bottleneck theorem therefore implies that, in the same sequential limit, the classifier at τ_{bot} agrees with that of $\mathbf{T}_{k_\star}$ with probability tending to one.

Intersecting the two classifier-agreement events and using a union bound gives (277). The predictor-norm and loss conclusions in (278) follow from Theorems B.7 and B.9.

It remains to establish late-time recovery. For every fixed member of the sequence,

$$H \geq \mathrm{rank}(\mathbf{T}) \tag{297}$$

and the positive decision gap of $\mathbf{T}$ allow us to apply Proposition B.10. Hence, for every fixed $\sigma > 0$, almost surely there exists a finite time after which the classifier agrees permanently with that of $\mathbf{T}$. Enlarging this time if necessary, we may choose it so that

$$\tau_{\mathrm{rec}}^{(\sigma)} > \tau_{\mathrm{bot}}. \tag{298}$$

This proves (280) and (281).

If $A_{\mathbf{T}} > A_{k_\star}$, then on the event in (277),

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{org}})\right) = A_{\mathbf{T}} > A_{k_\star} = \mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{bot}})\right). \tag{299}$$

Since the probability of this event tends to one, this proves (283) and completes the proof. □

Unary-operation targets. We now specialize the two-factor theory to unary-operation learning. Let $f : [p] \to [p]$, and consider the dataset

$$N = C = p, \qquad \mathbf{x}_i = \mathbf{e}_i, \qquad \mathbf{y}_i = \mathbf{e}_{f(i)}, \qquad i \in [p], \tag{300}$$

where $\{\mathbf{e}_i\}_{i=1}^p$ denotes the standard basis of $\mathbb{R}^p$. The empirical input covariance is

$$\mathbf{G} = \frac{1}{p}\sum_{i=1}^{p} \mathbf{e}_i \mathbf{e}_i^\top = \frac{1}{p}\mathbf{I}_p, \tag{301}$$

so the isotropy parameter in (87) is

$$\gamma = \frac{1}{p}. \tag{302}$$

Moreover,

$$\mathbf{S} = \frac{1}{p}\sum_{i=1}^{p} \mathbf{e}_{f(i)}\mathbf{e}_i^\top, \tag{303}$$

and hence the effective target $\mathbf{T} = \mathbf{S}/\gamma$ is

$$\mathbf{T} = \sum_{i=1}^{p} \mathbf{e}_{f(i)}\mathbf{e}_i^\top, \qquad T_{ci} = \mathbf{1}\{f(i) = c\}. \tag{304}$$

Thus

$$\mathbf{T}\mathbf{e}_i = \mathbf{e}_{f(i)} \tag{305}$$

for every $i \in [p]$. Hence $\mathbf{T}$ classifies every training example correctly. Moreover, the correct class has score one and every competing class has score zero, so the minimum decision gap $\delta_{\mathbf{T}}$ from (168) satisfies

$$A_{\mathbf{T}} = 1, \qquad \delta_{\mathbf{T}} = 1. \tag{306}$$

As throughout the preceding two-factor analysis, all times below are expressed in the rescaled variable $\tau = \gamma t = t/p$.

For each output $c \in [p]$, define its fiber and occupancy by

$$\mathcal{C}_c = \{i \in [p] : f(i) = c\}, \qquad n_c = |\mathcal{C}_c|, \tag{307}$$

and let

$$\mathcal{O} = \{c \in [p] : n_c > 0\} \tag{308}$$

be the set of occupied outputs. The occupancies determine the entire nonzero singular spectrum of $\mathbf{T}$.

Lemma B.12 (Singular structure of a unary target)**.** *For every $c \in \mathcal{O}$, define*

$$\mathbf{p}_c = \mathbf{e}_c, \qquad \mathbf{q}_c = \frac{1}{\sqrt{n_c}}\sum_{i\in\mathcal{C}_c} \mathbf{e}_i. \tag{309}$$

Then

$$\mathbf{T} = \sum_{c\in\mathcal{O}} \sqrt{n_c}\,\mathbf{p}_c\mathbf{q}_c^\top \tag{310}$$

is a compact singular-value decomposition of $\mathbf{T}$. Consequently, the nonzero singular values are

$$s_c = \sqrt{n_c}, \qquad c \in \mathcal{O}, \tag{311}$$

and

$$\operatorname{rank}(\mathbf{T}) = |\mathcal{O}|. \tag{312}$$

Proof. The vectors $\{\mathbf{p}_c\}_{c\in\mathcal{O}}$ are distinct standard basis vectors and hence form an orthonormal family. The vectors $\{\mathbf{q}_c\}_{c\in\mathcal{O}}$ are also orthonormal because the fibers $\{\mathcal{C}_c\}_{c\in\mathcal{O}}$ are pairwise disjoint. Finally,

$$\begin{aligned}\sum_{c\in\mathcal{O}} \sqrt{n_c}\,\mathbf{p}_c\mathbf{q}_c^\top &= \sum_{c\in\mathcal{O}} \mathbf{e}_c \sum_{i\in\mathcal{C}_c} \mathbf{e}_i^\top \\ &= \sum_{i=1}^{p} \mathbf{e}_{f(i)}\mathbf{e}_i^\top \\ &= \mathbf{T}.\end{aligned} \tag{313}$$

Therefore (310) is a compact singular-value decomposition, and the stated singular values and rank follow immediately. □

At initialization, the classifier is at chance level. Indeed, fix $i \in [p]$. Conditional on $\mathbf{U}^{(\sigma)}(0)\mathbf{e}_i$, the p scores

$$\left(\mathbf{M}^{(\sigma)}(0)\mathbf{e}_i\right)_c = \left(\mathbf{V}^{(\sigma)}(0)^\top \mathbf{U}^{(\sigma)}(0)\mathbf{e}_i\right)_c, \qquad c \in [p], \tag{314}$$

are independent and identically distributed centered Gaussian random variables. Their common variance is positive almost surely, so the maximizing class is almost surely unique and uniformly distributed on $[p]$. Consequently, for every $\sigma > 0$,

$$\mathbb{E}\left[\mathcal{A}\left(\mathbf{M}^{(\sigma)}(0)\right)\right] = \frac{1}{p}. \tag{315}$$

Therefore, for every fixed $\varepsilon > 0$, Markov's inequality gives

$$\mathbb{P}\left(\mathcal{A}\left(\mathbf{M}^{(\sigma)}(0)\right) > \varepsilon\right) \le \frac{1}{\varepsilon p}, \tag{316}$$

and hence

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}(0)\right) \longrightarrow 0 \qquad \text{in probability as } p \to \infty. \tag{317}$$

Despite this vanishing initial accuracy, the early target-driven organization established above rapidly recovers the entire unary target classifier. Define

$$n_{\max} = \max_{c \in [p]} n_c. \tag{318}$$

Proposition B.13 (Early perfect classification for unary targets)**.** *Suppose that $p, H \to \infty$ along a sequence such that*

$$n_{\max}^{3/2} \frac{\log(2p^2H)}{H} \longrightarrow 0. \tag{319}$$

Then there exists a universal constant $K_0 > 0$ such that, for every fixed $K > K_0$, at the time

$$\tau_{\mathrm{org}} = K\sqrt{\frac{\log(2p^2H)}{H}}, \tag{320}$$

the exact gradient flow satisfies

$$\lim_{p,H\to\infty} \liminf_{\sigma \downarrow 0} \mathbb{P}\left(\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{org}})\right) = 1\right) = 1. \tag{321}$$

Moreover, for each fixed p, H, almost surely as $\sigma \downarrow 0$,

$$\left\|\mathbf{M}^{(\sigma)}(\tau_{\mathrm{org}})\right\|_{\mathrm{op}} \longrightarrow 0, \qquad \mathcal{L}\left(\mathbf{U}^{(\sigma)}(\tau_{\mathrm{org}}), \mathbf{V}^{(\sigma)}(\tau_{\mathrm{org}})\right) - \mathcal{L}\left(\mathbf{U}^{(\sigma)}(0), \mathbf{V}^{(\sigma)}(0)\right) \longrightarrow 0. \tag{322}$$

Thus perfect accuracy is reached while the end-to-end predictor remains microscopic and the loss remains asymptotically at its initial value.

Proof. For the unary data in (300),

$$R = \max_{i \in [p]} \|\mathbf{x}_i\|_2 = 1. \tag{323}$$

By (306),

$$\delta_{\mathbf{T}} = 1. \tag{324}$$

Furthermore, Lemma B.12 gives

$$\|\mathbf{T}\|_{\mathrm{op}} = \sqrt{n_{\max}}. \tag{325}$$

Hence the parameter κ from (170) satisfies

$$\kappa = \frac{R\|\mathbf{T}\|_{\mathrm{op}}}{\delta_{\mathbf{T}}} = \sqrt{n_{\max}}. \tag{326}$$

Since $N = C = p$, the logarithmic factor in the same theorem is

$$\ell = \log(2NCH) = \log(2p^2H). \tag{327}$$

Therefore

$$\kappa^3 \frac{\ell}{H} = n_{\max}^{3/2} \frac{\log(2p^2H)}{H}, \tag{328}$$

so (319) is precisely the sufficient width condition (171) from Theorem B.7.

The organization time from that theorem specializes to

$$K \frac{R}{\delta_{\mathbf{T}}} \sqrt{\frac{\ell}{H}} = K \sqrt{\frac{\log(2p^2H)}{H}} = \tau_{\mathrm{org}}. \tag{329}$$

Finally, $A_{\mathbf{T}} = 1$ by (306). Hence Theorem B.7 gives (321), together with the microscopic predictor and flat-loss conclusions in (322). □

Combining (317) with Proposition B.13, the unary model therefore has asymptotically vanishing accuracy at initialization but reaches perfect accuracy at τ_{org}, entirely within the microscopic, asymptotically constant-loss phase.

What happens after the early perfect-classification stage is more subtle. To relate the unary geometry to the general two-factor theory, define the set of outputs with maximal occupancy by

$$\mathcal{O}_{\max} = \{c \in \mathcal{O} : n_c = n_{\max}\}. \tag{330}$$

By Lemma B.12, the largest singular value is

$$s_1 = \sqrt{n_{\max}}, \qquad k_\star = |\mathcal{O}_{\max}|, \tag{331}$$

and the complete top singular block from (95) is

$$\begin{aligned} \mathbf{T}_{k_\star} &= \sum_{c \in \mathcal{O}_{\max}} \sqrt{n_{\max}}\, \mathbf{p}_c \mathbf{q}_c^\top \\ &= \sum_{c \in \mathcal{O}_{\max}} \mathbf{e}_c \sum_{i \in \mathcal{C}_c} \mathbf{e}_i^\top. \end{aligned} \tag{332}$$

If the occupied outputs have heterogeneous occupancies, then there exists $c \in \mathcal{O} \setminus \mathcal{O}_{\max}$. For every $i \in \mathcal{C}_c$,

$$\mathbf{T}_{k_\star} \mathbf{e}_i = \mathbf{0}. \tag{333}$$

Thus all output scores of $\mathbf{T}_{k_\star}$ are tied on every submaximal fiber, and its minimum decision gap is zero:

$$\delta_{k_\star} = 0. \tag{334}$$

Consequently, Theorem B.9 does not determine the classifier on these examples. In the opposite case, when all occupied outputs have the same occupancy, the entire nonzero singular spectrum is flat and

$$k_\star = \mathrm{rank}(\mathbf{T}), \qquad \mathbf{T}_{k_\star} = \mathbf{T}, \tag{335}$$

so deterministic concentration onto the top singular block does not alter the classifier at all.

Unary targets therefore expose a classification-degenerate case of the general spectral picture. For heterogeneous occupancies, the deterministic wide predictor loses all score separation on the submaximal fibers, and finite-width terms become decisive. Since the corresponding singular values $s_c = \sqrt{n_c}$ grow at different microscopic rates, random cross-mode terms associated with faster modes can eventually overwhelm the slower correct modes. We analyze this spectral interference next.

Heterogeneous unary targets: spectral interference. We now characterize how the perfect classifier from Proposition B.13 can subsequently lose accuracy. The key point is that, on fibers outside $\mathcal{O}_{\max}$, classification is eventually controlled by finite-width cross-mode fluctuations rather than by the deterministic top-block predictor. Heterogeneous occupancies amplify these fluctuations at different exponential rates.

We first specialize the explicit microscopic solution to the unary singular basis. For every occupied output $c \in \mathcal{O}$, let

$$s_c = \sqrt{n_c}, \qquad \boldsymbol{\gamma}_c = \frac{\mathbf{A}\mathbf{q}_c + \mathbf{B}\mathbf{e}_c}{\sqrt{2}}, \qquad \boldsymbol{\xi}_c = \frac{\mathbf{A}\mathbf{q}_c - \mathbf{B}\mathbf{e}_c}{\sqrt{2}}. \tag{336}$$

The vectors

$$\{\boldsymbol{\gamma}_c, \boldsymbol{\xi}_c\}_{c\in\mathcal{O}} \tag{337}$$

are mutually independent standard Gaussian vectors in $\mathbb{R}^H$. Define

$$\mathbf{X}_c(\tau) = \frac{1}{\sqrt{2}}\left(e^{s_c\tau}\boldsymbol{\gamma}_c + e^{-s_c\tau}\boldsymbol{\xi}_c\right), \qquad \mathbf{Y}_c(\tau) = \frac{1}{\sqrt{2}}\left(e^{s_c\tau}\boldsymbol{\gamma}_c - e^{-s_c\tau}\boldsymbol{\xi}_c\right). \tag{338}$$

These are precisely the c-th active singular coordinates of $\mathbf{U}_0$ and $\mathbf{V}_0$, respectively.

For $i \in \mathcal{C}_c$, decompose

$$\mathbf{e}_i = \frac{\mathbf{q}_c}{\sqrt{n_c}} + \mathbf{h}_i, \qquad \mathbf{h}_i = \mathbf{e}_i - \frac{\mathbf{q}_c}{\sqrt{n_c}}. \tag{339}$$

Since $\mathbf{T}\mathbf{e}_i = \mathbf{e}_c$ and $\mathbf{T}\mathbf{q}_c = \sqrt{n_c}\,\mathbf{e}_c$,

$$\mathbf{T}\mathbf{h}_i = \mathbf{0}, \qquad \mathbf{h}_i \in \ker\mathbf{T}. \tag{340}$$

The explicit microscopic solution therefore gives

$$\mathbf{U}_0(\tau)\mathbf{e}_i = \frac{\mathbf{X}_c(\tau)}{\sqrt{n_c}} + \mathbf{A}\mathbf{h}_i. \tag{341}$$

For an occupied output $r \in \mathcal{O}$,

$$\left(\mathbf{M}_0(\tau)\mathbf{e}_i\right)_r = \frac{\mathbf{Y}_r(\tau)^\top\mathbf{X}_c(\tau)}{\sqrt{n_c}} + \mathbf{Y}_r(\tau)^\top\mathbf{A}\mathbf{h}_i. \tag{342}$$

For an unoccupied output $r \notin \mathcal{O}$, $\mathbf{e}_r \in \ker\mathbf{T}^\top$, so its output factor remains constant under the linearized dynamics and

$$\left(\mathbf{M}_0(\tau)\mathbf{e}_i\right)_r = \frac{(\mathbf{B}\mathbf{e}_r)^\top\mathbf{X}_c(\tau)}{\sqrt{n_c}} + (\mathbf{B}\mathbf{e}_r)^\top\mathbf{A}\mathbf{h}_i. \tag{343}$$

The correct active-mode contribution is

$$\mathbf{Y}_c(\tau)^\top\mathbf{X}_c(\tau) = \frac{1}{2}\left(e^{2s_c\tau}\|\boldsymbol{\gamma}_c\|_2^2 - e^{-2s_c\tau}\|\boldsymbol{\xi}_c\|_2^2\right). \tag{344}$$

For $r \neq c$, the corresponding cross-mode term is

$$\begin{aligned}\mathbf{Y}_r(\tau)^\top\mathbf{X}_c(\tau) = \frac{1}{2}\Big(&e^{(s_r+s_c)\tau}\boldsymbol{\gamma}_r^\top\boldsymbol{\gamma}_c + e^{(s_r-s_c)\tau}\boldsymbol{\gamma}_r^\top\boldsymbol{\xi}_c \\ &- e^{(s_c-s_r)\tau}\boldsymbol{\xi}_r^\top\boldsymbol{\gamma}_c - e^{-(s_r+s_c)\tau}\boldsymbol{\xi}_r^\top\boldsymbol{\xi}_c\Big).\end{aligned} \tag{345}$$

The first term in (345) is responsible for the interference transition. A typical diagonal inner product has size H, whereas a typical cross inner product has size $\sqrt{H}$. Hence, when $s_r > s_c$, the relative scale of the faster cross-mode contribution is

$$H^{-1/2}e^{(s_r-s_c)\tau}. \tag{346}$$

It therefore becomes comparable to the correct diagonal score on a logarithmic time scale.

We now make this transition precise for unary targets with finitely many occupancy levels. Fix integers

$$1 \le m_1 < \cdots < m_K, \tag{347}$$

and define

$$\mathcal{O}_k = \{c \in [p] : n_c = m_k\}, \qquad k \in [K]. \tag{348}$$

Suppose that every occupied output belongs to one of these strata and that, as $p, H \to \infty$,

$$\frac{H}{p} \longrightarrow \alpha \in (0,\infty), \qquad \frac{|\mathcal{O}_k|}{p} \longrightarrow \rho_k \in (0,\infty), \qquad k \in [K]. \tag{349}$$

Since every input belongs to exactly one occupied fiber,

$$\sum_{k=1}^{K} m_k|\mathcal{O}_k| = p, \tag{350}$$

and therefore

$$\sum_{k=1}^{K} m_k \rho_k = 1. \tag{351}$$

Define the asymptotic fraction of inputs whose target output lies in the k-th occupancy stratum by

$$\pi_k = m_k \rho_k, \qquad \sum_{k=1}^{K} \pi_k = 1. \tag{352}$$

For $k < K$, define the interference threshold

$$\zeta_k = \frac{1}{2\left(\sqrt{m_K} - \sqrt{m_k}\right)}, \qquad \zeta_K = +\infty. \tag{353}$$

Lemma B.14 (Spectral-interference threshold)**.** *Fix $\zeta > 0$ and set*

$$\tau_\zeta = \zeta \log H. \tag{354}$$

Then, with probability tending to one as $p, H \to \infty$, the following statements hold simultaneously for all $c \in \mathcal{O}_k$ and all $i \in \mathcal{C}_c$:

1. *if $\zeta < \zeta_k$, then the correct output c has strictly largest score at time τ_ζ;*
2. *if $k < K$ and $\zeta \geq \zeta_k$, then there exists an output $r \in \mathcal{O}_K$ whose score is strictly larger than that of c.*

In particular, examples whose target output belongs to the maximal-occupancy stratum $\mathcal{O}_K$ are correctly classified at every fixed logarithmic scale $\tau = \zeta \log H$.

Proof. Fix $k \in [K]$, $c \in \mathcal{O}_k$, and $i \in \mathcal{C}_c$, and write

$$m = m_k, \qquad s = \sqrt{m}. \tag{355}$$

We first determine the scale of the correct score.

Since $H/p \to \alpha > 0$, standard Gaussian concentration followed by a union bound over $O(p)$ occupied outputs gives

$$\max_{c \in \mathcal{O}} \left| \frac{\|\boldsymbol{\gamma}_c\|_2^2}{H} - 1 \right| + \max_{c \in \mathcal{O}} \left| \frac{\|\boldsymbol{\xi}_c\|_2^2}{H} - 1 \right| \longrightarrow 0 \tag{356}$$

in probability.

Moreover, $\mathbf{h}_i$ is orthogonal to every active right singular vector $\mathbf{q}_r$. Since $\mathbf{A}$ is Gaussian, $\mathbf{A}\mathbf{h}_i$ is therefore independent of $\{\boldsymbol{\gamma}_r, \boldsymbol{\xi}_r\}_{r \in \mathcal{O}}$. Gaussian product tail bounds and a union bound over all relevant pairs give, with probability tending to one,

$$\max_{\substack{r \in \mathcal{O} \\ i \in [p]}} e^{-s_r \tau_\zeta} \left| \mathbf{Y}_r(\tau_\zeta)^\top \mathbf{A}\mathbf{h}_i \right| \leq C_0 \sqrt{H \log p} \tag{357}$$

for a constant $C_0 < \infty$ independent of p, H.

Using (342) and (344), we therefore obtain, uniformly over $c \in \mathcal{O}_k$ and $i \in \mathcal{C}_c$,

$$\left(\mathbf{M}_0(\tau_\zeta)\mathbf{e}_i\right)_c = \frac{H}{2\sqrt{m}} e^{2s\tau_\zeta} \left(1 + o_{\mathbb{P}}(1)\right). \tag{358}$$

Indeed, the decaying contribution in (344) is smaller by the factor

$$e^{-4s\tau_\zeta}, \tag{359}$$

while the nullspace contribution relative to the leading correct-score scale is

$$O_{\mathbb{P}}\left(\sqrt{\frac{\log p}{H}}\, e^{-s\tau_\zeta} \right) = o_{\mathbb{P}}(1). \tag{360}$$

We next control incorrect occupied outputs. A union bound over $O(p^2)$ pairs of occupied modes gives

$$\max_{\substack{r,c\in\mathcal{O}\\ r\neq c}} \max\left\{\left|\boldsymbol{\gamma}_r^\top\boldsymbol{\gamma}_c\right|, \left|\boldsymbol{\gamma}_r^\top\boldsymbol{\xi}_c\right|, \left|\boldsymbol{\xi}_r^\top\boldsymbol{\gamma}_c\right|, \left|\boldsymbol{\xi}_r^\top\boldsymbol{\xi}_c\right|\right\} \le C_1\sqrt{H\log p} \tag{361}$$

with probability tending to one. Since all exponential factors in (345) are bounded above by $e^{(s_r+s)\tau_\zeta}$, equations (342), (357), and (361) give

$$\left|(\mathbf{M}_0(\tau_\zeta)\mathbf{e}_i)_r\right| \le C_2 e^{(s_r+s)\tau_\zeta}\sqrt{H\log p} + C_2 e^{s_r\tau_\zeta}\sqrt{H\log p} \tag{362}$$

uniformly over occupied $r \neq c$.

Using $s_r \le \sqrt{m_K}$, the ratio of the largest incorrect occupied score to the leading correct-score scale is therefore bounded by

$$C_3\sqrt{\frac{\log p}{H}}\, e^{(\sqrt{m_K}-\sqrt{m})\tau_\zeta}\,(1+o_{\mathbb{P}}(1)). \tag{363}$$

Since $H/p \to \alpha$,

$$\sqrt{\frac{\log p}{H}}\, e^{(\sqrt{m_K}-\sqrt{m})\tau_\zeta} = O\left(\sqrt{\log H}\, H^{-1/2+\zeta(\sqrt{m_K}-\sqrt{m})}\right). \tag{364}$$

If $\zeta < \zeta_k$, the exponent of H is strictly negative, so (364) tends to zero.

For unoccupied outputs, the representation (343) and the same Gaussian bounds yield

$$\max_{r\notin\mathcal{O}}\left|(\mathbf{M}_0(\tau_\zeta)\mathbf{e}_i)_r\right| = O_{\mathbb{P}}\left(e^{s\tau_\zeta}\sqrt{H\log p}\right) \tag{365}$$

uniformly over i. Relative to (358), this is $o_{\mathbb{P}}(1)$. Hence, if $\zeta < \zeta_k$, the correct score dominates every competing score, proving the first claim. For $k = K$, every occupied competitor satisfies $s_r - s \le 0$, so the same argument proves correctness at every fixed $\zeta > 0$.

It remains to prove destruction after the threshold. Fix $k < K$ and set

$$\Delta_k = \sqrt{m_K} - \sqrt{m_k} > 0. \tag{366}$$

For $r \in \mathcal{O}_K$, the leading term of its score on an example from $\mathcal{C}_c$ is

$$\frac{1}{2\sqrt{m}} e^{(\sqrt{m_K}+\sqrt{m})\tau_\zeta}\boldsymbol{\gamma}_r^\top\boldsymbol{\gamma}_c. \tag{367}$$

Conditional on $\boldsymbol{\gamma}_c$, the random variables

$$\frac{\boldsymbol{\gamma}_r^\top\boldsymbol{\gamma}_c}{\|\boldsymbol{\gamma}_c\|_2}, \qquad r \in \mathcal{O}_K, \tag{368}$$

are independent standard Gaussians. By (349),

$$|\mathcal{O}_K| \asymp p \asymp H, \tag{369}$$

while (356) implies

$$\min_{c\in\mathcal{O}} \frac{\|\boldsymbol{\gamma}_c\|_2}{\sqrt{H}} \longrightarrow 1 \tag{370}$$

in probability.

Consequently, for every fixed $L > 0$, with probability tending to one, simultaneously for every $c \notin \mathcal{O}_K$, there exists $r = r(c) \in \mathcal{O}_K$ such that

$$\boldsymbol{\gamma}_{r(c)}^\top\boldsymbol{\gamma}_c \ge L\sqrt{H}. \tag{371}$$

Indeed, conditional on $\boldsymbol{\gamma}_c$, the probability that none of the $\Theta(H)$ candidates exceeds the fixed Gaussian threshold corresponding to L decays exponentially in H. A union bound over $O(p)$ choices of c therefore preserves convergence to zero.

For the competitor $r(c)$, the leading interference contribution divided by the leading correct-score scale is at least

$$LH^{-1/2}e^{\Delta_k\tau_\zeta} = LH^{-1/2+\zeta\Delta_k}. \tag{372}$$

If

$$\zeta > \zeta_k = \frac{1}{2\Delta_k}, \tag{373}$$

this ratio diverges. At the boundary $\zeta = \zeta_k$, it is at least L.

For this selected competitor, the remaining three cross-mode terms in (345) and the nullspace contribution in (342) are smaller than the leading interference term by at most

$$O_{\mathbb{P}}\left(\sqrt{\log p}\, e^{-\sqrt{m}\,\tau_\zeta}\right) + O_{\mathbb{P}}\left(\sqrt{\log p}\, e^{-2\sqrt{m}\,\tau_\zeta}\right) = o_{\mathbb{P}}(1). \tag{374}$$

Thus, if $\zeta > \zeta_k$, the selected competitor dominates the correct score by a diverging factor. If $\zeta = \zeta_k$, choose the fixed constant L sufficiently large, for example $L > 2$; then the same conclusion follows from (372) and (374). Therefore, uniformly over $c \in \mathcal{O}_k$ and $i \in \mathcal{C}_c$,

$$\left(\mathbf{M}_0(\tau_\zeta)\mathbf{e}_i\right)_{r(c)} > \left(\mathbf{M}_0(\tau_\zeta)\mathbf{e}_i\right)_c \tag{375}$$

with probability tending to one whenever $\zeta \geq \zeta_k$. This proves the second claim. □

The stratum-wise transition gives an explicit microscopic accuracy profile.

Theorem B.15 (Accuracy profile for heterogeneous unary targets)**.** *Under the assumptions* (349)*, define*

$$\mathcal{A}_{\mathrm{int}}(\zeta) = \pi_K + \sum_{k=1}^{K-1} \pi_k \mathbf{1}\left\{\zeta < \zeta_k\right\}. \tag{376}$$

Then, for every fixed $\zeta > 0$,

$$\mathcal{A}\left(\mathbf{M}_0(\zeta \log H)\right) \longrightarrow \mathcal{A}_{\mathrm{int}}(\zeta) \qquad \textit{in probability}. \tag{377}$$

Moreover, for the exact gradient flow,

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\zeta \log H)\right) \longrightarrow \mathcal{A}_{\mathrm{int}}(\zeta) \tag{378}$$

in probability in the sequential limit $\sigma \downarrow 0$ followed by $p, H \to \infty$.

Proof. By Lemma B.14, with probability tending to one, every example whose target output belongs to $\mathcal{O}_k$ is correctly classified at time $\zeta \log H$ precisely when either $k = K$ or $\zeta < \zeta_k$. The fraction of training examples belonging to the k-th stratum is

$$\frac{m_k|\mathcal{O}_k|}{p} \longrightarrow m_k\rho_k = \pi_k. \tag{379}$$

Summing the masses of the correctly classified strata gives (376) and proves (377).

For the exact flow, fix p, H. The time $\zeta \log H$ is finite during the inner limit $\sigma \downarrow 0$. By Theorem B.6,

$$\frac{\mathbf{M}^{(\sigma)}(\zeta \log H)}{\sigma^2} \longrightarrow \mathbf{M}_0(\zeta \log H) \qquad \text{in Frobenius norm} \tag{380}$$

almost surely. On every realization for which the strict classification inequalities in Lemma B.14 hold, the finite set of relevant score differences therefore remains of the same sign for all sufficiently small σ. Multiplication by the positive scalar σ^2 does not affect these inequalities. Hence, for sufficiently small σ, every training example has the same correctness indicator under $\mathbf{M}^{(\sigma)}$ as under $\mathbf{M}_0$ at this time. Taking the outer limit proves (378). □

Thus heterogeneous unary targets exhibit a microscopic staircase after the initial perfect-accuracy stage. Outputs with smaller occupancy lose correct classification earlier because they have a larger spectral gap to the fastest modes. More precisely, the k-th occupancy stratum loses correct classification at the logarithmic scale

$$\tau \asymp \frac{\log H}{2\left(\sqrt{m_K} - \sqrt{m_k}\right)}, \qquad k < K, \tag{381}$$

while the maximal-occupancy stratum remains correctly classified at every fixed logarithmic scale. This staircase is specific to the structured unary geometry: it is generated by finite-width cross-mode interference in a regime where the deterministic top singular block has zero classification margin on the submaximal fibers.

Random unary targets. We now draw the unary target itself at random:

$$f(1), \dots, f(p) \overset{\text{iid}}{\sim} \text{Unif}[p], \qquad \frac{H}{p} \longrightarrow \alpha \in (0, \infty). \tag{382}$$

The occupancies $\{n_c\}_{c=1}^{p}$ then follow the classical balls-into-bins model with p balls and p bins. For $m \geq 0$, let

$$N_m = |\{c \in [p] : n_c = m\}| \tag{383}$$

be the number of outputs with occupancy m. We will use the following standard occupancy asymptotics; see, e.g., Mitzenmacher & Upfal (2017, Chapter 5) and Raab & Steger (1998).

Lemma B.16 (Occupancy statistics of a random unary target)**.** *For every fixed integer $m \geq 0$,*

$$\frac{N_m}{p} \longrightarrow \frac{e^{-1}}{m!} \qquad \textit{in probability}. \tag{384}$$

Consequently, for every fixed $m \geq 1$, the fraction of training examples whose target output has occupancy m satisfies

$$\frac{mN_m}{p} \longrightarrow \frac{e^{-1}}{(m-1)!} \qquad \textit{in probability}. \tag{385}$$

Moreover,

$$\frac{\text{rank}(\mathbf{T})}{p} \longrightarrow 1 - e^{-1} \qquad \textit{in probability}, \tag{386}$$

and

$$n_{\max} = O_{\mathbb{P}}\left(\frac{\log p}{\log \log p}\right). \tag{387}$$

Proof. The limits (384) and (387) are standard results for the classical occupancy problem with p balls placed independently and uniformly into p bins; see Mitzenmacher & Upfal (2017, Chapter 5) for the Poisson occupancy asymptotics and Raab & Steger (1998) for maximum-load estimates.

Equation (385) follows immediately from (384) by multiplying by m. By Lemma B.12,

$$\text{rank}(\mathbf{T}) = p - N_0, \tag{388}$$

so (386) follows from (384) with $m = 0$. □

The random target has occupancies extending over an increasing number of levels, so the finite-stratum result above cannot be applied directly with a fixed maximal occupancy. We therefore isolate the part of the interference argument that only requires an extensive collection of faster modes. For $m \geq 0$, write

$$\mathcal{O}^{(m)} = \{c \in [p] : n_c = m\}. \tag{389}$$

Lemma B.17 (Destruction by an extensive faster stratum)**.** *Fix integers $1 \leq m < n$, and consider a sequence of unary targets satisfying*

$$\frac{H}{p} \longrightarrow \alpha \in (0, \infty), \qquad \frac{|\mathcal{O}^{(n)}|}{p} \longrightarrow \rho_n > 0. \tag{390}$$

If

$$\zeta > \frac{1}{2(\sqrt{n} - \sqrt{m})}, \tag{391}$$

then at the time

$$\tau = \zeta \log H \tag{392}$$

the microscopic predictor $\mathbf{M}_0$ misclassifies every training example whose target output has occupancy m, with probability tending to one.

The same conclusion holds for the exact gradient flow in the sequential limit $\sigma \downarrow 0$ followed by $p, H \to \infty$.

Proof. The destruction part of the proof of Lemma B.14 uses only that the correct mode has singular value $\sqrt{m}$ and that there are $\Theta(p)$ independent competing modes with a strictly larger singular value. It does not require the competing modes to belong to the maximal-occupancy stratum.

Replacing the maximal stratum in that proof by $\mathcal{O}^{(n)}$, the leading interfering score relative to the correct score has scale

$$H^{-1/2}e^{(\sqrt{n}-\sqrt{m})\tau} = H^{-1/2+\zeta(\sqrt{n}-\sqrt{m})}. \tag{393}$$

Under (391), the exponent is strictly positive, so this relative scale diverges. The same conditional Gaussian maximum argument as in Lemma B.14 then produces, uniformly over all outputs of occupancy m, a faster competitor whose score exceeds the correct score with probability tending to one.

For the exact gradient flow, fix p, H. The time (392) is finite during the inner limit $\sigma \downarrow 0$. Hence Theorem B.6 transfers the strict score inequalities from $\mathbf{M}_0$ to $\mathbf{M}^{(\sigma)}$ for all sufficiently small σ. □

We can now combine the early target-driven organization with the random occupancy statistics and the interference mechanism to obtain the random-unary accuracy dynamics.

Theorem B.18 [4.3] **(Random unary target).** *Consider the random unary model* (382)*. For every $\alpha \in (0, \infty)$, there exist deterministic times*

$$\tau_{\mathrm{org}} < \tau_{\mathrm{int}} \tag{394}$$

for all sufficiently large p, H, such that, in the sequential limit $\sigma \downarrow 0$ followed by $p, H \to \infty$ with $H/p \to \alpha$,

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}(0)\right) \longrightarrow 0, \qquad \mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{org}})\right) \longrightarrow 1, \qquad \mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{int}})\right) \longrightarrow 0 \tag{395}$$

in probability.

The organization time may be chosen as

$$\tau_{\mathrm{org}} = K\sqrt{\frac{\log(2p^2H)}{H}} \tag{396}$$

for any sufficiently large fixed universal constant K. Moreover, for each fixed p, H and fixed realization of the random target, almost surely as $\sigma \downarrow 0$,

$$\left\|\mathbf{M}^{(\sigma)}(\tau)\right\|_{\mathrm{op}} \longrightarrow 0, \qquad \mathcal{L}\left(\mathbf{U}^{(\sigma)}(\tau), \mathbf{V}^{(\sigma)}(\tau)\right) - \mathcal{L}\left(\mathbf{U}^{(\sigma)}(0), \mathbf{V}^{(\sigma)}(0)\right) \longrightarrow 0 \tag{397}$$

for

$$\tau \in \{\tau_{\mathrm{org}}, \tau_{\mathrm{int}}\}. \tag{398}$$

If, in addition,

$$\alpha > 1 - e^{-1}, \tag{399}$$

then

$$\mathbb{P}_f\left(H \geq \mathrm{rank}(\mathbf{T})\right) \longrightarrow 1, \tag{400}$$

where $\mathbb{P}_f$ denotes probability over the random target. On this realizability event, for every fixed $\sigma > 0$, almost surely over the Gaussian initialization there exists a finite random time $\tau_{\mathrm{rec}}^{(\sigma)} > \tau_{\mathrm{int}}$ such that

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau)\right) = 1 \qquad \textit{for every } \tau \geq \tau_{\mathrm{rec}}^{(\sigma)}. \tag{401}$$

Proof. The initial convergence in (395) follows from (317).

Early target-driven organization. By Lemma B.16,

$$n_{\max} = O_{\mathbb{P}}\left(\frac{\log p}{\log\log p}\right). \tag{402}$$

Since $H/p \to \alpha > 0$,

$$n_{\max}^{3/2}\frac{\log(2p^2H)}{H} \longrightarrow 0 \qquad \text{in probability.} \tag{403}$$

Thus, with probability tending to one over the random target, the width condition of Proposition B.13 holds. For a sufficiently large fixed universal constant K, at

$$\tau_{\mathrm{org}} = K\sqrt{\frac{\log(2p^2H)}{H}}, \tag{404}$$

the exact gradient flow therefore has perfect accuracy with probability tending to one in the sequential limit $\sigma \downarrow 0$ followed by $p, H \to \infty$. Moreover,

$$\tau_{\mathrm{org}} \longrightarrow 0, \tag{405}$$

and the microscopic-predictor and flat-loss conclusions follow from the same proposition.

Microscopic interference collapse. We first obtain a collapse bound for a fixed occupancy cutoff. Fix an integer $K \geq 1$, and define

$$\zeta_K = \frac{1}{\sqrt{K+1} - \sqrt{K}}. \tag{406}$$

Then

$$\zeta_K > \frac{1}{2(\sqrt{K+1} - \sqrt{K})}. \tag{407}$$

By (384),

$$\frac{|\mathcal{O}^{(K+1)}|}{p} = \frac{N_{K+1}}{p} \longrightarrow \frac{e^{-1}}{(K+1)!} > 0 \qquad \text{in probability.} \tag{408}$$

Thus occupancy $K+1$ supplies an extensive collection of faster modes with probability tending to one.

Set

$$\tau_K = \zeta_K \log H. \tag{409}$$

For every $1 \leq m \leq K$,

$$\zeta_K > \frac{1}{2(\sqrt{K+1} - \sqrt{m})}, \tag{410}$$

because

$$\sqrt{K+1} - \sqrt{m} \geq \sqrt{K+1} - \sqrt{K}. \tag{411}$$

Lemma B.17, applied with $n = K+1$, therefore implies that, with probability tending to one in the sequential limit, every training example whose target output has occupancy

$$1 \leq m \leq K \tag{412}$$

is misclassified at time τ_K. Since K is fixed, the finitely many occupancy levels can be handled simultaneously by a union bound.

Consequently,

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_K)\right) \leq \frac{1}{p}\sum_{m>K} mN_m + o_{\mathbb{P}}(1) \tag{413}$$

in the sequential limit.

Since every input belongs to exactly one occupied fiber,

$$\sum_{m\geq 1} mN_m = p. \tag{414}$$

For every fixed K, Lemma B.16 therefore gives

$$\begin{aligned} \frac{1}{p}\sum_{m>K} mN_m &= 1 - \sum_{m=1}^{K} \frac{mN_m}{p} \\ &\longrightarrow 1 - e^{-1}\sum_{m=1}^{K} \frac{1}{(m-1)!} \\ &= e^{-1}\sum_{j=K}^{\infty} \frac{1}{j!} =: r_K \end{aligned} \tag{415}$$

in probability, where

$$r_K \longrightarrow 0 \qquad \text{as } K \to \infty. \tag{416}$$

Hence, for every fixed K and every $\varepsilon > 0$,

$$\lim_{\substack{p,H\to\infty \\ H/p\to\alpha}} \limsup_{\sigma\downarrow 0} \mathbb{P}\left(\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_K)\right) > r_K + \varepsilon\right) = 0. \tag{417}$$

We now choose the occupancy cutoff to increase sufficiently slowly with the system size. Let

$$K_j \uparrow \infty \tag{418}$$

be a deterministic sequence such that

$$r_{K_j} < \frac{1}{2j}. \tag{419}$$

For each j, the fixed-K_j result above allows us to choose a deterministic threshold p_j such that, along $H/p \to \alpha$,

$$\limsup_{\sigma\downarrow 0} \mathbb{P}\left(\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_{K_j})\right) > \frac{1}{j}\right) < \frac{1}{j} \tag{420}$$

whenever $p \geq p_j$. We may choose

$$p_1 < p_2 < \cdots. \tag{421}$$

Define

$$K(p) = K_j \qquad \text{for } p_j \leq p < p_{j+1}. \tag{422}$$

Then

$$K(p) \longrightarrow \infty. \tag{423}$$

Set

$$\tau_{\mathrm{int}} = \zeta_{K(p)} \log H. \tag{424}$$

By construction,

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{int}})\right) \longrightarrow 0 \tag{425}$$

in probability in the sequential limit $\sigma \downarrow 0$ followed by $p, H \to \infty$.

Since $\zeta_{K(p)} \geq \zeta_1 > 0$,

$$\tau_{\mathrm{int}} \geq \zeta_1 \log H \longrightarrow \infty, \tag{426}$$

whereas $\tau_{\mathrm{org}} \to 0$. Hence

$$\tau_{\mathrm{org}} < \tau_{\mathrm{int}} \tag{427}$$

for all sufficiently large p, H.

For every fixed p, H, τ_{int} is nevertheless a finite deterministic time during the inner limit $\sigma \downarrow 0$. Therefore Theorem B.6, applied with a finite horizon containing τ_{int}, gives

$$\left\|\mathbf{M}^{(\sigma)}(\tau_{\mathrm{int}})\right\|_{\mathrm{op}} \longrightarrow 0 \tag{428}$$

and

$$\mathcal{L}\left(\mathbf{U}^{(\sigma)}(\tau_{\mathrm{int}}), \mathbf{V}^{(\sigma)}(\tau_{\mathrm{int}})\right) - \mathcal{L}\left(\mathbf{U}^{(\sigma)}(0), \mathbf{V}^{(\sigma)}(0)\right) \longrightarrow 0 \tag{429}$$

almost surely as $\sigma \downarrow 0$. Together with the corresponding conclusions at τ_{org}, this proves (397).

Final recovery. By (386),

$$\frac{\mathrm{rank}(\mathbf{T})}{p} \longrightarrow 1 - e^{-1} \qquad \text{in probability.} \tag{430}$$

Suppose now that

$$\alpha > 1 - e^{-1}. \tag{431}$$

Choose $\varepsilon > 0$ such that

$$1 - e^{-1} + \varepsilon < \alpha. \tag{432}$$

Since $H/p \to \alpha$, for all sufficiently large p, H,

$$\frac{H}{p} > 1 - e^{-1} + \varepsilon, \tag{433}$$

while

$$\mathbb{P}_f\left(\frac{\operatorname{rank}(\mathbf{T})}{p} < 1 - e^{-1} + \varepsilon\right) \longrightarrow 1. \tag{434}$$

Therefore

$$\mathbb{P}_f\left(H \geq \operatorname{rank}(\mathbf{T})\right) \longrightarrow 1, \tag{435}$$

which proves (400).

Every unary target has $\delta_{\mathbf{T}} = 1$. Hence, on the realizability event in (435), Proposition B.10 applies. For every fixed $\sigma > 0$, almost surely over the Gaussian initialization there exists a finite random time after which $\mathbf{M}^{(\sigma)}$ permanently induces the classifier of $\mathbf{T}$. Since $A_{\mathbf{T}} = 1$, its training accuracy is then permanently equal to one. Enlarging the recovery time if necessary, it may be chosen so that

$$\tau_{\mathrm{rec}}^{(\sigma)} > \tau_{\mathrm{int}}. \tag{436}$$

This proves (401) and completes the proof. □

The theorem separates two distinct roles of width. Any fixed $\alpha > 0$ is sufficient for the microscopic target-driven organization and the subsequent interference-induced collapse: the classifier first becomes perfect and later loses essentially all accuracy while both the predictor and the change in loss remain microscopic in the small-initialization limit. The stronger condition $\alpha > 1 - e^{-1}$ is needed only for permanent late-time recovery, because it makes the random unary target realizable with probability tending to one.

Permutation targets: permanent perfect accuracy. We finally consider the opposite extreme of the unary setting, where $f : [p] \to [p]$ is a permutation. Then every output has occupancy one:

$$n_c = 1, \qquad c \in [p]. \tag{437}$$

By Lemma B.12, the target has rank p and all of its singular values are equal to one. Hence

$$s_1 = \cdots = s_p = 1, \qquad k_\star = p, \qquad \mathbf{T}_{k_\star} = \mathbf{T}. \tag{438}$$

Thus the deterministic top-block concentration discussed above does not change the predictor's target geometry: the complete top singular block is already the full target. At the same time, there are no faster and slower occupancy modes, so the spectral-interference mechanism responsible for the accuracy collapse of heterogeneous unary targets is absent. We will show that, in this flat-spectrum case, the training accuracy is instead asymptotically nondecreasing throughout the entire trajectory and, once it reaches one, remains equal to one permanently.

We first reduce an arbitrary permutation target to the identity. Let $\mathbf{P}_f \in \mathbb{R}^{p\times p}$ be the permutation matrix associated with f, so that

$$\mathbf{P}_f \mathbf{e}_i = \mathbf{e}_{f(i)}, \qquad \mathbf{T} = \mathbf{P}_f. \tag{439}$$

Define

$$\widetilde{\mathbf{V}} = \mathbf{V}\mathbf{P}_f, \qquad \widetilde{\mathbf{M}} = \widetilde{\mathbf{V}}^\top \mathbf{U} = \mathbf{P}_f^\top \mathbf{M}. \tag{440}$$

Since $\mathbf{P}_f$ is orthogonal,

$$\|\mathbf{M} - \mathbf{P}_f\|_{\mathrm{F}} = \left\|\widetilde{\mathbf{M}} - \mathbf{I}_p\right\|_{\mathrm{F}}. \tag{441}$$

The gradient-flow equations are equivariant under this change of output coordinates. Moreover, multiplication by $\mathbf{P}_f^\top$ merely relabels the output scores, sending the correct class $f(i)$ to i, and therefore leaves training accuracy unchanged. Finally, under isotropic Gaussian initialization, $\mathbf{V}\mathbf{P}_f$ has the same distribution as $\mathbf{V}$. It is therefore enough to analyze the identity target

$$\mathbf{T} = \mathbf{I}_p. \tag{442}$$

For the identity target, the rescaled gradient flow introduced above becomes

$$\dot{\mathbf{U}} = \mathbf{V}\left(\mathbf{I}_p - \mathbf{V}^\top \mathbf{U}\right), \qquad \dot{\mathbf{V}} = \mathbf{U}\left(\mathbf{I}_p - \mathbf{U}^\top \mathbf{V}\right). \tag{443}$$

It is useful to decompose the two factors into their symmetric and antisymmetric combinations,

$$\mathbf{W}_+ = \frac{\mathbf{U}+\mathbf{V}}{2}, \qquad \mathbf{W}_- = \frac{\mathbf{U}-\mathbf{V}}{2}. \tag{444}$$

Under Gaussian initialization

$$\mathbf{U}^{(\sigma)}(0) = \sigma\mathbf{A}, \qquad \mathbf{V}^{(\sigma)}(0) = \sigma\mathbf{B}, \tag{445}$$

with $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{H\times p}$ independent standard Gaussian matrices, we may equivalently write

$$\mathbf{W}_+^{(\sigma)}(0) = \frac{\sigma}{\sqrt{2}}\mathbf{X}, \qquad \mathbf{W}_-^{(\sigma)}(0) = \frac{\sigma}{\sqrt{2}}\mathbf{Y}, \tag{446}$$

where $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{H\times p}$ are again independent standard Gaussian matrices.

Near the origin, the linearized dynamics satisfy

$$\dot{\mathbf{W}}_+ = \mathbf{W}_+, \qquad \dot{\mathbf{W}}_- = -\mathbf{W}_-, \tag{447}$$

so the symmetric component grows as e^{τ}, whereas the antisymmetric component decays as $e^{-\tau}$. Introduce

$$\varepsilon = \frac{H\sigma^2}{2}, \qquad q(\tau) = \varepsilon e^{2\tau}. \tag{448}$$

The variable q measures the scale of the growing symmetric component: $q \ll 1$ corresponds to the small-q regime in which the factor dynamics remain close to their linearization, whereas $q = O(1)$ marks the onset of nonlinear evolution.

Since $n_{\max} = 1$, the width condition in Proposition B.13 is automatically satisfied when

$$\frac{H}{p} \longrightarrow \alpha \in (0,\infty). \tag{449}$$

Thus, for every sufficiently large fixed universal constant K, perfect accuracy is reached at the organization time

$$\tau_{\mathrm{org}} = K\sqrt{\frac{\log(2p^2H)}{H}}. \tag{450}$$

For each fixed p, H,

$$q(\tau_{\mathrm{org}}) = \frac{H\sigma^2}{2}e^{2\tau_{\mathrm{org}}} \longrightarrow 0 \qquad \text{as } \sigma \downarrow 0. \tag{451}$$

Hence the perfect classifier emerges strictly inside the small-q regime in the vanishing-initialization limit.

To prove that accuracy never subsequently decreases, we use two overlapping controls. First, we show that every correct-versus-incorrect classification margin is strictly increasing throughout a fixed small-q interval. Second, after shifting time so that q remains fixed as $\sigma \downarrow 0$, we show that the entire nonlinear tail converges uniformly to an explicit balanced trajectory whose classification margins are uniformly positive. The final permutation theorem will follow by combining these two regimes.

We first collect the high-dimensional Gaussian geometry of the initialization needed to control all classification margins simultaneously. Write

$$\mathbf{X} = \begin{pmatrix}\boldsymbol{\xi}_1 & \cdots & \boldsymbol{\xi}_p\end{pmatrix}, \qquad \mathbf{Y} = \begin{pmatrix}\boldsymbol{\eta}_1 & \cdots & \boldsymbol{\eta}_p\end{pmatrix}, \tag{452}$$

where $\boldsymbol{\xi}_i, \boldsymbol{\eta}_i \in \mathbb{R}^H$, and define the normalized Gram matrices

$$\mathbf{G}_+ = \frac{1}{H}\mathbf{X}^\top\mathbf{X}, \qquad \mathbf{G}_- = \frac{1}{H}\mathbf{Y}^\top\mathbf{Y}, \qquad \mathbf{G}_\pm = \frac{1}{H}\mathbf{X}^\top\mathbf{Y}. \tag{453}$$

Lemma B.19 (Gaussian initialization geometry)**.** *Suppose that*

$$\frac{H}{p} \longrightarrow \alpha \in (0,\infty). \tag{454}$$

Then there exist a constant $C_\alpha < \infty$ and a deterministic sequence

$$\eta_p = C_\alpha\sqrt{\frac{\log p}{H}} \longrightarrow 0 \tag{455}$$

such that, with probability tending to one,

$$\max_{i\in[p]} |(\mathbf{G}_+)_{ii} - 1| \le \eta_p, \qquad \max_{i\in[p]} |(\mathbf{G}_-)_{ii} - 1| \le \eta_p, \tag{456}$$

and

$$\max_{i\neq j} |(\mathbf{G}_+)_{ij}| \le \eta_p, \qquad \max_{i\neq j} |(\mathbf{G}_-)_{ij}| \le \eta_p, \qquad \max_{i,j} |(\mathbf{G}_\pm)_{ij}| \le \eta_p. \tag{457}$$

Moreover,

$$\frac{\|\mathbf{X}\|_{\mathrm{op}}}{\sqrt{H}} + \frac{\|\mathbf{Y}\|_{\mathrm{op}}}{\sqrt{H}} \le C_\alpha \tag{458}$$

with probability tending to one. In particular, after increasing C_α if necessary,

$$\|\mathbf{G}_+\|_{\mathrm{op}} + \|\mathbf{G}_-\|_{\mathrm{op}} + \|\mathbf{G}_\pm\|_{\mathrm{op}} \le C_\alpha^2 \tag{459}$$

with probability tending to one.

Proof. Let $\mathbf{g} \in \mathbb{R}^H$ be a standard Gaussian vector. The standard chi-square concentration bound implies that there exists a universal constant $c > 0$ such that, for $0 < t < 1$,

$$\mathbb{P}\left(\left|\frac{\|\mathbf{g}\|_2^2}{H} - 1\right| > t\right) \le 2e^{-cHt^2}. \tag{460}$$

Likewise, if $\mathbf{g}, \mathbf{h} \in \mathbb{R}^H$ are independent standard Gaussian vectors, then, after changing the universal constant c if necessary,

$$\mathbb{P}\left(\frac{|\mathbf{g}^\top \mathbf{h}|}{H} > t\right) \le 2e^{-cHt^2}, \qquad 0 < t < 1. \tag{461}$$

The columns $\{\boldsymbol{\xi}_i\}_{i=1}^p$ are independent standard Gaussian vectors, as are $\{\boldsymbol{\eta}_i\}_{i=1}^p$, and the two families are independent. Applying (460) to the $2p$ column norms and (461) to the $O(p^2)$ relevant inner products, followed by a union bound, gives (456) and (457) upon choosing

$$t = C_\alpha \sqrt{\frac{\log p}{H}} \tag{462}$$

with C_α sufficiently large. Since $H/p \to \alpha > 0$, the right-hand side of (462) tends to zero.

For the operator norms, the standard Gaussian matrix estimate gives

$$\mathbb{P}\left(\|\mathbf{X}\|_{\mathrm{op}} > \sqrt{H} + \sqrt{p} + u\right) \le e^{-u^2/2}, \qquad u \ge 0, \tag{463}$$

and the same estimate holds for $\mathbf{Y}$. Taking $u = \sqrt{H}$ and using $p/H \to \alpha^{-1}$ shows that (458) holds with probability tending to one for some finite constant depending only on α.

Finally,

$$\|\mathbf{G}_+\|_{\mathrm{op}} = \frac{\|\mathbf{X}\|_{\mathrm{op}}^2}{H}, \qquad \|\mathbf{G}_-\|_{\mathrm{op}} = \frac{\|\mathbf{Y}\|_{\mathrm{op}}^2}{H}, \tag{464}$$

while

$$\|\mathbf{G}_\pm\|_{\mathrm{op}} \le \frac{\|\mathbf{X}\|_{\mathrm{op}}\|\mathbf{Y}\|_{\mathrm{op}}}{H}. \tag{465}$$

Thus (459) follows after increasing C_α if necessary. □

The Gaussian geometry above provides the diagonal dominance needed to control the early dynamics uniformly over all classes. In the small-q regime, the growing symmetric factor remains close to its linearized trajectory $e^{\tau}\mathbf{W}_+(0)$, while the antisymmetric factor remains close to $e^{-\tau}\mathbf{W}_-(0)$. Consequently, the leading contribution to $\dot{\mathbf{M}}$ is governed by the Gram matrices $\mathbf{G}_+$ and $\mathbf{G}_-$, whose diagonal entries are close to one and whose off-diagonal entries are uniformly small. We now show that the nonlinear corrections are only quadratic in q, and therefore cannot overturn this diagonal dominance while q is sufficiently small. This implies that every correct-versus-incorrect classification margin increases throughout a fixed small-q regime.

Lemma B.20 (Small-q margin monotonicity)**.** *Suppose that*

$$\frac{H}{p} \longrightarrow \alpha \in (0, \infty), \tag{466}$$

and consider the identity target $\mathbf{T} = \mathbf{I}_p$*. There exists a constant*

$$q_{\mathrm{sm}} = q_{\mathrm{sm}}(\alpha) > 0 \tag{467}$$

such that, with probability tending to one over the Gaussian matrices $\mathbf{X}, \mathbf{Y}$*, the following statement holds simultaneously for every* $\sigma > 0$ *satisfying*

$$\varepsilon = \frac{H\sigma^2}{2} \leq q_{\mathrm{sm}}. \tag{468}$$

For $i \in [p]$ *and* $c \neq i$*, define the correct-versus-incorrect classification margin*

$$\Delta_{ci}(\tau) = M_{ii}(\tau) - M_{ci}(\tau). \tag{469}$$

Then

$$\frac{d}{d\tau}\Delta_{ci}(\tau) > 0 \tag{470}$$

simultaneously for every $i \in [p]$*,* $c \neq i$*, and every* $\tau \geq 0$ *such that*

$$q(\tau) = \varepsilon e^{2\tau} \leq q_{\mathrm{sm}}. \tag{471}$$

Consequently, the training accuracy

$$\tau \longmapsto \mathcal{A}\big(\mathbf{M}^{(\sigma)}(\tau)\big) \tag{472}$$

is nondecreasing throughout the interval on which $q(\tau) \leq q_{\mathrm{sm}}$*.*

Proof. We work on the high-probability event from Lemma B.19. For all sufficiently large p, H, this event satisfies

$$\eta_p \leq \frac{1}{8}, \qquad \frac{\|\mathbf{X}\|_{\mathrm{op}}}{\sqrt{H}} + \frac{\|\mathbf{Y}\|_{\mathrm{op}}}{\sqrt{H}} \leq C, \tag{473}$$

where $C = C_\alpha < \infty$ depends only on α. We show that the desired conclusion holds on this event whenever $q_{\mathrm{sm}} > 0$ is sufficiently small depending only on C.

Recall the decomposition

$$\mathbf{W}_+ = \frac{\mathbf{U} + \mathbf{V}}{2}, \qquad \mathbf{W}_- = \frac{\mathbf{U} - \mathbf{V}}{2}, \tag{474}$$

and define the symmetric and antisymmetric parts of the end-to-end predictor by

$$\mathbf{M}_{\mathrm{s}} = \frac{\mathbf{M} + \mathbf{M}^\top}{2}, \qquad \mathbf{M}_{\mathrm{a}} = \frac{\mathbf{M}^\top - \mathbf{M}}{2}. \tag{475}$$

Using the identity-target flow (443), a direct calculation gives

$$\dot{\mathbf{W}}_+ = \mathbf{W}_+ - \mathbf{W}_+\mathbf{M}_{\mathrm{s}} - \mathbf{W}_-\mathbf{M}_{\mathrm{a}}, \tag{476}$$

and

$$\dot{\mathbf{W}}_- = -\mathbf{W}_- + \mathbf{W}_+\mathbf{M}_{\mathrm{a}} + \mathbf{W}_-\mathbf{M}_{\mathrm{s}}. \tag{477}$$

Moreover,

$$\|\mathbf{M}_{\mathrm{s}}\|_{\mathrm{op}},\, \|\mathbf{M}_{\mathrm{a}}\|_{\mathrm{op}} \leq \|\mathbf{M}\|_{\mathrm{op}} \leq \left(\|\mathbf{W}_+\|_{\mathrm{op}} + \|\mathbf{W}_-\|_{\mathrm{op}}\right)^2. \tag{478}$$

Define

$$w(\tau) = \|\mathbf{W}_+(\tau)\|_{\mathrm{op}} + \|\mathbf{W}_-(\tau)\|_{\mathrm{op}}. \tag{479}$$

The nonlinear terms in each of (476) and (477) therefore have operator norm at most

$$w(\tau)^3. \tag{480}$$

We compare the exact factors with the linearized trajectories

$$\overline{\mathbf{W}}_+(\tau) = e^{\tau}\mathbf{W}_+(0) = \sqrt{\frac{q(\tau)}{H}}\,\mathbf{X}, \qquad \overline{\mathbf{W}}_-(\tau) = e^{-\tau}\mathbf{W}_-(0) = \sqrt{\frac{r(\tau)}{H}}\,\mathbf{Y}, \tag{481}$$

where

$$r(\tau) = \varepsilon e^{-2\tau}. \tag{482}$$

Since $\tau \geq 0$,

$$0 < r(\tau) \leq q(\tau). \tag{483}$$

On the event (473),

$$\|\overline{\mathbf{W}}_+(\tau)\|_{\mathrm{op}} + \|\overline{\mathbf{W}}_-(\tau)\|_{\mathrm{op}} \leq C\sqrt{q(\tau)}. \tag{484}$$

We next prove the bootstrap estimate

$$w(\tau) \leq 2C\sqrt{q(\tau)} \qquad \text{whenever } q(\tau) \leq q_{\mathrm{sm}}. \tag{485}$$

Work on the maximal existence interval of the exact flow and set

$$D = 2C. \tag{486}$$

Suppose temporarily that

$$w(u) \leq D\sqrt{q(u)} \qquad \text{for } 0 \leq u \leq \tau. \tag{487}$$

Variation of constants in (476) then gives

$$\begin{aligned} \left\|\mathbf{W}_+(\tau) - \overline{\mathbf{W}}_+(\tau)\right\|_{\mathrm{op}} &\leq D^3 \int_0^\tau e^{\tau-u} q(u)^{3/2}\, du \\ &\leq \frac{D^3}{2} q(\tau)^{3/2}, \end{aligned} \tag{488}$$

because $q(u) = \varepsilon e^{2u}$. Likewise,

$$\begin{aligned} \left\|\mathbf{W}_-(\tau) - \overline{\mathbf{W}}_-(\tau)\right\|_{\mathrm{op}} &\leq D^3 \int_0^\tau e^{-(\tau-u)} q(u)^{3/2}\, du \\ &\leq \frac{D^3}{4} q(\tau)^{3/2}. \end{aligned} \tag{489}$$

Combining (484), (488), and (489), we obtain

$$w(\tau) \leq C\sqrt{q(\tau)} + \frac{3D^3}{4} q(\tau)^{3/2}. \tag{490}$$

Choose q_{sm} small enough that

$$q_{\mathrm{sm}} \leq \frac{1}{12C^2}. \tag{491}$$

Then, whenever $q(\tau) \leq q_{\mathrm{sm}}$, the right-hand side of (490) is at most

$$\frac{3}{2} C\sqrt{q(\tau)} < D\sqrt{q(\tau)}. \tag{492}$$

Since the bootstrap bound holds strictly at $\tau = 0$, a first-exit argument proves (485) throughout the portion of the maximal existence interval on which $q \leq q_{\mathrm{sm}}$. The bound keeps both factors finite there, so the standard continuation theorem for finite-dimensional ODEs shows that the solution exists through the entire interval $q(\tau) \leq q_{\mathrm{sm}}$.

We now turn to the classification margins. Differentiating

$$\mathbf{M} = \mathbf{V}^\top \mathbf{U} \tag{493}$$

along the exact flow gives

$$\begin{aligned} \dot{\mathbf{M}} &= (\mathbf{I}_p - \mathbf{M})\mathbf{U}^\top \mathbf{U} + \mathbf{V}^\top \mathbf{V}(\mathbf{I}_p - \mathbf{M}) \\ &= \mathbf{U}^\top \mathbf{U} + \mathbf{V}^\top \mathbf{V} - \mathbf{M}\mathbf{U}^\top \mathbf{U} - \mathbf{V}^\top \mathbf{V}\mathbf{M}. \end{aligned} \tag{494}$$

Since $\mathbf{U} = \mathbf{W}_+ + \mathbf{W}_-$ and $\mathbf{V} = \mathbf{W}_+ - \mathbf{W}_-$,

$$\mathbf{U}^\top \mathbf{U} + \mathbf{V}^\top \mathbf{V} = 2\mathbf{W}_+^\top \mathbf{W}_+ + 2\mathbf{W}_-^\top \mathbf{W}_-. \tag{495}$$

For either sign,

$$\begin{aligned}\left\|\mathbf{W}_{\pm}^{\top}\mathbf{W}_{\pm}-\overline{\mathbf{W}}_{\pm}^{\top}\overline{\mathbf{W}}_{\pm}\right\|_{\mathrm{op}} \\ \leq \left(\|\mathbf{W}_{\pm}\|_{\mathrm{op}}+\|\overline{\mathbf{W}}_{\pm}\|_{\mathrm{op}}\right)\left\|\mathbf{W}_{\pm}-\overline{\mathbf{W}}_{\pm}\right\|_{\mathrm{op}}.\end{aligned} \tag{496}$$

Using (484), (485), (488), and (489), we obtain

$$\left\|\mathbf{W}_{+}^{\top}\mathbf{W}_{+}-\overline{\mathbf{W}}_{+}^{\top}\overline{\mathbf{W}}_{+}\right\|_{\mathrm{op}} \leq \frac{(D+C)D^3}{2}q^2 \tag{497}$$

and

$$\left\|\mathbf{W}_{-}^{\top}\mathbf{W}_{-}-\overline{\mathbf{W}}_{-}^{\top}\overline{\mathbf{W}}_{-}\right\|_{\mathrm{op}} \leq \frac{(D+C)D^3}{4}q^2. \tag{498}$$

Furthermore,

$$\|\mathbf{U}\|_{\mathrm{op}},\ \|\mathbf{V}\|_{\mathrm{op}} \leq D\sqrt{q}, \tag{499}$$

and hence

$$\|\mathbf{M}\|_{\mathrm{op}},\ \|\mathbf{U}^{\top}\mathbf{U}\|_{\mathrm{op}},\ \|\mathbf{V}^{\top}\mathbf{V}\|_{\mathrm{op}} \leq D^2 q. \tag{500}$$

Therefore

$$\left\|\mathbf{M}\mathbf{U}^{\top}\mathbf{U}+\mathbf{V}^{\top}\mathbf{V}\mathbf{M}\right\|_{\mathrm{op}} \leq 2D^4q^2. \tag{501}$$

Combining the preceding estimates with (494) gives

$$\left\|\dot{\mathbf{M}}-2\overline{\mathbf{W}}_{+}^{\top}\overline{\mathbf{W}}_{+}-2\overline{\mathbf{W}}_{-}^{\top}\overline{\mathbf{W}}_{-}\right\|_{\mathrm{op}} \leq C_1q^2, \tag{502}$$

where

$$C_1=\frac{3}{2}(D+C)D^3+2D^4. \tag{503}$$

Since $D=2C$ and $C=C_\alpha$, the constant C_1 depends only on α.

By (481),

$$2\overline{\mathbf{W}}_{+}^{\top}\overline{\mathbf{W}}_{+}=2q\mathbf{G}_{+}, \qquad 2\overline{\mathbf{W}}_{-}^{\top}\overline{\mathbf{W}}_{-}=2r\mathbf{G}_{-}. \tag{504}$$

Thus we may write

$$\dot{\mathbf{M}}=2q\mathbf{G}_{+}+2r\mathbf{G}_{-}+\mathbf{E}, \qquad \|\mathbf{E}\|_{\mathrm{op}} \leq C_1q^2. \tag{505}$$

Fix $i\in[p]$ and $c\neq i$. From (469) and (505),

$$\begin{aligned}\dot{\Delta}_{ci}=2q\left[(\mathbf{G}_{+})_{ii}-(\mathbf{G}_{+})_{ci}\right]+2r\left[(\mathbf{G}_{-})_{ii}-(\mathbf{G}_{-})_{ci}\right] \\ +(\mathbf{e}_i-\mathbf{e}_c)^{\top}\mathbf{E}\mathbf{e}_i.\end{aligned} \tag{506}$$

The Gaussian geometry event gives

$$(\mathbf{G}_{\pm})_{ii}-(\mathbf{G}_{\pm})_{ci} \geq 1-2\eta_p \geq \frac{3}{4}. \tag{507}$$

Moreover,

$$\left|(\mathbf{e}_i-\mathbf{e}_c)^{\top}\mathbf{E}\mathbf{e}_i\right| \leq \sqrt{2}\|\mathbf{E}\|_{\mathrm{op}} \leq \sqrt{2}C_1q^2. \tag{508}$$

Since $r\geq 0$,

$$\dot{\Delta}_{ci} \geq \frac{3}{2}q-\sqrt{2}C_1q^2. \tag{509}$$

Decrease q_{sm}, if necessary, so that in addition to (491),

$$\sqrt{2}C_1q_{\mathrm{sm}}<\frac{1}{2}. \tag{510}$$

Then, whenever $q(\tau)\leq q_{\mathrm{sm}}$,

$$\dot{\Delta}_{ci}(\tau) \geq q(\tau)>0 \tag{511}$$

simultaneously for every $i\in[p]$ and $c\neq i$.

Thus every correct-versus-incorrect margin is strictly increasing throughout the small-q regime. Once a training example becomes correctly classified, all of its margins are positive and remain positive for the rest of this regime. Hence the set of correctly classified training examples can only grow, and the training accuracy is nondecreasing whenever $q(\tau)\leq q_{\mathrm{sm}}$. □

The lemma controls the trajectory from initialization through a fixed small-q interval. In particular, once the early target-driven organization reaches perfect accuracy, that perfect classifier cannot be lost before $q = q_{\mathrm{sm}}$. We next identify the nonlinear trajectory selected by vanishing initialization and use it to control all later times.

The small-q monotonicity lemma controls the trajectory from initialization through a fixed microscopic scale. To control all later times, we now identify the nonlinear trajectory selected by vanishing initialization. The key observation is that the antisymmetric component $\mathbf{W}_-$ is a decaying direction near the origin, whereas the growing component $\mathbf{W}_+$ selects, after an appropriate time shift, a balanced trajectory with $\mathbf{U} = \mathbf{V}$.

Define

$$\mathbf{C} = \frac{1}{H}\mathbf{X}^\top\mathbf{X} \tag{512}$$

and, for $q > 0$,

$$\mathbf{M}_{\mathrm{bal}}(q) = q\mathbf{C}\left(\mathbf{I}_p + q\mathbf{C}\right)^{-1}. \tag{513}$$

We will show that, after reparametrizing time by the growing scale q, the exact predictor converges to $\mathbf{M}_{\mathrm{bal}}(q)$ uniformly not only on compact q-intervals, but over the entire tail bounded away from zero.

Lemma B.21 (Global reduction to the balanced trajectory)**.** *Fix $p, H \in \mathbb{N}$ and fixed matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{H\times p}$, with*

$$\mathrm{rank}(\mathbf{X}) = \min\{H, p\}. \tag{514}$$

For each $\sigma > 0$, initialize

$$\mathbf{U}^{(\sigma)}(0) = \frac{\sigma}{\sqrt{2}}\left(\mathbf{X} + \mathbf{Y}\right), \qquad \mathbf{V}^{(\sigma)}(0) = \frac{\sigma}{\sqrt{2}}\left(\mathbf{X} - \mathbf{Y}\right), \tag{515}$$

and let

$$\mathbf{M}^{(\sigma)}(\tau) = \mathbf{V}^{(\sigma)}(\tau)^\top\mathbf{U}^{(\sigma)}(\tau) \tag{516}$$

evolve under the identity-target flow (443)*.*

Set

$$\varepsilon = \frac{H\sigma^2}{2}, \qquad \tau_\sigma(q) = \frac{1}{2}\log\frac{q}{\varepsilon}, \qquad q \geq \varepsilon, \tag{517}$$

and define the predictor in q-coordinates by

$$\widehat{\mathbf{M}}_\sigma(q) = \mathbf{M}^{(\sigma)}\left(\tau_\sigma(q)\right). \tag{518}$$

Then, for every fixed $\delta > 0$,

$$\limsup_{\sigma\downarrow 0}\sup_{q\geq\delta}\left\|\widehat{\mathbf{M}}_\sigma(q) - \mathbf{M}_{\mathrm{bal}}(q)\right\|_{\mathrm{F}} = 0. \tag{519}$$

Consequently,

$$\limsup_{\sigma\downarrow 0}\sup_{q\geq\delta}\left\|\widehat{\mathbf{M}}_\sigma(q) - \mathbf{M}_{\mathrm{bal}}(q)\right\|_{\max} = 0. \tag{520}$$

Here the suprema are understood for sufficiently small σ such that $\varepsilon < \delta$.

Since a standard Gaussian $H \times p$ matrix has rank $\min\{H, p\}$ almost surely, the conclusion applies almost surely to the Gaussian initialization considered above.

Proof. We divide the proof into three steps. First, we identify the exact balanced reference trajectory. Second, we prove convergence to this trajectory on every compact q-interval bounded away from zero. Finally, we establish stability near its limiting critical point, which extends the comparison uniformly over the entire infinite-time tail.

Step 1: the balanced reference trajectory.

We first note that the exact gradient flow exists for every $\tau \geq 0$. The identity-target dynamics are gradient flow for

$$\Phi(\mathbf{U}, \mathbf{V}) = \frac{1}{2}\left\|\mathbf{V}^\top\mathbf{U} - \mathbf{I}_p\right\|_{\mathrm{F}}^2. \tag{521}$$

Hence Φ is nonincreasing along the trajectory, and therefore

$$\left\|\mathbf{M}^{(\sigma)}(\tau)\right\|_{\mathrm{F}} \tag{522}$$

remains bounded for all times on the maximal existence interval.

In addition, the balance difference

$$\mathbf{D} = \mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top \tag{523}$$

is conserved. Indeed, direct differentiation using (443) gives

$$\frac{d}{d\tau}\left(\mathbf{U}\mathbf{U}^\top\right) = \frac{d}{d\tau}\left(\mathbf{V}\mathbf{V}^\top\right). \tag{524}$$

Let

$$\mathbf{G}_U = \mathbf{U}\mathbf{U}^\top, \qquad \mathbf{G}_V = \mathbf{V}\mathbf{V}^\top. \tag{525}$$

Since

$$\|\mathbf{M}\|_{\mathrm{F}}^2 = \mathrm{Tr}\left(\mathbf{G}_U\mathbf{G}_V\right), \tag{526}$$

we have

$$\begin{aligned}\|\mathbf{G}_U + \mathbf{G}_V\|_{\mathrm{F}}^2 &= \|\mathbf{G}_U - \mathbf{G}_V\|_{\mathrm{F}}^2 + 4\,\mathrm{Tr}\left(\mathbf{G}_U\mathbf{G}_V\right)\\ &= \|\mathbf{D}\|_{\mathrm{F}}^2 + 4\|\mathbf{M}\|_{\mathrm{F}}^2.\end{aligned} \tag{527}$$

The right-hand side remains bounded. Since $\mathbf{G}_U, \mathbf{G}_V \succeq 0$, this bounds the Frobenius norms of $\mathbf{U}$ and $\mathbf{V}$ on every finite time interval. The standard continuation theorem therefore rules out finite-time blow-up, so the exact solution exists for all $\tau \geq 0$.

The balanced submanifold

$$\mathbf{U} = \mathbf{V} = \mathbf{W} \tag{528}$$

is invariant. On this submanifold,

$$\dot{\mathbf{W}} = \mathbf{W}\left(\mathbf{I}_p - \mathbf{W}^\top\mathbf{W}\right). \tag{529}$$

Since

$$\frac{dq}{d\tau} = 2q, \tag{530}$$

the same equation in q-coordinates is

$$2q\frac{d\mathbf{W}}{dq} = \mathbf{W}\left(\mathbf{I}_p - \mathbf{W}^\top\mathbf{W}\right). \tag{531}$$

Define

$$\mathbf{W}_*(q) = \sqrt{\frac{q}{H}}\,\mathbf{X}\left(\mathbf{I}_p + q\mathbf{C}\right)^{-1/2}. \tag{532}$$

Since $\mathbf{C} = \mathbf{X}^\top\mathbf{X}/H$,

$$\begin{aligned}\mathbf{W}_*(q)^\top\mathbf{W}_*(q) &= \left(\mathbf{I}_p + q\mathbf{C}\right)^{-1/2} q\mathbf{C}\left(\mathbf{I}_p + q\mathbf{C}\right)^{-1/2}\\ &= q\mathbf{C}\left(\mathbf{I}_p + q\mathbf{C}\right)^{-1}\\ &= \mathbf{M}_{\mathrm{bal}}(q).\end{aligned} \tag{533}$$

Because $\mathbf{C}$ commutes with every matrix function of $\mathbf{C}$, direct differentiation gives

$$\begin{aligned}2q\frac{d\mathbf{W}_*}{dq} &= \mathbf{W}_* - \mathbf{W}_* q\mathbf{C}\left(\mathbf{I}_p + q\mathbf{C}\right)^{-1}\\ &= \mathbf{W}_*\left(\mathbf{I}_p - \mathbf{M}_{\mathrm{bal}}(q)\right)\\ &= \mathbf{W}_*\left(\mathbf{I}_p - \mathbf{W}_*^\top\mathbf{W}_*\right).\end{aligned} \tag{534}$$

Thus $(\mathbf{W}_*(q), \mathbf{W}_*(q))$ is an exact balanced trajectory of the factor dynamics.

As $q \downarrow 0$,

$$\mathbf{W}_*(q) = \sqrt{\frac{q}{H}}\,\mathbf{X} + O(q^{3/2}), \tag{535}$$

where throughout the remainder of this proof the constants in $O(\cdot)$ may depend on the fixed $p, H, \mathbf{X}, \mathbf{Y}$, but not on σ.

Step 2: convergence on compact q-intervals.

Introduce the stacked factor variable

$$\mathbf{Z} = \begin{pmatrix} \mathbf{U} \\ \mathbf{V} \end{pmatrix}. \tag{536}$$

The identity-target flow can be written as

$$\dot{\mathbf{Z}} = \mathcal{A}\mathbf{Z} + \mathcal{N}(\mathbf{Z}), \qquad \mathcal{A} = \begin{pmatrix} \mathbf{0} & \mathbf{I}_H \\ \mathbf{I}_H & \mathbf{0} \end{pmatrix}, \tag{537}$$

where

$$\mathcal{N}(\mathbf{Z}) = -\begin{pmatrix} \mathbf{V}\mathbf{M} \\ \mathbf{U}\mathbf{M}^\top \end{pmatrix}, \qquad \mathbf{M} = \mathbf{V}^\top \mathbf{U}. \tag{538}$$

The nonlinear term is cubic. Hence, for the fixed dimensions p, H, there exists $K_0 < \infty$ such that

$$\|\mathcal{N}(\mathbf{Z})\|_{\mathrm{F}} \le K_0 \|\mathbf{Z}\|_{\mathrm{F}}^3. \tag{539}$$

Moreover,

$$\left\|e^{\tau\mathcal{A}}\right\|_{\mathrm{op}} = e^{\tau}, \qquad \tau \ge 0. \tag{540}$$

Set

$$z(\tau) = \|\mathbf{Z}(\tau)\|_{\mathrm{F}}. \tag{541}$$

Variation of constants gives

$$z(\tau) \le e^{\tau} z(0) + K_0 \int_0^\tau e^{\tau - u} z(u)^3 \, du. \tag{542}$$

From (515), there exists a finite constant c_0, depending only on the fixed $p, H, \mathbf{X}, \mathbf{Y}$, such that

$$z(0) = c_0 \sqrt{\varepsilon}. \tag{543}$$

Define

$$a(\tau) = e^{-\tau} z(\tau). \tag{544}$$

Then

$$a(\tau) \le z(0) + K_0 \int_0^\tau e^{2u} a(u)^3 \, du. \tag{545}$$

For the fixed $p, H, \mathbf{X}, \mathbf{Y}$, choose $q_{\mathrm{loc}} > 0$ sufficiently small. If

$$a(u) \le 2z(0) \qquad \text{for } 0 \le u \le \tau, \tag{546}$$

and $q(\tau) \le q_{\mathrm{loc}}$, then

$$\begin{aligned} K_0 \int_0^\tau e^{2u} a(u)^3 \, du &\le 4K_0 z(0)^3 e^{2\tau} \\ &= 4K_0 c_0^2 q(\tau) z(0). \end{aligned} \tag{547}$$

Choosing q_{loc} so that

$$4K_0 c_0^2 q_{\mathrm{loc}} < \frac{1}{2} \tag{548}$$

closes the usual first-exit argument and yields

$$z(\tau) \le 2c_0 \sqrt{q(\tau)} \qquad \text{whenever } q(\tau) \le q_{\mathrm{loc}}. \tag{549}$$

Using (549) once more in variation of constants gives

$$\left\|\mathbf{Z}(\tau) - e^{\tau\mathcal{A}}\mathbf{Z}(0)\right\|_{\mathrm{F}} \le K_1 q(\tau)^{3/2} \tag{550}$$

for some finite K_1 depending only on the fixed $p, H, \mathbf{X}, \mathbf{Y}$.

The corresponding linearized factors, expressed in terms of q, are

$$\mathbf{U}_{\mathrm{lin}}(q) = \sqrt{\frac{q}{H}}\,\mathbf{X} + \frac{\varepsilon}{\sqrt{qH}}\,\mathbf{Y}, \tag{551}$$

and

$$\mathbf{V}_{\mathrm{lin}}(q) = \sqrt{\frac{q}{H}}\,\mathbf{X} - \frac{\varepsilon}{\sqrt{qH}}\,\mathbf{Y}. \tag{552}$$

Choose the intermediate scale

$$q_0 = \sqrt{\varepsilon}. \tag{553}$$

At $q = q_0$, the decaying terms in (551) and (552) have size $O(\varepsilon^{3/4})$, and (550) has the same order. Together with (535), this gives

$$E_\sigma(q_0) = O(\varepsilon^{3/4}), \tag{554}$$

where

$$E_\sigma(q) = \left\| \begin{pmatrix} \mathbf{U}^{(\sigma)}(\tau_\sigma(q)) - \mathbf{W}_*(q) \\ \mathbf{V}^{(\sigma)}(\tau_\sigma(q)) - \mathbf{W}_*(q) \end{pmatrix} \right\|_{\mathrm{F}}. \tag{555}$$

We next propagate this comparison from the vanishing scale q_0 to a fixed positive value of q. By decreasing q_{loc}, if necessary, the exact and balanced trajectories for $q \le q_{\mathrm{loc}}$ remain in a fixed neighborhood of the origin on which the vector field is 2-Lipschitz. Since both trajectories solve the same autonomous ODE, Gronwall's inequality gives

$$\begin{aligned} E_\sigma(q) &\le E_\sigma(q_0) \exp\left(2[\tau_\sigma(q) - \tau_\sigma(q_0)]\right) \\ &= E_\sigma(q_0)\frac{q}{q_0}, \qquad q_0 \le q \le q_{\mathrm{loc}}. \end{aligned} \tag{556}$$

Consequently,

$$E_\sigma(q_{\mathrm{loc}}) = O(\varepsilon^{1/4}) \longrightarrow 0. \tag{557}$$

Now fix any finite $Q > q_{\mathrm{loc}}$. The balanced trajectory on $[q_{\mathrm{loc}}, Q]$ is compact. Choose a compact neighborhood of this trajectory on which the factor vector field has Lipschitz constant $L_Q < \infty$. Up to the first exit of the exact trajectory from this neighborhood, Gronwall's inequality gives

$$E_\sigma(q) \le E_\sigma(q_{\mathrm{loc}}) \left(\frac{q}{q_{\mathrm{loc}}}\right)^{L_Q/2}, \qquad q_{\mathrm{loc}} \le q \le Q. \tag{558}$$

The right-hand side tends uniformly to zero as $\sigma \downarrow 0$, so for sufficiently small σ such a first exit cannot occur. Hence

$$\sup_{q_{\mathrm{loc}} \le q \le Q} E_\sigma(q) \longrightarrow 0. \tag{559}$$

If $0 < \delta < q_{\mathrm{loc}}$, then once $q_0 < \delta$, (556) also gives

$$\sup_{\delta \le q \le q_{\mathrm{loc}}} E_\sigma(q) = O(\varepsilon^{1/4}). \tag{560}$$

Combining this with (559), we conclude that for every fixed $0 < \delta < Q < \infty$,

$$\sup_{\delta \le q \le Q} E_\sigma(q) \longrightarrow 0. \tag{561}$$

Since the factors remain bounded on the fixed interval $[\delta, Q]$, it follows that

$$\sup_{\delta \le q \le Q} \left\| \widehat{\mathbf{M}}_\sigma(q) - \mathbf{M}_{\mathrm{bal}}(q) \right\|_{\mathrm{F}} \longrightarrow 0. \tag{562}$$

Step 3: uniform control of the infinite-time tail.

Let

$$r = \min\{H, p\}, \tag{563}$$

and take a compact singular-value decomposition

$$\mathbf{X} = \mathbf{L}\mathbf{\Sigma}\mathbf{R}^\top, \tag{564}$$

where $\mathbf{L} \in \mathbb{R}^{H\times r}$ and $\mathbf{R} \in \mathbb{R}^{p\times r}$ have orthonormal columns and $\mathbf{\Sigma} \in \mathbb{R}^{r\times r}$ is positive diagonal. From (532),

$$\mathbf{W}_*(q) = \mathbf{L}\,\mathrm{diag}\left(\frac{\sqrt{q/H}\,\Sigma_{kk}}{\sqrt{1+(q/H)\Sigma_{kk}^2}}\right)_{k=1}^{r}\mathbf{R}^\top. \tag{565}$$

Hence

$$\mathbf{W}_*(q) \longrightarrow \mathbf{W}_\infty := \mathbf{L}\mathbf{R}^\top \qquad \text{as } q\to\infty, \tag{566}$$

and

$$\mathbf{M}_{\mathrm{bal}}(q) \longrightarrow \mathbf{P}_X := \mathbf{W}_\infty^\top\mathbf{W}_\infty = \mathbf{R}\mathbf{R}^\top. \tag{567}$$

If $H \geq p$, then

$$\mathbf{P}_X = \mathbf{I}_p, \tag{568}$$

whereas if $H < p$, then $\mathbf{P}_X$ is the orthogonal projector onto the row space of $\mathbf{X}$.

We next establish the local stability property needed to control the exact trajectory after it has entered a neighborhood of $(\mathbf{W}_\infty, \mathbf{W}_\infty)$. Specifically, we claim that for every $\eta > 0$ there exists $\rho > 0$ such that

$$\|\mathbf{U}(\tau_0) - \mathbf{W}_\infty\|_{\mathrm{F}} + \|\mathbf{V}(\tau_0) - \mathbf{W}_\infty\|_{\mathrm{F}} < \rho \tag{569}$$

implies

$$\left\|\mathbf{V}(\tau)^\top\mathbf{U}(\tau) - \mathbf{P}_X\right\|_{\mathrm{F}} < \eta \qquad \text{for every } \tau \geq \tau_0. \tag{570}$$

All constants and neighborhoods in this argument may depend on the fixed $p, H, \mathbf{X}, \mathbf{Y}$. This causes no difficulty because the limit $\sigma \downarrow 0$ is taken before the large-system limit.

We verify (570) from the local geometry of the potential Φ in (521).

Suppose first that $H < p$. By orthogonal changes of hidden and input coordinates, which preserve both the Euclidean metric and Φ, we may assume

$$\mathbf{W}_\infty = \begin{pmatrix}\mathbf{I}_H & \mathbf{0}\end{pmatrix}. \tag{571}$$

The global minimizers in a neighborhood of this point form the smooth manifold

$$\mathcal{M} = \left\{\left((\mathbf{V}\mathbf{V}^\top)^{-1}\mathbf{V}, \mathbf{V}\right) : \mathrm{rank}(\mathbf{V}) = H\right\}. \tag{572}$$

Indeed, every end-to-end matrix $\mathbf{V}^\top\mathbf{U}$ has rank at most H, so the optimal rank-H approximation error to $\mathbf{I}_p$ is

$$p - H. \tag{573}$$

For a fixed full-row-rank $\mathbf{V}$, define

$$\mathbf{P}_V = \mathbf{V}^\top(\mathbf{V}\mathbf{V}^\top)^{-1}\mathbf{V}. \tag{574}$$

Since the columns of $\mathbf{V}^\top\mathbf{U}$ lie in the range of $\mathbf{P}_V$, orthogonality gives

$$\left\|\mathbf{V}^\top\mathbf{U} - \mathbf{I}_p\right\|_{\mathrm{F}}^2 = \left\|\mathbf{V}^\top\mathbf{U} - \mathbf{P}_V\right\|_{\mathrm{F}}^2 + p - H. \tag{575}$$

Equality with the minimum value therefore holds precisely when

$$\mathbf{V}^\top\mathbf{U} = \mathbf{P}_V, \tag{576}$$

equivalently,

$$\mathbf{U} = (\mathbf{V}\mathbf{V}^\top)^{-1}\mathbf{V}. \tag{577}$$

Thus every point of $\mathcal{M}$ is a global minimum and hence a critical point. Since the full-row-rank matrices form an open set and (577) defines a smooth graph, $\mathcal{M}$ is an embedded smooth manifold.

Write

$$\mathbf{U} = \mathbf{W}_\infty + \mathbf{E}, \qquad \mathbf{V} = \mathbf{W}_\infty + \mathbf{F}, \tag{578}$$

with block decompositions

$$\mathbf{E} = \begin{pmatrix}\mathbf{E}_1 & \mathbf{E}_2\end{pmatrix}, \qquad \mathbf{F} = \begin{pmatrix}\mathbf{F}_1 & \mathbf{F}_2\end{pmatrix}, \tag{579}$$

where $\mathbf{E}_1, \mathbf{F}_1 \in \mathbb{R}^{H\times H}$. A second-order expansion gives

$$\begin{aligned}&\Phi(\mathbf{W}_\infty + \mathbf{E}, \mathbf{W}_\infty + \mathbf{F}) - \Phi(\mathbf{W}_\infty, \mathbf{W}_\infty)\\ &\quad = \frac{1}{2}\left\|\mathbf{E}_1 + \mathbf{F}_1^\top\right\|_{\mathrm{F}}^2 + \frac{1}{2}\left\|\mathbf{E}_2 - \mathbf{F}_2\right\|_{\mathrm{F}}^2 + O\left((\|\mathbf{E}\|_{\mathrm{F}} + \|\mathbf{F}\|_{\mathrm{F}})^3\right).\end{aligned} \tag{580}$$

Hence the Hessian is positive semidefinite, with kernel characterized by

$$\mathbf{E}_1 = -\mathbf{F}_1^\top, \qquad \mathbf{E}_2 = \mathbf{F}_2. \tag{581}$$

Differentiating (577) at $\mathbf{W}_\infty$ gives

$$D\mathbf{U}(\mathbf{W}_\infty)[\mathbf{F}] = \mathbf{F} - \left(\mathbf{F}\mathbf{W}_\infty^\top + \mathbf{W}_\infty\mathbf{F}^\top\right)\mathbf{W}_\infty = \begin{pmatrix}-\mathbf{F}_1^\top & \mathbf{F}_2\end{pmatrix}. \tag{582}$$

Therefore

$$\ker \nabla^2\Phi(\mathbf{W}_\infty, \mathbf{W}_\infty) = T_{(\mathbf{W}_\infty, \mathbf{W}_\infty)}\mathcal{M}. \tag{583}$$

Thus the Hessian is positive definite on the Euclidean normal space to $\mathcal{M}$.

Suppose next that $H \geq p$. We may choose coordinates such that

$$\mathbf{W}_\infty = \begin{pmatrix}\mathbf{I}_p \\ \mathbf{0}\end{pmatrix}. \tag{584}$$

The global-minimum manifold is now

$$\mathcal{M} = \left\{(\mathbf{U}, \mathbf{V}) : \mathbf{V}^\top\mathbf{U} = \mathbf{I}_p\right\}. \tag{585}$$

This is a smooth manifold. Indeed, the derivative of the constraint map is

$$(\mathbf{E}, \mathbf{F}) \longmapsto \mathbf{V}^\top\mathbf{E} + \mathbf{F}^\top\mathbf{U}, \tag{586}$$

and is surjective at every point of $\mathcal{M}$: for an arbitrary $\mathbf{J} \in \mathbb{R}^{p\times p}$, one may take

$$\mathbf{E} = \mathbf{V}(\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{J}, \qquad \mathbf{F} = \mathbf{0}. \tag{587}$$

The implicit function theorem therefore applies.

Writing

$$\mathbf{E} = \begin{pmatrix}\mathbf{E}_1 \\ \mathbf{E}_2\end{pmatrix}, \qquad \mathbf{F} = \begin{pmatrix}\mathbf{F}_1 \\ \mathbf{F}_2\end{pmatrix}, \tag{588}$$

with $\mathbf{E}_1, \mathbf{F}_1 \in \mathbb{R}^{p\times p}$, we obtain

$$\Phi(\mathbf{W}_\infty + \mathbf{E}, \mathbf{W}_\infty + \mathbf{F}) = \frac{1}{2}\left\|\mathbf{E}_1 + \mathbf{F}_1^\top\right\|_{\mathrm{F}}^2 + O\left((\|\mathbf{E}\|_{\mathrm{F}} + \|\mathbf{F}\|_{\mathrm{F}})^3\right). \tag{589}$$

Its quadratic kernel is

$$\mathbf{E}_1 + \mathbf{F}_1^\top = \mathbf{0}, \tag{590}$$

with $\mathbf{E}_2, \mathbf{F}_2$ arbitrary, exactly the tangent space to $\mathcal{M}$. Thus in this case as well,

$$\ker \nabla^2\Phi(\mathbf{W}_\infty, \mathbf{W}_\infty) = T_{(\mathbf{W}_\infty, \mathbf{W}_\infty)}\mathcal{M}, \tag{591}$$

and the Hessian is strictly positive in every normal direction.

We now use this common local structure. Regard

$$\mathbf{Z} = (\mathbf{U}, \mathbf{V}) \tag{592}$$

as a point in Euclidean factor space with norm

$$\|\mathbf{Z}\|^2 = \|\mathbf{U}\|_{\mathrm{F}}^2 + \|\mathbf{V}\|_{\mathrm{F}}^2, \qquad \mathbf{Z}_\infty = (\mathbf{W}_\infty, \mathbf{W}_\infty). \tag{593}$$

On a sufficiently small patch of $\mathcal{M}$ around $\mathbf{Z}_\infty$, define

$$\mathcal{H}_\Phi(\boldsymbol{\theta}) = \nabla^2\Phi(\boldsymbol{\theta}). \tag{594}$$

Since $\nabla\Phi$ vanishes identically on $\mathcal{M}$, $\mathcal{H}_\Phi(\boldsymbol{\theta})$ annihilates $T_{\boldsymbol{\theta}}\mathcal{M}$. By symmetry, it preserves the normal space $N_{\boldsymbol{\theta}}\mathcal{M}$. The positivity established above, together with continuity of the Hessian and of the tangent and normal spaces, implies that after shrinking the patch there exists $\mu > 0$ such that

$$\langle \mathbf{v}, \mathcal{H}_\Phi(\boldsymbol{\theta})\mathbf{v}\rangle \geq \mu\|\mathbf{v}\|^2 \qquad \text{for every } \mathbf{v} \in N_{\boldsymbol{\theta}}\mathcal{M}. \tag{595}$$

After decreasing the normal radius if necessary, continuity further gives

$$\langle \mathbf{v}, \nabla^2 \Phi(\boldsymbol{\theta} + \boldsymbol{\zeta})\mathbf{v} \rangle \geq \frac{\mu}{2}\|\mathbf{v}\|^2 \tag{596}$$

whenever

$$\mathbf{v}, \boldsymbol{\zeta} \in N_{\boldsymbol{\theta}}\mathcal{M} \tag{597}$$

and $\|\boldsymbol{\zeta}\|$ is sufficiently small.

The normal map

$$(\boldsymbol{\theta}, \boldsymbol{\zeta}) \longmapsto \boldsymbol{\theta} + \boldsymbol{\zeta}, \qquad \boldsymbol{\zeta} \in N_{\boldsymbol{\theta}}\mathcal{M}, \tag{598}$$

provides smooth tubular coordinates near $\mathbf{Z}_\infty$. After restricting to a smaller patch with compact closure, there exists a uniform tubular radius and a smooth nearest-manifold projection

$$\boldsymbol{\pi}(\mathbf{Z}) = \boldsymbol{\theta}. \tag{599}$$

At a point $\boldsymbol{\theta} \in \mathcal{M}$,

$$D\boldsymbol{\pi}(\boldsymbol{\theta}) = \mathbf{P}_{T_{\boldsymbol{\theta}}\mathcal{M}}. \tag{600}$$

Uniform Taylor estimates therefore give a constant $K_{\mathrm{loc}} < \infty$ such that

$$\|D\boldsymbol{\pi}(\boldsymbol{\theta} + \boldsymbol{\zeta}) - \mathbf{P}_{T_{\boldsymbol{\theta}}\mathcal{M}}\|_{\mathrm{op}} \leq K_{\mathrm{loc}}\|\boldsymbol{\zeta}\| \tag{601}$$

and

$$\|\nabla\Phi(\boldsymbol{\theta} + \boldsymbol{\zeta}) - \mathcal{H}_\Phi(\boldsymbol{\theta})\boldsymbol{\zeta}\| \leq K_{\mathrm{loc}}\|\boldsymbol{\zeta}\|^2. \tag{602}$$

Along gradient flow, write

$$\boldsymbol{\theta}(\tau) = \boldsymbol{\pi}(\mathbf{Z}(\tau)), \qquad \boldsymbol{\zeta}(\tau) = \mathbf{Z}(\tau) - \boldsymbol{\theta}(\tau). \tag{603}$$

Because $\mathcal{H}_\Phi(\boldsymbol{\theta})\boldsymbol{\zeta}$ is normal to $\mathcal{M}$, the chain rule and the preceding Taylor bounds imply

$$\dot{\boldsymbol{\zeta}} = -\mathcal{H}_\Phi(\boldsymbol{\theta})\boldsymbol{\zeta} + \mathbf{R}_\zeta, \qquad \|\mathbf{R}_\zeta\| \leq K_{\mathrm{loc}}\|\boldsymbol{\zeta}\|^2, \tag{604}$$

and

$$\|\dot{\boldsymbol{\theta}}\| \leq K_{\mathrm{loc}}\|\boldsymbol{\zeta}\|^2. \tag{605}$$

Choose a normal radius $b > 0$ inside the tubular neighborhood such that

$$2K_{\mathrm{loc}}b \leq \mu. \tag{606}$$

Then, as long as $\|\boldsymbol{\zeta}\| < b$,

$$\begin{aligned} \frac{d}{d\tau}\|\boldsymbol{\zeta}\|^2 &\leq -2\mu\|\boldsymbol{\zeta}\|^2 + 2K_{\mathrm{loc}}\|\boldsymbol{\zeta}\|^3 \\ &\leq -\mu\|\boldsymbol{\zeta}\|^2. \end{aligned} \tag{607}$$

Therefore, up to any possible exit time,

$$\|\boldsymbol{\zeta}(\tau)\|^2 \leq e^{-\mu(\tau-\tau_0)}\|\boldsymbol{\zeta}(\tau_0)\|^2 \tag{608}$$

and

$$\int_{\tau_0}^{\tau} \|\dot{\boldsymbol{\theta}}(u)\|\, du \leq \frac{K_{\mathrm{loc}}}{\mu}\|\boldsymbol{\zeta}(\tau_0)\|^2 \left(1 - e^{-\mu(\tau-\tau_0)}\right). \tag{609}$$

It remains to rule out exit from the tubular neighborhood. Choose $a > 0$ such that the closure of

$$\{\boldsymbol{\theta} \in \mathcal{M} : \|\boldsymbol{\theta} - \mathbf{Z}_\infty\| < a\} \tag{610}$$

lies inside the chosen coordinate patch, and define

$$\beta = 1 + \frac{K_{\mathrm{loc}}}{\mu}. \tag{611}$$

After decreasing b if necessary, consider the tapered tubular neighborhood

$$\mathcal{T} = \left\{ \mathbf{Z} = \boldsymbol{\theta} + \boldsymbol{\zeta} : \begin{array}{ll} \boldsymbol{\theta} \in \mathcal{M}, & \boldsymbol{\zeta} \in N_{\boldsymbol{\theta}}\mathcal{M}, \\ \|\boldsymbol{\zeta}\| < b, & \|\boldsymbol{\theta} - \mathbf{Z}_\infty\| + \beta\|\boldsymbol{\zeta}\|^2 < a \end{array} \right\}. \tag{612}$$

Its closure may be chosen to lie inside the tubular coordinate neighborhood. Moreover, $\mathcal{T}$ is forward invariant. Indeed, (607) shows that $\|\boldsymbol{\zeta}\|^2$ is nonincreasing, while (605) and (607) give

$$\frac{d^+}{d\tau}\left(\|\boldsymbol{\theta}-\mathbf{Z}_\infty\|+\beta\|\boldsymbol{\zeta}\|^2\right)\le\left(K_{\mathrm{loc}}-\beta\mu\right)\|\boldsymbol{\zeta}\|^2\le 0. \tag{613}$$

Here $d^+/d\tau$ denotes the upper-right Dini derivative, used only to avoid differentiability issues at $\boldsymbol{\theta}=\mathbf{Z}_\infty$. A first-exit argument, together with global existence established above, therefore shows that every trajectory starting in $\mathcal{T}$ remains there for all later times.

In particular, the normal component decays exponentially and the tangential drift is integrable. Thus $\boldsymbol{\theta}(\tau)$ converges to some $\boldsymbol{\theta}_\infty\in\mathcal{M}$, and

$$\|\boldsymbol{\theta}(\tau)-\boldsymbol{\theta}_\infty\|\le\frac{K_{\mathrm{loc}}}{\mu}\|\boldsymbol{\zeta}(\tau_0)\|^2e^{-\mu(\tau-\tau_0)}. \tag{614}$$

Finally, by continuity of

$$(\mathbf{U},\mathbf{V})\longmapsto\mathbf{V}^\top\mathbf{U} \tag{615}$$

at $\mathbf{Z}_\infty$, we may choose the tubular neighborhood sufficiently small that

$$\left\|\mathbf{V}^\top\mathbf{U}-\mathbf{P}_X\right\|_{\mathrm{F}}<\eta \tag{616}$$

for every $(\mathbf{U},\mathbf{V})\in\mathcal{T}$. Since $\mathcal{T}$ is open and contains $\mathbf{Z}_\infty$, it contains an ambient ball of some radius $\rho>0$. Therefore the entry condition (569) places the factors inside $\mathcal{T}$, and forward invariance proves (570).

We now finish the global comparison. Fix $\eta>0$, and let $\rho>0$ be the stability radius corresponding to the output tolerance $\eta/2$. Since

$$\mathbf{W}_*(q)\longrightarrow\mathbf{W}_\infty,\qquad\mathbf{M}_{\mathrm{bal}}(q)\longrightarrow\mathbf{P}_X, \tag{617}$$

we may choose a finite $Q>\max\{\delta,q_{\mathrm{loc}}\}$ such that

$$2\left\|\mathbf{W}_*(Q)-\mathbf{W}_\infty\right\|_{\mathrm{F}}<\frac{\rho}{2} \tag{618}$$

and

$$\sup_{q\ge Q}\left\|\mathbf{M}_{\mathrm{bal}}(q)-\mathbf{P}_X\right\|_{\mathrm{F}}<\frac{\eta}{2}. \tag{619}$$

By (561),

$$E_\sigma(Q)\longrightarrow 0\qquad\text{as }\sigma\downarrow 0. \tag{620}$$

Hence, for all sufficiently small σ,

$$E_\sigma(Q)<\frac{\rho}{2\sqrt{2}}. \tag{621}$$

Using (555), (618), and (621), we obtain

$$\begin{aligned}&\left\|\mathbf{U}^{(\sigma)}(\tau_\sigma(Q))-\mathbf{W}_\infty\right\|_{\mathrm{F}}+\left\|\mathbf{V}^{(\sigma)}(\tau_\sigma(Q))-\mathbf{W}_\infty\right\|_{\mathrm{F}}\\&\le\sqrt{2}E_\sigma(Q)+2\left\|\mathbf{W}_*(Q)-\mathbf{W}_\infty\right\|_{\mathrm{F}}<\rho.\end{aligned} \tag{622}$$

The stability property (570) therefore yields

$$\sup_{q\ge Q}\left\|\widehat{\mathbf{M}}_\sigma(q)-\mathbf{P}_X\right\|_{\mathrm{F}}<\frac{\eta}{2}. \tag{623}$$

Combining this with (619), we obtain

$$\sup_{q\ge Q}\left\|\widehat{\mathbf{M}}_\sigma(q)-\mathbf{M}_{\mathrm{bal}}(q)\right\|_{\mathrm{F}}<\eta. \tag{624}$$

On the remaining compact interval $[\delta,Q]$, (562) gives

$$\sup_{\delta\le q\le Q}\left\|\widehat{\mathbf{M}}_\sigma(q)-\mathbf{M}_{\mathrm{bal}}(q)\right\|_{\mathrm{F}}\longrightarrow 0. \tag{625}$$

Since $\eta>0$ was arbitrary, combining (624) and (625) proves (519).

Finally,

$$\|\mathbf{A}\|_{\max}\le\|\mathbf{A}\|_{\mathrm{F}} \tag{626}$$

for every matrix $\mathbf{A}$, so (520) follows immediately. □

The global reduction above turns the late-time classification problem into a property of the explicit balanced predictor $\mathbf{M}_{\mathrm{bal}}(q)$. It remains to show that this reference trajectory itself classifies every training example correctly, with a margin bounded uniformly away from zero for all q bounded away from zero. Although the eigenvalues of the Gaussian Gram matrix $\mathbf{C}$ need not be close to one, its eigenvectors are delocalized. Consequently, the diagonal entries of $\mathbf{M}_{\mathrm{bal}}(q)$ are close to their common average, while all off-diagonal entries are uniformly small. The following lemma makes this uniform in the entire range $q \geq \delta$.

Lemma B.22 (Uniform accuracy margin of the balanced trajectory)**.** *Suppose that*

$$\frac{H}{p} \longrightarrow \alpha \in (0, \infty), \tag{627}$$

and let

$$\mathbf{C} = \frac{1}{H}\mathbf{X}^\top \mathbf{X}, \qquad \mathbf{M}_{\mathrm{bal}}(q) = q\mathbf{C}\left(\mathbf{I}_p + q\mathbf{C}\right)^{-1}, \tag{628}$$

where $\mathbf{X} \in \mathbb{R}^{H\times p}$ has independent standard Gaussian entries. For every fixed $\delta > 0$, there exists a constant $c_{\alpha,\delta} > 0$ such that, with probability tending to one,

$$\inf_{q\geq\delta} \min_{i\in[p]} \min_{c\neq i} \left[(\mathbf{M}_{\mathrm{bal}}(q))_{ii} - (\mathbf{M}_{\mathrm{bal}}(q))_{ci}\right] \geq c_{\alpha,\delta}. \tag{629}$$

In particular, with probability tending to one, $\mathbf{M}_{\mathrm{bal}}(q)$ has perfect training accuracy simultaneously for every $q \geq \delta$.

Proof. We separately control the average diagonal size and the fluctuations of individual entries around it.

By orthogonal invariance of the Gaussian Wishart law, we may represent $\mathbf{C}$ as

$$\mathbf{C} = \mathbf{O}\mathbf{\Lambda}\mathbf{O}^\top, \qquad \mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_p), \qquad 0 \leq \lambda_1 \leq \cdots \leq \lambda_p, \tag{630}$$

where, conditional on the ordered eigenvalues, $\mathbf{O}$ may be taken Haar distributed on the orthogonal group and independent of $\mathbf{\Lambda}$. When $H < p$, the zero eigenvalue has multiplicity $p - H$; choosing an independent Haar basis inside this nullspace gives the same representation.

For $q > 0$, define

$$d_j(q) = \frac{q\lambda_j}{1 + q\lambda_j}, \qquad \overline{d}(q) = \frac{1}{p}\sum_{j=1}^{p} d_j(q). \tag{631}$$

Then

$$\mathbf{M}_{\mathrm{bal}}(q) = \mathbf{O}\,\mathrm{diag}\big(d_1(q), \ldots, d_p(q)\big)\mathbf{O}^\top. \tag{632}$$

For every $q > 0$, the sequence $j \mapsto d_j(q)$ is nondecreasing and takes values in $[0, 1]$.

We first lower-bound $\overline{d}(q)$. On the high-probability event from Lemma B.19, there exists $L_\alpha < \infty$, depending only on α, such that for all sufficiently large p, H,

$$\|\mathbf{C}\|_{\mathrm{op}} \leq L_\alpha, \qquad \frac{1}{p}\,\mathrm{Tr}\,\mathbf{C} \geq \frac{1}{2}. \tag{633}$$

Indeed, one may take $L_\alpha = C_\alpha^2$ in the notation of that lemma. Since $q \mapsto d_j(q)$ is nondecreasing, for every $q \geq \delta$,

$$\begin{aligned} \overline{d}(q) &\geq \overline{d}(\delta) \\ &= \frac{1}{p}\sum_{j=1}^{p} \frac{\delta\lambda_j}{1 + \delta\lambda_j} \\ &\geq \frac{\delta}{1 + \delta L_\alpha}\frac{1}{p}\sum_{j=1}^{p} \lambda_j \\ &\geq \frac{\delta}{2(1 + \delta L_\alpha)} =: m_{\alpha,\delta} > 0. \end{aligned} \tag{634}$$

It remains to show that the individual entries are uniformly close to this average. We claim that

$$\sup_{q>0}\left[\max_{i\in[p]}\left|(\mathbf{M}_{\mathrm{bal}}(q))_{ii}-\overline{d}(q)\right|+\max_{i\neq c}|(\mathbf{M}_{\mathrm{bal}}(q))_{ci}|\right]=o_{\mathbb{P}}(1). \tag{635}$$

The uniformity over the continuous and unbounded range of q follows from the monotonicity of the weights $d_j(q)$.

We first record a concentration estimate for Haar rows. If $\mathbf{z}$ is uniform on the unit sphere $\mathbb{S}^{p-1}$, then for a universal constant $c>0$, uniformly over $1\le k\le p$,

$$\mathbb{P}\left(\left|\sum_{j=1}^{k}z_j^2-\frac{k}{p}\right|>t\right)\le 4e^{-cpt^2},\qquad 0<t<1. \tag{636}$$

To see this, write $\mathbf{z}=\mathbf{g}/\|\mathbf{g}\|_2$ with $\mathbf{g}\sim\mathcal{N}(\mathbf{0},\mathbf{I}_p)$. The standard chi-square bound

$$\mathbb{P}\left(|\chi_k^2-k|>u\right)\le 2\exp\left[-c\min\left\{\frac{u^2}{k},u\right\}\right] \tag{637}$$

applied to the partial sum $\sum_{j=1}^{k}g_j^2$ and to $\|\mathbf{g}\|_2^2$ gives (636).

Let O_{ij} denote the entries of $\mathbf{O}$. Applying (636) to every row of $\mathbf{O}$, followed by a union bound over i and k, gives

$$\max_{i,k}\left|\sum_{j=1}^{k}\left(O_{ij}^2-\frac{1}{p}\right)\right|\le\omega_p \tag{638}$$

with probability tending to one, where

$$\omega_p=K\sqrt{\frac{\log p}{p}}\longrightarrow 0 \tag{639}$$

for a sufficiently large universal constant K.

We also need an analogous estimate for two distinct rows. Fix $i\neq c$. Since the i-th and c-th rows of $\mathbf{O}$ are orthonormal,

$$\mathbf{z}^{(i,c)}=\frac{\mathbf{O}_{i\cdot}+\mathbf{O}_{c\cdot}}{\sqrt{2}} \tag{640}$$

is a unit vector which, by Haar invariance, is itself uniformly distributed on $\mathbb{S}^{p-1}$. For every k,

$$\sum_{j=1}^{k}O_{ij}O_{cj}=\sum_{j=1}^{k}\left(z_j^{(i,c)}\right)^2-\frac{1}{2}\sum_{j=1}^{k}O_{ij}^2-\frac{1}{2}\sum_{j=1}^{k}O_{cj}^2. \tag{641}$$

Applying (636) to these vectors and taking a union bound over i,c,k, after increasing K if necessary, gives

$$\max_{i\neq c}\max_{k}\left|\sum_{j=1}^{k}O_{ij}O_{cj}\right|\le\omega_p \tag{642}$$

with probability tending to one.

We now convert these partial-sum estimates into bounds uniform in q. Fix i and define

$$a_j=O_{ij}^2-\frac{1}{p},\qquad A_k=\sum_{j=1}^{k}a_j. \tag{643}$$

Since the i-th row of $\mathbf{O}$ has unit norm, $A_p=0$. Summation by parts gives

$$\begin{aligned}(\mathbf{M}_{\mathrm{bal}}(q))_{ii}-\overline{d}(q)&=\sum_{j=1}^{p}d_j(q)a_j\\&=-\sum_{k=1}^{p-1}A_k\left[d_{k+1}(q)-d_k(q)\right].\end{aligned} \tag{644}$$

Because the weights are nondecreasing and belong to $[0,1]$,

$$\sum_{k=1}^{p-1} [d_{k+1}(q) - d_k(q)] \le 1. \tag{645}$$

Therefore (638) implies

$$\sup_{q>0} \max_i \left|(\mathbf{M}_{\mathrm{bal}}(q))_{ii} - \overline{d}(q)\right| \le \omega_p. \tag{646}$$

Similarly, for $i \neq c$, define

$$b_j = O_{ij}O_{cj}, \qquad B_k = \sum_{j=1}^{k} b_j. \tag{647}$$

Orthogonality of the two rows gives $B_p = 0$, and another application of summation by parts yields

$$(\mathbf{M}_{\mathrm{bal}}(q))_{ci} = -\sum_{k=1}^{p-1} B_k \left[d_{k+1}(q) - d_k(q)\right]. \tag{648}$$

Hence (642) gives

$$\sup_{q>0} \max_{i\neq c} |(\mathbf{M}_{\mathrm{bal}}(q))_{ci}| \le \omega_p. \tag{649}$$

The events (638) and (642) hold simultaneously with probability tending to one. On this single event, the summation-by-parts estimates are deterministic for every ordered eigenvalue sequence and every $q > 0$. This proves (635) without any discretization or union bound over q.

Finally, combining (634), (646), and (649), for every $q \ge \delta$, $i \in [p]$, and $c \neq i$,

$$\begin{aligned}(\mathbf{M}_{\mathrm{bal}}(q))_{ii} - (\mathbf{M}_{\mathrm{bal}}(q))_{ci} &\ge \overline{d}(q) - 2\omega_p \\ &\ge m_{\alpha,\delta} - 2\omega_p.\end{aligned} \tag{650}$$

Since $\omega_p \to 0$, for all sufficiently large p the right-hand side is at least $m_{\alpha,\delta}/2$. Thus one may take

$$c_{\alpha,\delta} = \frac{\delta}{4(1+\delta L_\alpha)} > 0, \tag{651}$$

which proves (629). □

We can now combine the small-q margin monotonicity with the global balanced-trajectory reduction. The two controls overlap at any fixed positive value of q, which allows us to cover the entire training trajectory without leaving an intermediate time interval uncontrolled.

Theorem B.23 [4.4] **(Permutation target: monotone accuracy and permanent perfect classification).** *Let $f : [p] \to [p]$ be any permutation, and suppose that*

$$\frac{H}{p} \longrightarrow \alpha \in (0,\infty) \tag{652}$$

as $p, H \to \infty$. Let

$$\tau_{\mathrm{org}} = K\sqrt{\frac{\log(2p^2H)}{H}} \tag{653}$$

with K a sufficiently large universal constant.

Then, in the sequential limit $\sigma \downarrow 0$ followed by $p, H \to \infty$ with $H/p \to \alpha$,

$$\mathbb{P}\left(\mathcal{A}\big(\mathbf{M}^{(\sigma)}(\tau_1)\big) \le \mathcal{A}\big(\mathbf{M}^{(\sigma)}(\tau_2)\big) \quad \textit{for every } 0 \le \tau_1 \le \tau_2\right) \longrightarrow 1. \tag{654}$$

Moreover,

$$\mathbb{P}\left(\mathcal{A}\big(\mathbf{M}^{(\sigma)}(\tau)\big) = 1 \quad \textit{for every } \tau \ge \tau_{\mathrm{org}}\right) \longrightarrow 1. \tag{655}$$

At the organization time, the perfect classifier is still microscopic: for every fixed p, H,

$$\left\|\mathbf{M}^{(\sigma)}(\tau_{\mathrm{org}})\right\|_{\mathrm{op}} \longrightarrow 0 \qquad \text{as } \sigma \downarrow 0, \tag{656}$$

and

$$\mathcal{L}\left(\mathbf{U}^{(\sigma)}(\tau_{\mathrm{org}}), \mathbf{V}^{(\sigma)}(\tau_{\mathrm{org}})\right) - \mathcal{L}\left(\mathbf{U}^{(\sigma)}(0), \mathbf{V}^{(\sigma)}(0)\right) \longrightarrow 0. \tag{657}$$

Finally,

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}(0)\right) \longrightarrow 0 \tag{658}$$

in probability in the large-system limit. Thus the permutation target exhibits an early microscopic rise from vanishing accuracy to perfect accuracy, but no subsequent accuracy loss.

Proof. By the permutation reduction preceding (442), it is enough to consider the identity target

$$\mathbf{T} = \mathbf{I}_p. \tag{659}$$

Let

$$\delta = q_{\mathrm{sm}}(\alpha) > 0 \tag{660}$$

be the constant from Lemma B.20.

Early organization and monotonicity in the small-q regime. Since $n_{\max} = 1$, Proposition B.13 implies that, for sufficiently large K,

$$\mathbb{P}\left(\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_{\mathrm{org}})\right) = 1\right) \longrightarrow 1 \tag{661}$$

in the sequential limit $\sigma \downarrow 0$ followed by $p, H \to \infty$.

For every fixed p, H,

$$q(\tau_{\mathrm{org}}) = \frac{H\sigma^2}{2} e^{2\tau_{\mathrm{org}}} \longrightarrow 0 \qquad \text{as } \sigma \downarrow 0. \tag{662}$$

Hence, for all sufficiently small σ,

$$q(\tau_{\mathrm{org}}) < \delta. \tag{663}$$

On the high-probability event of Lemma B.20, every margin

$$\Delta_{ci}(\tau) = M_{ii}(\tau) - M_{ci}(\tau), \qquad c \neq i, \tag{664}$$

is strictly increasing for every τ such that $q(\tau) \leq \delta$. Therefore the training accuracy is nondecreasing on the entire interval

$$0 \leq \tau \leq \tau_\sigma(\delta), \qquad \tau_\sigma(\delta) = \frac{1}{2} \log \frac{\delta}{H\sigma^2/2}. \tag{665}$$

On the intersection with the early-perfect event (661), all classification margins are strictly positive at τ_{org}, and therefore remain strictly positive throughout

$$\tau_{\mathrm{org}} \leq \tau \leq \tau_\sigma(\delta). \tag{666}$$

Permanent classification on the nonlinear tail. By Lemma B.22, with probability tending to one over $\mathbf{X}$, there exists $c_{\alpha,\delta} > 0$ such that

$$(\mathbf{M}_{\mathrm{bal}}(q))_{ii} - (\mathbf{M}_{\mathrm{bal}}(q))_{ci} \geq c_{\alpha,\delta} \tag{667}$$

simultaneously for every

$$q \geq \delta, \qquad i \in [p], \qquad c \neq i. \tag{668}$$

A standard Gaussian $\mathbf{X} \in \mathbb{R}^{H \times p}$ has rank $\min\{H, p\}$ almost surely, so Lemma B.21 applies almost surely. For each fixed p, H,

$$\sup_{q \geq \delta} \left\|\widehat{\mathbf{M}}_\sigma(q) - \mathbf{M}_{\mathrm{bal}}(q)\right\|_{\max} \longrightarrow 0 \qquad \text{as } \sigma \downarrow 0. \tag{669}$$

Hence, on the balanced-margin event, for all sufficiently small σ,

$$\sup_{q\ge\delta}\left\|\widehat{\mathbf{M}}_\sigma(q)-\mathbf{M}_{\mathrm{bal}}(q)\right\|_{\max}<\frac{c_{\alpha,\delta}}{4}. \tag{670}$$

It follows that, for every $q\ge\delta$, $i\in[p]$, and $c\neq i$,

$$\begin{aligned}(\widehat{\mathbf{M}}_\sigma(q))_{ii}-(\widehat{\mathbf{M}}_\sigma(q))_{ci}&\ge(\mathbf{M}_{\mathrm{bal}}(q))_{ii}-(\mathbf{M}_{\mathrm{bal}}(q))_{ci}\\&\quad-2\left\|\widehat{\mathbf{M}}_\sigma(q)-\mathbf{M}_{\mathrm{bal}}(q)\right\|_{\max}\\&\ge\frac{c_{\alpha,\delta}}{2}>0.\end{aligned} \tag{671}$$

Since $q(\tau)$ increases from δ to $+\infty$ as τ ranges from $\tau_\sigma(\delta)$ to $+\infty$, this proves

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau)\right)=1\qquad\text{for every }\tau\ge\tau_\sigma(\delta) \tag{672}$$

for all sufficiently small σ on the balanced-margin event.

Combining the two regimes. The two controlled intervals

$$[0,\tau_\sigma(\delta)]\qquad\text{and}\qquad[\tau_\sigma(\delta),\infty) \tag{673}$$

cover the entire trajectory. On the first interval, the accuracy is nondecreasing by Lemma B.20. With probability tending to one it has already reached one by $\tau_{\mathrm{org}}<\tau_\sigma(\delta)$. It therefore remains equal to one until the end of the first interval, while (672) keeps it equal to one throughout the second interval. Thus, on the intersection of these events,

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_1)\right)\le\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau_2)\right)\qquad\text{for every }0\le\tau_1\le\tau_2, \tag{674}$$

and

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}(\tau)\right)=1\qquad\text{for every }\tau\ge\tau_{\mathrm{org}}. \tag{675}$$

For completeness, let $\mathcal{G}_{p,H}$ denote the high-probability event from Lemma B.20, and let $\mathcal{B}_{p,H}$ denote the uniform balanced-margin event from Lemma B.22. Both events are independent of σ, and

$$\mathbb{P}(\mathcal{G}_{p,H})\longrightarrow1,\qquad\mathbb{P}(\mathcal{B}_{p,H})\longrightarrow1 \tag{676}$$

as $p,H\to\infty$ with $H/p\to\alpha$. The almost-sure convergence in (669) implies, for each fixed p,H,

$$\mathbb{P}\left(\sup_{q\ge\delta}\left\|\widehat{\mathbf{M}}_\sigma(q)-\mathbf{M}_{\mathrm{bal}}(q)\right\|_{\max}\ge\frac{c_{\alpha,\delta}}{4}\right)\longrightarrow0 \tag{677}$$

as $\sigma\downarrow0$. Combining this with (661) and taking the limits in the prescribed order proves (654) and (655).

The microscopic-predictor and flat-loss statements (656) and (657) follow directly from Proposition B.13. Finally, (658) is (317). □

The permanence result is particularly informative in the undercomplete regime

$$0<\alpha<1. \tag{678}$$

Here

$$\mathrm{rank}(\mathbf{T})=p>H, \tag{679}$$

so the identity target cannot be represented exactly by an end-to-end matrix $\mathbf{V}^\top\mathbf{U}$ of rank at most H. Nevertheless, the theorem shows that perfect classification is reached microscopically and then maintained for all later training times.

In the vanishing-initialization limit, the balanced trajectory converges to the orthogonal projector onto the row space of $\mathbf{X}$,

$$\mathbf{P}_X=\mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top\right)^{-1}\mathbf{X}, \tag{680}$$

which satisfies

$$\|\mathbf{I}_p - \mathbf{P}_X\|_{\mathrm{F}}^2 = p - H. \tag{681}$$

For the unary identity target,

$$\mathcal{L}(\mathbf{M}) = \frac{1}{2p}\|\mathbf{M} - \mathbf{I}_p\|_{\mathrm{F}}^2. \tag{682}$$

For every fixed p, H with $H < p$, the global balanced reduction therefore implies

$$\lim_{\sigma\downarrow 0}\limsup_{\tau\to\infty}\left|\mathcal{L}\left(\mathbf{M}^{(\sigma)}(\tau)\right) - \frac{p-H}{2p}\right| = 0. \tag{683}$$

Taking the large-system limit afterward with $H/p \to \alpha < 1$ gives

$$\frac{p-H}{2p} \longrightarrow \frac{1-\alpha}{2}. \tag{684}$$

Thus perfect classification does not require convergence to the least-squares target itself: in the undercomplete regime, a nonvanishing least-squares error remains even though the training accuracy is permanently equal to one.

B.3 Deeper linear networks: $L \geq 3$

We now turn to depths $L \geq 3$. In contrast to the two-factor case, the leading near-origin dynamics are already nonlinear, and escape from vanishing initialization occurs on a diverging time scale. Nevertheless, the same qualitative separation between target-driven organization and a later predictive bottleneck persists. On the first near-origin scale, the normalized dynamics develop a coherent component in the direction of the full target $\mathbf{T}$, so that the target classifier can emerge while the end-to-end predictor remains microscopic and the loss is asymptotically unchanged.

This distinction is important when relating our picture to the saddle-escape analysis of Jacot et al. (2021). From the viewpoint of the loss and the parameter-scale escape from the rank-zero saddle, the width-one trajectory identified there is the first nontrivial escape dynamics. For classification, however, nontrivial behavior appears strictly earlier: the predictor can already reorganize enough to recover the classifier of the full target while remaining microscopic and producing a vanishing change in loss.

At a later scale, when the largest singular value of $\mathbf{T}$ is simple,

$$s_1 > s_2, \tag{685}$$

we use Jacot et al. (2021, Theorem 2) to show that the first escape from the origin is asymptotically governed by a width-one trajectory aligned with the leading singular mode. In our accuracy-based picture, this escape is therefore not the first stage of the dynamics, but rather the subsequent degeneracy bottleneck.

Recovery of the same Gaussian-initialized trajectory is not known in the generality needed here for $L \geq 3$. We therefore treat the first two stages rigorously and return at the end of the subsection to a spectrally aligned setting in which recruitment of the remaining modes and final recovery can be followed exactly.

Near-origin dynamics. We retain the isotropic setting and the rescaled time

$$\tau = \gamma t, \tag{686}$$

so that, up to an additive constant, the loss is

$$\mathcal{L} = \frac{\gamma}{2}\|\mathbf{M} - \mathbf{T}\|_{\mathrm{F}}^2, \qquad \mathbf{M} = \mathbf{W}_L \cdots \mathbf{W}_1. \tag{687}$$

For simplicity, all hidden layers have width H:

$$\mathbf{W}_1 \in \mathbb{R}^{H\times p}, \qquad \mathbf{W}_\ell \in \mathbb{R}^{H\times H}, \quad 2 \leq \ell \leq L-1, \qquad \mathbf{W}_L \in \mathbb{R}^{C\times H}. \tag{688}$$

For $b \geq a$, write

$$\mathbf{W}_{b:a} = \mathbf{W}_b \mathbf{W}_{b-1} \cdots \mathbf{W}_a, \tag{689}$$

with the natural identity convention for empty products. Gradient flow in the rescaled time is

$$\dot{\mathbf{W}}_\ell = \mathbf{W}_{L:\ell+1}^\top (\mathbf{T} - \mathbf{M}) \mathbf{W}_{\ell-1:1}^\top, \qquad \ell = 1, \dots, L. \tag{690}$$

We initialize

$$\mathbf{W}_\ell(0) = \sigma \mathbf{G}_\ell, \qquad \ell = 1, \dots, L, \tag{691}$$

where all entries of $\mathbf{G}_1, \dots, \mathbf{G}_L$ are independent standard Gaussian variables. For $L \geq 3$, the right-hand side of (690) is of order σ^{L-1} near the origin. Since the factors themselves have size $O(\sigma)$, the first $O(1)$ evolution of their normalized values occurs on the diverging physical-time scale

$$s = \sigma^{L-2} \tau. \tag{692}$$

Define normalized factors by

$$\mathbf{W}_\ell(\tau) = \sigma \mathbf{Z}_\ell^{(\sigma)}(s), \qquad s = \sigma^{L-2}\tau, \tag{693}$$

and let

$$\mathbf{P}^{(\sigma)}(s) = \mathbf{Z}_L^{(\sigma)}(s) \cdots \mathbf{Z}_1^{(\sigma)}(s). \tag{694}$$

Substitution into (690) gives the exact normalized dynamics

$$\frac{d\mathbf{Z}_\ell^{(\sigma)}}{ds} = \left(\mathbf{Z}_{L:\ell+1}^{(\sigma)}\right)^\top \left(\mathbf{T} - \sigma^L \mathbf{P}^{(\sigma)}\right) \left(\mathbf{Z}_{\ell-1:1}^{(\sigma)}\right)^\top, \qquad \mathbf{Z}_\ell^{(\sigma)}(0) = \mathbf{G}_\ell. \tag{695}$$

Hence, unlike for $L = 2$, the vanishing-initialization limit is already a nonlinear dynamical system:

$$\frac{d\mathbf{Z}_\ell}{ds} = \mathbf{Z}_{L:\ell+1}^\top \mathbf{T} \mathbf{Z}_{\ell-1:1}^\top, \qquad \mathbf{Z}_\ell(0) = \mathbf{G}_\ell. \tag{696}$$

We refer to (696) as the homogeneous near-origin flow. The following lemma justifies its use on every compact interval preceding a possible blow-up of the limiting system.

Lemma B.24 (Near-origin reduction for $L \geq 3$)**.** *Fix p, C, H, L, the target $\mathbf{T}$, and a realization of $\mathbf{G}_1, \dots, \mathbf{G}_L$. Let $(\mathbf{Z}_1, \dots, \mathbf{Z}_L)$ be the maximal solution of* (696)*, with maximal forward existence time $s_* \in (0, \infty]$. Then, for every $S < s_*$,*

$$\max_{1 \leq \ell \leq L} \sup_{0 \leq s \leq S} \left\| \mathbf{Z}_\ell^{(\sigma)}(s) - \mathbf{Z}_\ell(s) \right\|_{\mathrm{F}} \longrightarrow 0 \qquad \text{as } \sigma \downarrow 0. \tag{697}$$

Consequently, if

$$\mathbf{P}(s) = \mathbf{Z}_L(s) \cdots \mathbf{Z}_1(s), \tag{698}$$

then

$$\sup_{0 \leq s \leq S} \left\| \frac{\mathbf{M}^{(\sigma)}\left(s/\sigma^{L-2}\right)}{\sigma^L} - \mathbf{P}(s) \right\|_{\mathrm{F}} \longrightarrow 0. \tag{699}$$

In particular,

$$\sup_{0 \leq s \leq S} \left\| \mathbf{M}^{(\sigma)}\left(s/\sigma^{L-2}\right) \right\|_{\mathrm{F}} = O(\sigma^L), \tag{700}$$

and

$$\sup_{0 \leq s \leq S} \left| \mathcal{L}\left(\mathbf{M}^{(\sigma)}\left(s/\sigma^{L-2}\right)\right) - \mathcal{L}\left(\mathbf{M}^{(\sigma)}(0)\right) \right| \longrightarrow 0. \tag{701}$$

Proof. The result follows from continuous dependence of the normalized ODE on the parameter σ. Let

$$\mathcal{Z} = (\mathbf{Z}_1, \dots, \mathbf{Z}_L), \qquad \|\mathcal{Z}\| = \max_{1 \leq \ell \leq L} \|\mathbf{Z}_\ell\|_{\mathrm{F}}, \tag{702}$$

and denote the vector fields in (695) and (696) by

$$\dot{\mathcal{Z}}^{(\sigma)} = F_\sigma\left(\mathcal{Z}^{(\sigma)}\right), \qquad \dot{\mathcal{Z}} = F_0(\mathcal{Z}). \tag{703}$$

Fix $S < s_*$. Since the limiting solution is continuous on the compact interval $[0, S]$,

$$R_0 = \sup_{0 \leq s \leq S} \|\mathcal{Z}(s)\| < \infty. \tag{704}$$

Choose $R > R_0 + 1$. On the closed ball

$$\|\mathcal{Z}\| \le R, \tag{705}$$

the polynomial vector field F_0 is Lipschitz, so there exists $K_R < \infty$ such that

$$\left\|F_0(\mathcal{Z}) - F_0(\widetilde{\mathcal{Z}})\right\| \le K_R \left\|\mathcal{Z} - \widetilde{\mathcal{Z}}\right\|. \tag{706}$$

The only difference between F_σ and F_0 is the term containing $\sigma^L \mathbf{P}^{(\sigma)}$. Each component of this remainder is a product of $2L-1$ normalized factors. Hence there exists $C_R < \infty$ such that

$$\sup_{\|\mathcal{Z}\| \le R} \|F_\sigma(\mathcal{Z}) - F_0(\mathcal{Z})\| \le C_R \sigma^L. \tag{707}$$

Let s_σ be the first time in $[0, S]$ at which

$$\left\|\mathcal{Z}^{(\sigma)}(s) - \mathcal{Z}(s)\right\| = 1, \tag{708}$$

with $s_\sigma = S$ if no such time exists. On $[0, s_\sigma]$, both trajectories remain inside the radius-R ball. Therefore, in the upper-Dini-derivative sense,

$$\frac{d^+}{ds}\left\|\mathcal{Z}^{(\sigma)} - \mathcal{Z}\right\| \le K_R \left\|\mathcal{Z}^{(\sigma)} - \mathcal{Z}\right\| + C_R \sigma^L. \tag{709}$$

The two trajectories have the same initial condition, so Gronwall's inequality yields

$$\sup_{0 \le s \le s_\sigma} \left\|\mathcal{Z}^{(\sigma)}(s) - \mathcal{Z}(s)\right\| \le C_R \sigma^L \frac{e^{K_R S} - 1}{K_R}, \tag{710}$$

with the usual interpretation when $K_R = 0$. The right-hand side tends to zero. Hence, for all sufficiently small σ, it is strictly smaller than one, so the first exit cannot occur before S. Thus $s_\sigma = S$, proving (697).

By (693),

$$\mathbf{M}^{(\sigma)}\left(\frac{s}{\sigma^{L-2}}\right) = \sigma^L \mathbf{Z}_L^{(\sigma)}(s) \cdots \mathbf{Z}_1^{(\sigma)}(s). \tag{711}$$

Uniform convergence of the factors on $[0, S]$ therefore gives (699), and hence (700).

Finally, using (687),

$$\begin{aligned} &\mathcal{L}\left(\mathbf{M}^{(\sigma)}\left(s/\sigma^{L-2}\right)\right) - \mathcal{L}\left(\mathbf{M}^{(\sigma)}(0)\right) \\ &\quad = \gamma\left[\frac{1}{2}\left(\left\|\mathbf{M}^{(\sigma)}\left(s/\sigma^{L-2}\right)\right\|_{\mathrm{F}}^2 - \left\|\mathbf{M}^{(\sigma)}(0)\right\|_{\mathrm{F}}^2\right)\right. \\ &\qquad \left. - \left\langle \mathbf{M}^{(\sigma)}\left(s/\sigma^{L-2}\right) - \mathbf{M}^{(\sigma)}(0), \mathbf{T}\right\rangle_{\mathrm{F}}\right]. \end{aligned} \tag{712}$$

Both predictors are $O(\sigma^L)$ uniformly on $s \in [0, S]$, while the outer problem is fixed in the inner limit. The right-hand side of (712) therefore tends uniformly to zero, proving (701). □

Thus vanishing initialization separates the near-origin evolution into two scales. The original factors remain of size $O(\sigma)$, and the end-to-end predictor remains microscopic, $\mathbf{M} = O(\sigma^L)$, while the normalized factors undergo a nontrivial nonlinear evolution on the slow time $s = \sigma^{L-2}\tau$. We next show that, at large width, the first classifier selected by this homogeneous flow is governed by the full target $\mathbf{T}$.

Target-driven organization. We now show that the homogeneous near-origin dynamics can already recover the classifier induced by the full target $\mathbf{T}$, before the rank-one escape analyzed by Jacot et al. (2021). The mechanism combines the early gradient-flow dynamics selected by vanishing initialization with concentration at large width. Although the initial end-to-end product is centered, the homogeneous gradient flow immediately generates a coherent first-order drift in the direction of the

full target $\mathbf{T}$. At large width, fluctuations of both the initial product and this first-order drift concentrate sufficiently strongly that the target-directed component dominates on an early microscopic time scale. Thus the target classifier can emerge through genuine gradient-flow evolution while the end-to-end predictor remains microscopic and the loss is asymptotically unchanged.

It is convenient to introduce a second rescaling that absorbs the natural width dependence of the homogeneous system. Define

$$u = H^{(L-2)/2}s = H^{(L-2)/2}\sigma^{L-2}\tau, \qquad \overline{\mathbf{Z}}_\ell(u) = \frac{1}{\sqrt{H}}\mathbf{Z}_\ell(s). \tag{713}$$

Then the homogeneous flow (696) becomes

$$\frac{d\overline{\mathbf{Z}}_\ell}{du} = \overline{\mathbf{Z}}_{L:\ell+1}^\top \mathbf{T}\overline{\mathbf{Z}}_{\ell-1:1}^\top, \qquad \overline{\mathbf{Z}}_\ell(0) = \overline{\mathbf{G}}_\ell := \frac{1}{\sqrt{H}}\mathbf{G}_\ell. \tag{714}$$

Let

$$\mathbf{Q}(u) = \overline{\mathbf{Z}}_L(u)\cdots\overline{\mathbf{Z}}_1(u). \tag{715}$$

Since

$$\mathbf{M}^{(\sigma)}(\tau) = \sigma^L H^{L/2}\mathbf{Q}(u) \tag{716}$$

in the vanishing-initialization limit, the positive scalar factor $\sigma^L H^{L/2}$ does not affect classification.

For each training example, define the class predicted by the target,

$$\widehat{c}_i = \arg\max_{c\in[C]} (\mathbf{T}\mathbf{x}_i)_c\,, \tag{717}$$

and assume that these maximizers are unique. Let

$$\delta_{\mathbf{T}} = \min_i \min_{c\neq\widehat{c}_i} \left[(\mathbf{T}\mathbf{x}_i)_{\widehat{c}_i} - (\mathbf{T}\mathbf{x}_i)_c\right], \qquad R = \max_i \|\mathbf{x}_i\|_2. \tag{718}$$

Theorem B.25 (Target-driven organization for $L \geq 3$)**.** *Fix the depth $L \geq 3$. Consider a sequence of problems indexed by H such that*

$$\frac{p}{H} = O(1), \qquad \frac{C}{H} = O(1), \qquad R = O(1), \qquad \|\mathbf{T}\|_{\mathrm{op}} = O(1), \qquad \delta_{\mathbf{T}} \geq \delta_0 > 0. \tag{719}$$

Let

$$\ell_H = \log(2NCH), \tag{720}$$

and suppose that

$$\frac{\ell_H^{L-1}}{\sqrt{H}} \longrightarrow 0. \tag{721}$$

Choose any deterministic sequence $b_H \to \infty$ sufficiently slowly that

$$u_{\mathrm{org}} := b_H \frac{\ell_H^{L/2}}{\sqrt{H}} \longrightarrow 0. \tag{722}$$

Then there exist events $\mathcal{E}_H$, depending only on the Gaussian initialization, such that

$$\mathbb{P}(\mathcal{E}_H) \longrightarrow 1, \tag{723}$$

and, for all sufficiently large H, on $\mathcal{E}_H$ the homogeneous flow (714) *is defined up to time u_{org} and satisfies*

$$\arg\max_{c\in[C]} (\mathbf{Q}(u_{\mathrm{org}})\mathbf{x}_i)_c = \widehat{c}_i \qquad \textit{for every } i. \tag{724}$$

Equivalently, $\mathbf{Q}(u_{\mathrm{org}})$ induces exactly the same classifier as the full target $\mathbf{T}$ on the training set.

Couple the exact gradient flows for different initialization scales by

$$\mathbf{W}_\ell^{(\sigma)}(0) = \sigma\mathbf{G}_\ell, \qquad \ell = 1,\dots,L. \tag{725}$$

Define

$$\tau_{\mathrm{org}}^{(\sigma,H)} = \sigma^{-(L-2)}H^{-(L-2)/2}u_{\mathrm{org}}. \tag{726}$$

For every sufficiently large H and every realization in $\mathcal{E}_H$, the exact gradient flow satisfies, for all sufficiently small $\sigma > 0$,

$$\arg\max_{c\in[C]} \left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right)\mathbf{x}_i\right)_c = \widehat{c}_i \qquad \textit{for every } i, \tag{727}$$

and, as $\sigma \downarrow 0$,

$$\left\|\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right)\right\|_{\mathrm{F}} \longrightarrow 0, \tag{728}$$

together with

$$\mathcal{L}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right)\right) - \mathcal{L}\left(\mathbf{M}^{(\sigma)}(0)\right) \longrightarrow 0. \tag{729}$$

Consequently, for every fixed $\varepsilon > 0$,

$$\begin{aligned}\lim_{H\to\infty} \liminf_{\sigma\downarrow 0} \mathbb{P}\Bigg(&\arg\max_{c\in[C]} \left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right)\mathbf{x}_i\right)_c = \widehat{c}_i \textit{ for every } i, \\ &\left\|\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right)\right\|_{\mathrm{F}} < \varepsilon, \\ &\left|\mathcal{L}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right)\right) - \mathcal{L}\left(\mathbf{M}^{(\sigma)}(0)\right)\right| < \varepsilon\Bigg) = 1.\end{aligned} \tag{730}$$

In particular,

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right)\right) = A_{\mathbf{T}} \tag{731}$$

with probability tending to one in the sequential limit $\sigma \downarrow 0$ followed by $H \to \infty$. Thus the classifier can reorganize to that of the full target while both the end-to-end predictor and the corresponding change in loss remain asymptotically vanishing.

Proof. We prove that, at the organization time u_{org}, the homogeneous predictor satisfies

$$\mathbf{a}_{ic}^{\top}\mathbf{Q}(u_{\mathrm{org}})\mathbf{x}_i = L u_{\mathrm{org}} \mathbf{a}_{ic}^{\top}\mathbf{T}\mathbf{x}_i + o(u_{\mathrm{org}}) \tag{732}$$

simultaneously over all score comparisons, on events whose probability tends to one.

For every training example i and every $c \neq \widehat{c}_i$, define

$$\mathbf{a}_{ic} = \mathbf{e}_{\widehat{c}_i} - \mathbf{e}_c. \tag{733}$$

Then

$$\|\mathbf{a}_{ic}\|_2 = \sqrt{2}, \qquad \mathbf{a}_{ic}^{\top}\mathbf{T}\mathbf{x}_i \geq \delta_{\mathbf{T}} \geq \delta_0. \tag{734}$$

It therefore suffices to prove

$$\mathbf{a}_{ic}^{\top}\mathbf{Q}(u_{\mathrm{org}})\mathbf{x}_i > 0 \tag{735}$$

simultaneously for every i and every $c \neq \widehat{c}_i$.

Step 1: the first-order drift is $L\mathbf{T}$.

Let

$$\mathbf{Q}_0 = \mathbf{Q}(0) = \overline{\mathbf{G}}_L \cdots \overline{\mathbf{G}}_1, \qquad \mathbf{Q}_1 = \left.\frac{d\mathbf{Q}}{du}\right|_{u=0}. \tag{736}$$

Differentiating the product $\mathbf{Q}(u)$ and using (714) gives

$$\mathbf{Q}_1 = \sum_{\ell=1}^{L} \left(\overline{\mathbf{G}}_{L:\ell+1}\overline{\mathbf{G}}_{L:\ell+1}^{\top}\right)\mathbf{T}\left(\overline{\mathbf{G}}_{\ell-1:1}^{\top}\overline{\mathbf{G}}_{\ell-1:1}\right), \tag{737}$$

with the natural identity convention for empty products.

For each ℓ, the Gaussian layers on the left and right of $\mathbf{T}$ are disjoint and hence independent. Moreover,

$$\mathbb{E}\left[\overline{\mathbf{G}}_{L:\ell+1}\overline{\mathbf{G}}_{L:\ell+1}^{\top}\right] = \mathbf{I}_C, \qquad \mathbb{E}\left[\overline{\mathbf{G}}_{\ell-1:1}^{\top}\overline{\mathbf{G}}_{\ell-1:1}\right] = \mathbf{I}_p. \tag{738}$$

Indeed, these identities follow by successively conditioning on the independent normalized Gaussian layers, each of which preserves the expected squared norm in the relevant orientation. Therefore every summand in (737) has expectation $\mathbf{T}$, and

$$\mathbb{E}\,\mathbf{Q}_1 = L\mathbf{T}. \tag{739}$$

Step 2: scalar concentration for the Gaussian products.

We first establish the particular Gram-matrix estimate required below. Fix an integer $m \geq 1$, let

$$\mathbf{X}_1 \in \mathbb{R}^{H\times d}, \qquad \mathbf{X}_j \in \mathbb{R}^{H\times H}, \quad 2 \leq j \leq m, \tag{740}$$

have mutually independent entries distributed as $\mathcal{N}(0, 1/H)$, and define

$$\mathbf{B}_m = \mathbf{X}_m \cdots \mathbf{X}_1. \tag{741}$$

For every deterministic $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$, there exists a constant $C_m < \infty$, depending only on m, such that

$$\mathbb{E}\left|\mathbf{u}^\top \left(\mathbf{B}_m^\top \mathbf{B}_m - \mathbf{I}_d\right)\mathbf{v}\right|^2 \leq \frac{C_m}{H}\|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2. \tag{742}$$

To prove (742), set

$$\mathbf{z}_0 = \mathbf{u}, \qquad \mathbf{w}_0 = \mathbf{v}, \qquad \mathbf{z}_j = \mathbf{X}_j\mathbf{z}_{j-1}, \qquad \mathbf{w}_j = \mathbf{X}_j\mathbf{w}_{j-1}, \tag{743}$$

and let

$$S_j = \mathbf{z}_j^\top \mathbf{w}_j. \tag{744}$$

Conditional on $\mathbf{z}_{j-1}, \mathbf{w}_{j-1}$, the Gaussian second and fourth moment identities give

$$\mathbb{E}\left[S_j \mid \mathbf{z}_{j-1}, \mathbf{w}_{j-1}\right] = S_{j-1}, \tag{745}$$

and

$$\mathrm{Var}\left(S_j \mid \mathbf{z}_{j-1}, \mathbf{w}_{j-1}\right) = \frac{1}{H}\left(\|\mathbf{z}_{j-1}\|_2^2\|\mathbf{w}_{j-1}\|_2^2 + S_{j-1}^2\right). \tag{746}$$

Thus $(S_j)_{j=0}^m$ is a martingale.

For every deterministic vector $\mathbf{z}$,

$$\mathbb{E}\left[\|\mathbf{X}_j\mathbf{z}\|_2^4 \mid \mathbf{z}\right] = \left(1 + \frac{2}{H}\right)\|\mathbf{z}\|_2^4. \tag{747}$$

Since m is fixed, iteration gives

$$\max_{0\leq j\leq m} \mathbb{E}\|\mathbf{z}_j\|_2^4 \leq C_m\|\mathbf{u}\|_2^4, \qquad \max_{0\leq j\leq m} \mathbb{E}\|\mathbf{w}_j\|_2^4 \leq C_m\|\mathbf{v}\|_2^4. \tag{748}$$

Using orthogonality of martingale increments, (746), Cauchy–Schwarz, and (748), we obtain

$$\begin{aligned}\mathbb{E}|S_m - S_0|^2 &= \sum_{j=1}^m \mathbb{E}\,\mathrm{Var}\left(S_j \mid \mathbf{z}_{j-1}, \mathbf{w}_{j-1}\right) \\ &\leq \frac{C_m}{H}\|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2.\end{aligned} \tag{749}$$

Since

$$S_m - S_0 = \mathbf{u}^\top\left(\mathbf{B}_m^\top\mathbf{B}_m - \mathbf{I}_d\right)\mathbf{v}, \tag{750}$$

this proves (742).

We now apply this estimate to the products in (737). For each ℓ, write

$$\mathbf{L}_\ell = \overline{\mathbf{G}}_{L:\ell+1}\overline{\mathbf{G}}_{L:\ell+1}^\top, \qquad \mathbf{R}_\ell = \overline{\mathbf{G}}_{\ell-1:1}^\top\overline{\mathbf{G}}_{\ell-1:1}. \tag{751}$$

The estimate (742) applies directly to $\mathbf{R}_\ell$, and it applies to $\mathbf{L}_\ell$ after transposing the corresponding Gaussian chain. Thus, for deterministic compatible vectors,

$$\mathbb{E}\left|\mathbf{u}^\top(\mathbf{L}_\ell - \mathbf{I}_C)\mathbf{v}\right|^2 \leq \frac{C_L}{H}\|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2, \tag{752}$$

and

$$\mathbb{E}\left|\mathbf{u}^\top(\mathbf{R}_\ell - \mathbf{I}_p)\mathbf{v}\right|^2 \le \frac{C_L}{H}\|\mathbf{u}\|_2^2\|\mathbf{v}\|_2^2. \tag{753}$$

We also require a second-moment bound for $\mathbf{R}_\ell\mathbf{x}$. By (753),

$$\begin{aligned}\mathbb{E}\,\|(\mathbf{R}_\ell - \mathbf{I}_p)\mathbf{x}\|_2^2 &= \sum_{j=1}^{p}\mathbb{E}\left|\mathbf{e}_j^\top(\mathbf{R}_\ell - \mathbf{I}_p)\mathbf{x}\right|^2 \\ &\le C_L\frac{p}{H}\|\mathbf{x}\|_2^2.\end{aligned} \tag{754}$$

Since $p/H = O(1)$,

$$\mathbb{E}\,\|\mathbf{R}_\ell\mathbf{x}\|_2^2 \le C_L\|\mathbf{x}\|_2^2. \tag{755}$$

Now fix deterministic $\mathbf{a} \in \mathbb{R}^C$ and $\mathbf{x} \in \mathbb{R}^p$. For one summand in (737),

$$\mathbf{L}_\ell\mathbf{T}\mathbf{R}_\ell - \mathbf{T} = (\mathbf{L}_\ell - \mathbf{I}_C)\mathbf{T}\mathbf{R}_\ell + \mathbf{T}(\mathbf{R}_\ell - \mathbf{I}_p). \tag{756}$$

The left and right Gram products are independent. Therefore, conditioning first on $\mathbf{R}_\ell$ and using (752),

$$\begin{aligned}&\mathbb{E}\left|\mathbf{a}^\top(\mathbf{L}_\ell - \mathbf{I}_C)\mathbf{T}\mathbf{R}_\ell\mathbf{x}\right|^2 \\ &\quad\le \frac{C_L}{H}\|\mathbf{a}\|_2^2\mathbb{E}\,\|\mathbf{T}\mathbf{R}_\ell\mathbf{x}\|_2^2 \\ &\quad\le \frac{C_L}{H}\|\mathbf{a}\|_2^2\|\mathbf{T}\|_{\mathrm{op}}^2\|\mathbf{x}\|_2^2.\end{aligned} \tag{757}$$

Likewise, (753) gives

$$\mathbb{E}\left|\mathbf{a}^\top\mathbf{T}(\mathbf{R}_\ell - \mathbf{I}_p)\mathbf{x}\right|^2 \le \frac{C_L}{H}\|\mathbf{a}\|_2^2\|\mathbf{T}\|_{\mathrm{op}}^2\|\mathbf{x}\|_2^2. \tag{758}$$

Since L is fixed, summing the L terms in (737) and using the triangle inequality in L^2 yields

$$\left\|\mathbf{a}^\top\left(\mathbf{Q}_1 - L\mathbf{T}\right)\mathbf{x}\right\|_{L^2} \le \frac{C_L}{\sqrt{H}}\|\mathbf{a}\|_2\|\mathbf{T}\|_{\mathrm{op}}\|\mathbf{x}\|_2. \tag{759}$$

For the initial product, repeated conditioning through the independent Gaussian layers gives the exact second moment

$$\mathbb{E}\left|\mathbf{a}^\top\mathbf{Q}_0\mathbf{x}\right|^2 = \frac{1}{H}\|\mathbf{a}\|_2^2\|\mathbf{x}\|_2^2. \tag{760}$$

Both scalar quantities are polynomials of fixed degree in the underlying standard Gaussian variables: $\mathbf{a}^\top\mathbf{Q}_0\mathbf{x}$ has degree L, while $\mathbf{a}^\top(\mathbf{Q}_1 - L\mathbf{T})\mathbf{x}$ has degree at most $2L-2$. Gaussian hypercontractivity therefore gives, for every $q \ge 2$,

$$\left\|\mathbf{a}^\top\mathbf{Q}_0\mathbf{x}\right\|_{L^q} \le (q-1)^{L/2}\left\|\mathbf{a}^\top\mathbf{Q}_0\mathbf{x}\right\|_{L^2}, \tag{761}$$

and

$$\left\|\mathbf{a}^\top\left(\mathbf{Q}_1 - L\mathbf{T}\right)\mathbf{x}\right\|_{L^q} \le (q-1)^{L-1}\left\|\mathbf{a}^\top\left(\mathbf{Q}_1 - L\mathbf{T}\right)\mathbf{x}\right\|_{L^2}. \tag{762}$$

Taking

$$q_H = \max\{2, 2\ell_H\}, \tag{763}$$

and applying Markov's inequality followed by a union bound over at most NC score comparisons, we obtain constants $C_L < \infty$ such that, with probability tending to one,

$$\max_i\max_{c\neq\widehat{c}_i}\left|\mathbf{a}_{ic}^\top\mathbf{Q}_0\mathbf{x}_i\right| \le C_L R\frac{\ell_H^{L/2}}{\sqrt{H}}, \tag{764}$$

and

$$\max_i\max_{c\neq\widehat{c}_i}\left|\mathbf{a}_{ic}^\top\left(\mathbf{Q}_1 - L\mathbf{T}\right)\mathbf{x}_i\right| \le C_L R\|\mathbf{T}\|_{\mathrm{op}}\frac{\ell_H^{L-1}}{\sqrt{H}}. \tag{765}$$

By (721), the right-hand side of (765) tends to zero.

Step 3: uniform second-order control of the homogeneous flow.

Under the aspect-ratio assumptions in (719), standard Gaussian operator-norm bounds imply that there exists a deterministic constant $K < \infty$ such that

$$\mathbb{P}\left(\max_{1\le\ell\le L}\left\|\overline{\mathbf{G}}_\ell\right\|_{\mathrm{op}} \le K\right) \longrightarrow 1. \tag{766}$$

Work on this event. Because the vector field in (714) is polynomial of fixed degree and $\|\mathbf{T}\|_{\mathrm{op}}$ is uniformly bounded, there exist constants $u_0 > 0$ and $K_1 < \infty$, independent of H, such that the homogeneous solution exists on $[0, u_0]$ and

$$\max_{\ell}\sup_{0\le u\le u_0}\left\|\overline{\mathbf{Z}}_\ell(u)\right\|_{\mathrm{op}} \le 2K. \tag{767}$$

Indeed, while all factor norms are at most $2K$,

$$\left\|\frac{d\overline{\mathbf{Z}}_\ell}{du}\right\|_{\mathrm{op}} \le (2K)^{L-1}\|\mathbf{T}\|_{\mathrm{op}}, \tag{768}$$

so (767) follows from a standard first-exit argument after choosing u_0 sufficiently small.

Moreover,

$$\left\|\overline{\mathbf{Z}}_\ell(u) - \overline{\mathbf{G}}_\ell\right\|_{\mathrm{op}} \le K_1 u, \qquad 0 \le u \le u_0. \tag{769}$$

Since $\mathbf{Q}'(u)$ is a polynomial of the normalized factors and one copy of $\mathbf{T}$, it is uniformly Lipschitz on the bounded set (767). Hence there exists $K_2 < \infty$, independent of H, such that

$$\|\mathbf{Q}'(u) - \mathbf{Q}_1\|_{\mathrm{op}} \le K_2 u, \qquad 0 \le u \le u_0. \tag{770}$$

Integrating,

$$\mathbf{Q}(u) = \mathbf{Q}_0 + u\mathbf{Q}_1 + \boldsymbol{\mathcal{R}}(u), \qquad \|\boldsymbol{\mathcal{R}}(u)\|_{\mathrm{op}} \le \frac{K_2}{2}u^2. \tag{771}$$

Since $u_{\mathrm{org}} \to 0$, this expansion applies at $u = u_{\mathrm{org}}$ for all sufficiently large H.

Step 4: positivity of every target score margin.

Let $\mathcal{E}_H$ be the intersection of the operator-norm event (766) with the two uniform concentration events (764) and (765). Then

$$\mathbb{P}(\mathcal{E}_H) \longrightarrow 1. \tag{772}$$

Fix a realization in $\mathcal{E}_H$, an example i, and a class $c \ne \widehat{c}_i$. Using (771),

$$\begin{aligned}\mathbf{a}_{ic}^\top\mathbf{Q}(u_{\mathrm{org}})\mathbf{x}_i = {} & Lu_{\mathrm{org}}\mathbf{a}_{ic}^\top\mathbf{T}\mathbf{x}_i + \mathbf{a}_{ic}^\top\mathbf{Q}_0\mathbf{x}_i \\ & + u_{\mathrm{org}}\mathbf{a}_{ic}^\top\left(\mathbf{Q}_1 - L\mathbf{T}\right)\mathbf{x}_i + \mathbf{a}_{ic}^\top\boldsymbol{\mathcal{R}}(u_{\mathrm{org}})\mathbf{x}_i.\end{aligned} \tag{773}$$

The target-driven term satisfies

$$Lu_{\mathrm{org}}\mathbf{a}_{ic}^\top\mathbf{T}\mathbf{x}_i \ge L\delta_0 u_{\mathrm{org}}. \tag{774}$$

For the initial random product, (764) and (722) give

$$\frac{\max_i\max_{c\ne\widehat{c}_i}\left|\mathbf{a}_{ic}^\top\mathbf{Q}_0\mathbf{x}_i\right|}{u_{\mathrm{org}}} \le \frac{C_L R}{b_H} \longrightarrow 0. \tag{775}$$

Similarly, (765) and (721) give

$$\max_i\max_{c\ne\widehat{c}_i}\left|\mathbf{a}_{ic}^\top\left(\mathbf{Q}_1 - L\mathbf{T}\right)\mathbf{x}_i\right| \longrightarrow 0 \tag{776}$$

uniformly on $\mathcal{E}_H$.

Finally,

$$\begin{aligned}\left|\mathbf{a}_{ic}^\top\boldsymbol{\mathcal{R}}(u_{\mathrm{org}})\mathbf{x}_i\right| &\le \|\mathbf{a}_{ic}\|_2\|\mathbf{x}_i\|_2\left\|\boldsymbol{\mathcal{R}}(u_{\mathrm{org}})\right\|_{\mathrm{op}} \\ &\le \sqrt{2}R\frac{K_2}{2}u_{\mathrm{org}}^2 = o(u_{\mathrm{org}}).\end{aligned} \tag{777}$$

Combining these estimates gives, uniformly over every i and $c \neq \widehat{c}_i$,

$$\mathbf{a}_{ic}^{\top}\mathbf{Q}(u_{\mathrm{org}})\mathbf{x}_i \geq L\delta_0 u_{\mathrm{org}} - o(u_{\mathrm{org}}) > 0 \tag{778}$$

for all sufficiently large H, on $\mathcal{E}_H$. This proves (724).

Step 5: transfer to the exact gradient flow.

Define the corresponding homogeneous slow time

$$s_{\mathrm{org}} = H^{-(L-2)/2}u_{\mathrm{org}}. \tag{779}$$

On $\mathcal{E}_H$, the homogeneous flow exists on the entire interval $[0, u_0]$, and $u_{\mathrm{org}} < u_0$ for all sufficiently large H. Equivalently, the original homogeneous flow (696) exists at least until s_{org}.

Lemma B.24 can therefore be applied, for every fixed sufficiently large H and every realization in $\mathcal{E}_H$, on a compact interval containing s_{org}. Using the relation between the s- and u-rescalings gives

$$\frac{\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right)}{\sigma^L H^{L/2}} \longrightarrow \mathbf{Q}(u_{\mathrm{org}}) \qquad \text{as } \sigma \downarrow 0. \tag{780}$$

By (778), every limiting score comparison has strictly positive margin on $\mathcal{E}_H$. Hence, for every such realization and all sufficiently small σ, the exact predictor has the same predicted class as $\mathbf{Q}(u_{\mathrm{org}})$, and therefore as $\mathbf{T}$, on every training example. This proves (727).

For fixed H, the quantity $H^{L/2}\|\mathbf{Q}(u_{\mathrm{org}})\|_{\mathrm{F}}$ is finite. Thus (780) implies

$$\left\|\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right)\right\|_{\mathrm{F}} = O(\sigma^L) \longrightarrow 0, \tag{781}$$

which proves (728).

Finally, using (687),

$$\begin{aligned}
&\mathcal{L}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right)\right) - \mathcal{L}\left(\mathbf{M}^{(\sigma)}(0)\right) \\
&\quad = \gamma\left[\frac{1}{2}\left(\left\|\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right)\right\|_{\mathrm{F}}^2 - \left\|\mathbf{M}^{(\sigma)}(0)\right\|_{\mathrm{F}}^2\right)\right. \\
&\qquad \left. - \left\langle \mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right) - \mathbf{M}^{(\sigma)}(0), \mathbf{T}\right\rangle_{\mathrm{F}}\right].
\end{aligned} \tag{782}$$

At fixed outer problem, both predictors are $O(\sigma^L)$, while γ and $\mathbf{T}$ are fixed. Therefore (782) tends to zero as $\sigma \downarrow 0$, proving (729).

All three conclusions hold for every realization in $\mathcal{E}_H$, and $\mathbb{P}(\mathcal{E}_H) \to 1$. Taking first the inner limit $\sigma \downarrow 0$ and then the outer limit $H \to \infty$ proves (730) and (731). □

The theorem identifies a stage of the dynamics that is invisible at the level of macroscopic loss escape. Starting from the centered random end-to-end product, the homogeneous gradient flow generates a coherent target-directed component, and large width suppresses the competing initialization fluctuations sufficiently for this component to determine all training predictions. Consequently, the classifier induced by the full target can emerge strictly before the rank-one escape from the origin.

Rank-one escape and the predictive bottleneck. The target-driven organization above takes place while the end-to-end predictor is still microscopic. We now turn to the later escape from the rank-zero saddle. This is the scale analyzed by Jacot et al. (2021, Theorem 2). Their result shows that, for a deep linear network initialized at vanishing scale, the first nontrivial parameter-scale escape is asymptotically described by a width-one gradient-flow trajectory, provided that the largest singular value of the gradient of the end-to-end objective at the origin is simple.

In our setting, after the time rescaling used throughout this section, the end-to-end objective is

$$\mathcal{C}(\mathbf{M}) = \frac{1}{2}\|\mathbf{M} - \mathbf{T}\|_{\mathrm{F}}^2, \qquad \nabla\mathcal{C}(\mathbf{0}) = -\mathbf{T}. \tag{783}$$

Write

$$\mathbf{T} = \sum_{k=1}^{r} s_k \mathbf{p}_k \mathbf{q}_k^\top, \qquad s_1 > s_2, \tag{784}$$

where we set $s_2 := 0$ when $r = 1$. We denote the unit leading singular direction by

$$\mathbf{D}_1 := \mathbf{p}_1 \mathbf{q}_1^\top. \tag{785}$$

Thus the usual rank-one truncation is

$$\mathbf{T}_1 = s_1 \mathbf{D}_1. \tag{786}$$

The strict inequality $s_1 > s_2$ is exactly the nondegeneracy condition required for the first saddle escape: the largest singular value of $\nabla \mathcal{C}(\mathbf{0}) = -\mathbf{T}$ is simple. The distinguished escape direction is therefore determined by $\mathbf{p}_1, \mathbf{q}_1$.

Proposition B.26 (Rank-one escape path)**.** *Fix $L \geq 3$, the dimensions p, C, H, and a target $\mathbf{T}$ satisfying* (784)*. Couple the Gaussian initializations for all $\sigma > 0$ by first sampling independent standard Gaussian matrices $\mathbf{G}_1, \ldots, \mathbf{G}_L$ and then setting*

$$\mathbf{W}_\ell^{(\sigma)}(0) = \sigma \mathbf{G}_\ell, \qquad \ell = 1, \ldots, L. \tag{787}$$

There exists an event $\mathcal{E}_{\mathrm{esc}}$, depending only on $(\mathbf{G}_1, \ldots, \mathbf{G}_L)$, such that

$$\mathbb{P}(\mathcal{E}_{\mathrm{esc}}) = 1 \quad \text{if } L = 3, \qquad \mathbb{P}(\mathcal{E}_{\mathrm{esc}}) \geq \frac{1}{2} \quad \text{if } L > 3. \tag{788}$$

On $\mathcal{E}_{\mathrm{esc}}$, there exist time shifts $\tau_{\mathrm{esc}}^{(\sigma)}$ and a deterministic function

$$m : \mathbb{R} \to (0, s_1) \tag{789}$$

such that, for every fixed $t \in \mathbb{R}$,

$$\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{esc}}^{(\sigma)} + t\right) \longrightarrow m(t)\mathbf{D}_1 \qquad \text{as } \sigma \downarrow 0. \tag{790}$$

The scalar amplitude is strictly increasing and satisfies

$$\dot{m} = L\,(s_1 - m)\, m^{2-2/L}, \tag{791}$$

with

$$\lim_{t \to -\infty} m(t) = 0, \qquad \lim_{t \to +\infty} m(t) = s_1. \tag{792}$$

Moreover, define the Euclidean parameter norm by

$$\left\| \boldsymbol{\theta}^{(\sigma)}(\tau) \right\|_2^2 := \sum_{\ell=1}^{L} \left\| \mathbf{W}_\ell^{(\sigma)}(\tau) \right\|_{\mathrm{F}}^2. \tag{793}$$

There exists $\rho_0 > 0$ such that, for every fixed $\rho \in (0, \rho_0)$, the first hitting time

$$\tau_\rho^{(\sigma)} := \inf\left\{ \tau \geq 0 : \left\| \boldsymbol{\theta}^{(\sigma)}(\tau) \right\|_2 = \rho \right\} \tag{794}$$

is finite for all sufficiently small σ on $\mathcal{E}_{\mathrm{esc}}$, and

$$\mathbf{M}^{(\sigma)}\left(\tau_\rho^{(\sigma)}\right) \longrightarrow \left(\frac{\rho}{\sqrt{L}}\right)^L \mathbf{D}_1 \qquad \text{as } \sigma \downarrow 0. \tag{795}$$

Proof. We apply Jacot et al. (2021, Theorem 2) to the objective (783). The initialization (787) has precisely the form of a fixed Gaussian direction in parameter space multiplied by a vanishing scalar initialization scale.

Moreover,

$$\nabla \mathcal{C}(\mathbf{0}) = -\mathbf{T}, \tag{796}$$

and the assumption $s_1 > s_2$ makes its largest singular value simple. The cited theorem therefore gives an escape event satisfying (788), realization-dependent time shifts $\tau_{\mathrm{esc}}^{(\sigma)}$, hidden-layer orthogonal changes of basis, and a limiting width-one gradient-flow trajectory.

The hidden-layer orthogonal transformations and the canonical embedding of a width-one network into a width-H network leave both the end-to-end matrix and the Euclidean parameter norm invariant. The optimal first escape direction is determined by the leading left and right singular vectors of $-\mathbf{T}$, with the orientation corresponding to descent of the loss. Consequently, the end-to-end matrix of the limiting trajectory is a positive multiple of $\mathbf{D}_1$. We may therefore write

$$\mathbf{M}_{\mathrm{esc}}(t) = m(t)\mathbf{D}_1, \qquad m(t) > 0. \tag{797}$$

This proves (790) once the scalar trajectory m is identified.

We next show that the limiting width-one trajectory is balanced. Write its active factors as

$$\mathbf{W}_1^*(t) = \alpha_1(t)\mathbf{q}_1^\top, \qquad \mathbf{W}_\ell^*(t) = \alpha_\ell(t), \quad 2 \le \ell \le L-1, \qquad \mathbf{W}_L^*(t) = \alpha_L(t)\mathbf{p}_1, \tag{798}$$

where the middle factors are scalars. Then

$$m(t) = \prod_{\ell=1}^{L} \alpha_\ell(t). \tag{799}$$

Deep-linear gradient flow preserves the adjacent-layer imbalance quantities

$$\mathbf{W}_{\ell+1}^\top \mathbf{W}_{\ell+1} - \mathbf{W}_\ell \mathbf{W}_\ell^\top. \tag{800}$$

Restricted to the width-one trajectory (798), these invariants become

$$\alpha_{\ell+1}(t)^2 - \alpha_\ell(t)^2 = \text{constant}, \qquad \ell = 1, \dots, L-1. \tag{801}$$

The limiting escape trajectory approaches the rank-zero saddle in backward time, so

$$\alpha_\ell(t) \longrightarrow 0 \qquad \text{as } t \to -\infty \tag{802}$$

for every layer. Hence every constant in (801) is zero, and therefore

$$|\alpha_1(t)| = \cdots = |\alpha_L(t)| = a(t) \tag{803}$$

for some $a(t) \ge 0$.

Since the end-to-end trajectory is oriented in the descending direction, its product is positive. Combining (799) and (803) therefore gives

$$m(t) = a(t)^L. \tag{804}$$

We can now derive the scalar end-to-end dynamics directly. Along the rank-one manifold,

$$\mathcal{C} = \frac{1}{2}(s_1 - m)^2 + \mathcal{C}_{\mathrm{const}}. \tag{805}$$

For each scalar factor, gradient flow gives

$$\dot{\alpha}_\ell = (s_1 - m) \prod_{j \ne \ell} \alpha_j. \tag{806}$$

Differentiating the product $m = \prod_{\ell=1}^{L} \alpha_\ell$, we obtain

$$\begin{aligned} \dot{m} &= \sum_{\ell=1}^{L} \left(\prod_{j \ne \ell} \alpha_j \right) \dot{\alpha}_\ell \\ &= (s_1 - m) \sum_{\ell=1}^{L} \left(\prod_{j \ne \ell} \alpha_j \right)^2. \end{aligned} \tag{807}$$

Using (803) and (804),

$$\left(\prod_{j \ne \ell} \alpha_j \right)^2 = a^{2L-2} = m^{2-2/L}, \tag{808}$$

and hence

$$\dot{m} = L\,(s_1 - m)\,m^{2-2/L}. \tag{809}$$

This proves (791).

Because the limiting trajectory is a nontrivial escape from the rank-zero saddle,

$$m(t) \longrightarrow 0 \qquad \text{as } t \to -\infty. \tag{810}$$

For $0 < m < s_1$, the right-hand side of (791) is strictly positive, so $m(t)$ is strictly increasing. It is bounded above by s_1: if the trajectory reached s_1 at a finite time, uniqueness of the scalar autonomous ODE would force it to remain identically equal to s_1 thereafter, whereas the escape trajectory lies in the nonstationary branch.

Thus $m(t)$ has a limit $m_\infty \in (0, s_1]$ as $t \to +\infty$. If $m_\infty < s_1$, then the right-hand side of (791) would converge to the strictly positive number

$$L\,(s_1 - m_\infty)\,m_\infty^{2-2/L}, \tag{811}$$

contradicting convergence of $m(t)$. Therefore

$$m(t) \longrightarrow s_1 \qquad \text{as } t \to +\infty. \tag{812}$$

This proves (792).

It remains to record a first-hitting-time form of the escape result. The construction in Jacot et al. (2021, Theorem 2 and Section 5.5) starts by fixing a sufficiently small radius around the origin and taking the escape time to be the first time at which the parameter trajectory reaches that radius. Hence there exists $\rho_0 > 0$ such that, for every fixed $\rho \in (0, \rho_0)$, the time $\tau_\rho^{(\sigma)}$ in (794) is finite for all sufficiently small σ on $\mathcal{E}_{\mathrm{esc}}$, and the corresponding first-hit parameters converge, up to the same hidden-layer orthogonal transformations, to the unique optimal width-one escape path at parameter norm ρ.

Along the balanced limiting trajectory, (803) gives

$$\|\boldsymbol{\theta}^*(t)\|_2^2 = \sum_{\ell=1}^{L} \|\mathbf{W}_\ell^*(t)\|_{\mathrm{F}}^2 = La(t)^2. \tag{813}$$

Thus the unique point of the limiting escape path with parameter norm ρ has

$$a = \frac{\rho}{\sqrt{L}}. \tag{814}$$

By (804), its end-to-end amplitude is therefore

$$m = a^L = \left(\frac{\rho}{\sqrt{L}}\right)^L. \tag{815}$$

Since the hidden-layer orthogonal transformations and the width-one embedding preserve the end-to-end matrix, convergence of the first-hit parameters implies

$$\mathbf{M}^{(\sigma)}\left(\tau_\rho^{(\sigma)}\right) \longrightarrow \left(\frac{\rho}{\sqrt{L}}\right)^L \mathbf{D}_1 \qquad \text{as } \sigma \downarrow 0. \tag{816}$$

This proves (795) and completes the proof. □

The rank-one escape immediately yields a classification-level degeneracy bottleneck. Recall the unit leading singular direction

$$\mathbf{D}_1 = \mathbf{p}_1 \mathbf{q}_1^\top, \tag{817}$$

so that the usual rank-one truncation is

$$\mathbf{T}_1 = s_1 \mathbf{D}_1. \tag{818}$$

Since $s_1 > 0$, the two matrices induce exactly the same classifier. Define

$$A_1 = \mathcal{A}(\mathbf{D}_1) = \mathcal{A}(\mathbf{T}_1). \tag{819}$$

Assume that, for every member of the large-system sequence, $\mathbf{D}_1$ has a unique top-scoring class on every training example. Writing

$$\widehat{c}_i^{(1)} = \arg\max_{c\in[C]} \left(\mathbf{D}_1\mathbf{x}_i\right)_c, \tag{820}$$

define its finite-problem minimum decision gap by

$$\delta_{1,H} = \min_i \min_{c\neq\widehat{c}_i^{(1)}} \left[\left(\mathbf{D}_1\mathbf{x}_i\right)_{\widehat{c}_i^{(1)}} - \left(\mathbf{D}_1\mathbf{x}_i\right)_c\right] > 0. \tag{821}$$

No uniform lower bound on $\delta_{1,H}$ is required, because the limit $\sigma \downarrow 0$ is taken at fixed H before the outer large-system limit.

Theorem B.27 (Predictive bottleneck for $L \geq 3$)**.** *Consider a sequence of finite problems satisfying* (784)*, and suppose that for every member of the sequence the unit leading direction $\mathbf{D}_1$ has a unique top-scoring class on every training example.*

Then there exist realization-dependent bottleneck times $\tau_{\mathrm{bot}}^{(\sigma,H)}$ such that, for every fixed $\varepsilon > 0$,

$$\begin{aligned}\liminf_{H\to\infty} \liminf_{\sigma\downarrow 0} \mathbb{P}\Bigg(&\mathcal{A}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{bot}}^{(\sigma,H)}\right)\right) = A_1, \\ &\sup_{0\leq\tau\leq\tau_{\mathrm{bot}}^{(\sigma,H)}} \left\|\mathbf{M}^{(\sigma)}(\tau)\right\|_{\mathrm{F}} < \varepsilon, \\ &\sup_{0\leq\tau\leq\tau_{\mathrm{bot}}^{(\sigma,H)}} \left|\mathcal{L}\left(\mathbf{M}^{(\sigma)}(\tau)\right) - \mathcal{L}(\mathbf{0})\right| < \varepsilon\Bigg) \geq \pi_L,\end{aligned} \tag{822}$$

where

$$\pi_L = \begin{cases} 1, & L = 3, \\ \frac{1}{2}, & L > 3. \end{cases} \tag{823}$$

Thus, with asymptotic probability one for $L = 3$, and with asymptotic probability at least $1/2$ for $L > 3$, the trajectory reaches a microscopic stage whose classifier is that of the leading singular mode, while the entire trajectory up to that stage remains microscopic and its loss remains asymptotically equal to the loss of the zero predictor.

Proof. For the H-th finite problem, let $\gamma = \gamma_H > 0$ denote the isotropy constant, let $s_1 = s_{1,H} > 0$ be the leading singular value, and let $\rho_{0,H} > 0$ be the local first-escape radius supplied by Proposition B.26. Choose the deterministic amplitude

$$a_H = \min\left\{\frac{s_1}{2}, \frac{1}{H}, \frac{1}{H\gamma(1+s_1)}, \left(\frac{\rho_{0,H}}{2\sqrt{L}}\right)^L\right\}. \tag{824}$$

Then

$$0 < a_H < s_1, \qquad a_H \longrightarrow 0, \tag{825}$$

and

$$\gamma a_H s_1 \leq \frac{1}{H}, \qquad \gamma a_H^2 \leq \frac{1}{H}. \tag{826}$$

Define the corresponding parameter-space radius by

$$\rho_H := \sqrt{L}\, a_H^{1/L}. \tag{827}$$

By construction,

$$0 < \rho_H < \rho_{0,H}. \tag{828}$$

On the escape event $\mathcal{E}_{\mathrm{esc}}$ of Proposition B.26, define

$$\tau_{\mathrm{bot}}^{(\sigma,H)} := \inf\left\{\tau \geq 0 : \left\|\boldsymbol{\theta}^{(\sigma)}(\tau)\right\|_2 = \rho_H\right\}. \tag{829}$$

On the complement $\mathcal{E}_{\mathrm{esc}}^c$, define $\tau_{\mathrm{bot}}^{(\sigma,H)}$ arbitrarily, for example to be zero.

For every fixed H, Proposition B.26 gives, on $\mathcal{E}_{\mathrm{esc}}$,

$$\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{bot}}^{(\sigma,H)}\right) \longrightarrow \left(\frac{\rho_H}{\sqrt{L}}\right)^L \mathbf{D}_1 = a_H \mathbf{D}_1 \qquad \text{as } \sigma \downarrow 0. \tag{830}$$

We first consider classification. Since $a_H > 0$, multiplication by a_H does not change score orderings. For every training example i and every $c \neq \widehat{c}_i^{(1)}$,

$$\begin{aligned}(a_H \mathbf{D}_1 \mathbf{x}_i)_{\widehat{c}_i^{(1)}} - (a_H \mathbf{D}_1 \mathbf{x}_i)_c \\ \geq a_H \delta_{1,H} > 0.\end{aligned} \tag{831}$$

For fixed H, there are only finitely many score comparisons. Therefore (830) implies that, on $\mathcal{E}_{\mathrm{esc}}$, for all sufficiently small σ,

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{bot}}^{(\sigma,H)}\right)\right) = A_1. \tag{832}$$

No uniform lower bound on $\delta_{1,H}$ is required.

We next prove microscopicity of the entire preceding trajectory. For every fixed H, the coupled initialization satisfies

$$\left\|\boldsymbol{\theta}^{(\sigma)}(0)\right\|_2 = \sigma\left(\sum_{\ell=1}^{L} \|\mathbf{G}_\ell\|_{\mathrm{F}}^2\right)^{1/2} \longrightarrow 0 \qquad \text{as } \sigma \downarrow 0. \tag{833}$$

Hence, almost surely, it lies inside the sphere of radius ρ_H for all sufficiently small σ. By the definition of the first hitting time,

$$\left\|\boldsymbol{\theta}^{(\sigma)}(\tau)\right\|_2 \leq \rho_H \qquad \text{for all } 0 \leq \tau \leq \tau_{\mathrm{bot}}^{(\sigma,H)}. \tag{834}$$

Using submultiplicativity and the arithmetic–geometric mean inequality,

$$\begin{aligned}\left\|\mathbf{M}^{(\sigma)}(\tau)\right\|_{\mathrm{F}} &\leq \prod_{\ell=1}^{L} \left\|\mathbf{W}_\ell^{(\sigma)}(\tau)\right\|_{\mathrm{F}} \\ &\leq \left(\frac{1}{L}\sum_{\ell=1}^{L} \left\|\mathbf{W}_\ell^{(\sigma)}(\tau)\right\|_{\mathrm{F}}^2\right)^{L/2} \\ &\leq \left(\frac{\rho_H}{\sqrt{L}}\right)^L = a_H.\end{aligned} \tag{835}$$

Consequently,

$$\sup_{0\leq\tau\leq\tau_{\mathrm{bot}}^{(\sigma,H)}} \left\|\mathbf{M}^{(\sigma)}(\tau)\right\|_{\mathrm{F}} \leq a_H \longrightarrow 0 \qquad \text{as } H \to \infty. \tag{836}$$

It remains to establish uniform flatness of the loss. Since

$$\|\mathbf{D}_1\|_{\mathrm{F}} = 1, \qquad \langle \mathbf{T}, \mathbf{D}_1 \rangle_{\mathrm{F}} = s_1, \tag{837}$$

we have

$$\begin{aligned}\mathcal{L}\left(a_H \mathbf{D}_1\right) - \mathcal{L}(\mathbf{0}) &= \frac{\gamma}{2}\left[\|\mathbf{T} - a_H \mathbf{D}_1\|_{\mathrm{F}}^2 - \|\mathbf{T}\|_{\mathrm{F}}^2\right] \\ &= \gamma\left(-a_H s_1 + \frac{1}{2} a_H^2\right).\end{aligned} \tag{838}$$

Therefore

$$\left|\mathcal{L}\left(a_H \mathbf{D}_1\right) - \mathcal{L}(\mathbf{0})\right| \leq \frac{3}{2H} \longrightarrow 0. \tag{839}$$

By continuity of the loss and (830), for every fixed H, on $\mathcal{E}_{\mathrm{esc}}$,

$$\mathcal{L}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{bot}}^{(\sigma,H)}\right)\right) \longrightarrow \mathcal{L}\left(a_H \mathbf{D}_1\right) \qquad \text{as } \sigma \downarrow 0. \tag{840}$$

Also,

$$\mathbf{M}^{(\sigma)}(0) = \sigma^L \mathbf{G}_L \cdots \mathbf{G}_1 \longrightarrow \mathbf{0} \qquad \text{as } \sigma \downarrow 0, \tag{841}$$

so

$$\mathcal{L}\left(\mathbf{M}^{(\sigma)}(0)\right) \longrightarrow \mathcal{L}(\mathbf{0}). \tag{842}$$

Gradient flow makes the loss nonincreasing. Hence, for every $0 \le \tau \le \tau_{\mathrm{bot}}^{(\sigma,H)}$,

$$\mathcal{L}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{bot}}^{(\sigma,H)}\right)\right) \le \mathcal{L}\left(\mathbf{M}^{(\sigma)}(\tau)\right) \le \mathcal{L}\left(\mathbf{M}^{(\sigma)}(0)\right). \tag{843}$$

It follows that

$$\begin{aligned} &\sup_{0\le\tau\le\tau_{\mathrm{bot}}^{(\sigma,H)}} \left|\mathcal{L}\left(\mathbf{M}^{(\sigma)}(\tau)\right) - \mathcal{L}(\mathbf{0})\right| \\ &\qquad \le \max\left\{\left|\mathcal{L}\left(\mathbf{M}^{(\sigma)}(0)\right) - \mathcal{L}(\mathbf{0})\right|, \left|\mathcal{L}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{bot}}^{(\sigma,H)}\right)\right) - \mathcal{L}(\mathbf{0})\right|\right\}. \end{aligned} \tag{844}$$

The right-hand side vanishes in the sequential limit $\sigma \downarrow 0$ followed by $H \to \infty$, by (839)–(842).

Finally, Proposition B.26 gives

$$\mathbb{P}(\mathcal{E}_{\mathrm{esc}}) \ge \pi_L. \tag{845}$$

On every realization in $\mathcal{E}_{\mathrm{esc}}$, the accuracy statement holds for all sufficiently small σ at fixed H, while (836) and (844) give the two uniform conclusions after the outer limit. This proves (822). □

The preceding two results can now be combined into a direct statement about the first two stages of the accuracy dynamics. The first stage is the target-driven organization established by Theorem B.25; the second is the rank-one predictive bottleneck generated by the first saddle escape. The only depth-dependent qualification comes from the probability with which the escape result of Jacot et al. (2021, Theorem 2) applies.

Theorem B.28 [4.5] **(Target-driven organization and predictive bottleneck for $L \ge 3$).** *Fix $L \ge 3$, and consider the coupled Gaussian initialization* (787)*. Suppose that the assumptions of Theorem B.25 hold and that*

$$\mathbf{T} = \sum_{k=1}^{r} s_k \mathbf{p}_k \mathbf{q}_k^\top, \qquad s_1 > s_2, \tag{846}$$

where $s_2 := 0$ when $r = 1$.

Let

$$\mathbf{D}_1 = \mathbf{p}_1 \mathbf{q}_1^\top, \qquad \mathbf{T}_1 = s_1 \mathbf{D}_1, \tag{847}$$

and suppose that, for every finite member of the sequence, $\mathbf{D}_1$ has a unique top-scoring class on every training example. Define

$$A_{\mathbf{T}} = \mathcal{A}(\mathbf{T}), \qquad A_1 = \mathcal{A}(\mathbf{D}_1) = \mathcal{A}(\mathbf{T}_1). \tag{848}$$

Let $\tau_{\mathrm{org}}^{(\sigma,H)}$ be the deterministic organization time from Theorem B.25. Then there exist realization-dependent bottleneck times $\tau_{\mathrm{bot}}^{(\sigma,H)}$ such that, for every fixed $\varepsilon > 0$,

$$\begin{aligned} \liminf_{H\to\infty} \liminf_{\sigma\downarrow 0} \mathbb{P}\Bigg(& 0 < \tau_{\mathrm{org}}^{(\sigma,H)} < \tau_{\mathrm{bot}}^{(\sigma,H)}, \\ & \mathcal{A}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right)\right) = A_{\mathbf{T}}, \\ & \mathcal{A}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{bot}}^{(\sigma,H)}\right)\right) = A_1, \\ & \sup_{0\le\tau\le\tau_{\mathrm{bot}}^{(\sigma,H)}} \left\|\mathbf{M}^{(\sigma)}(\tau)\right\|_{\mathrm{F}} < \varepsilon, \\ & \sup_{0\le\tau\le\tau_{\mathrm{bot}}^{(\sigma,H)}} \left|\mathcal{L}\left(\mathbf{M}^{(\sigma)}(\tau)\right) - \mathcal{L}(\mathbf{0})\right| < \varepsilon \Bigg) \ge \pi_L, \end{aligned} \tag{849}$$

where

$$\pi_L = \begin{cases} 1, & L = 3, \\ \frac{1}{2}, & L > 3. \end{cases} \tag{850}$$

Thus, whenever

$$A_{\mathbf{T}} > A_1, \tag{851}$$

the trajectory reaches an earlier stage with target accuracy $A_{\mathbf{T}}$ and a later bottleneck with the strictly smaller accuracy A_1, while the entire trajectory through the bottleneck remains microscopic and its loss remains asymptotically equal to that of the zero predictor. For $L = 3$, this conclusion holds with probability tending to one; for $L > 3$, its asymptotic probability is bounded below by $1/2$.

Proof. Write

$$\mathcal{L}_{\text{zero}} := \mathcal{L}(\mathbf{0}). \tag{852}$$

Let $\mathcal{E}_H^{\text{org}}$ be the event supplied by Theorem B.25. Then

$$\mathbb{P}\left(\mathcal{E}_H^{\text{org}}\right) \longrightarrow 1. \tag{853}$$

For every sufficiently large H, on $\mathcal{E}_H^{\text{org}}$, the theorem gives, for all sufficiently small σ,

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\text{org}}^{(\sigma,H)}\right)\right) = A_{\mathbf{T}}, \tag{854}$$

and

$$\mathcal{L}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\text{org}}^{(\sigma,H)}\right)\right) \longrightarrow \mathcal{L}_{\text{zero}} \qquad \text{as } \sigma \downarrow 0. \tag{855}$$

Moreover,

$$\tau_{\text{org}}^{(\sigma,H)} > 0. \tag{856}$$

Let $\mathcal{E}_{\text{esc},H}$ be the escape event from Proposition B.26, and take $\tau_{\text{bot}}^{(\sigma,H)}$ to be the first-hitting bottleneck time constructed in Theorem B.27. Then

$$\mathbb{P}\left(\mathcal{E}_{\text{esc},H}\right) \geq \pi_L. \tag{857}$$

Let a_H be the deterministic bottleneck amplitude used in the proof of Theorem B.27. On $\mathcal{E}_{\text{esc},H}$, for every fixed H,

$$\mathbf{M}^{(\sigma)}\left(\tau_{\text{bot}}^{(\sigma,H)}\right) \longrightarrow a_H \mathbf{D}_1 \qquad \text{as } \sigma \downarrow 0. \tag{858}$$

Therefore, for all sufficiently small σ,

$$\mathcal{A}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\text{bot}}^{(\sigma,H)}\right)\right) = A_1. \tag{859}$$

It remains to verify that organization occurs before the bottleneck on the simultaneous good event. For every fixed H, define

$$d_H := \mathcal{L}_{\text{zero}} - \mathcal{L}\left(a_H \mathbf{D}_1\right) = \gamma\left(a_H s_1 - \frac{1}{2} a_H^2\right). \tag{860}$$

Since $0 < a_H < s_1$,

$$d_H > 0. \tag{861}$$

By (858) and continuity of the loss, on $\mathcal{E}_{\text{esc},H}$,

$$\mathcal{L}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\text{bot}}^{(\sigma,H)}\right)\right) \longrightarrow \mathcal{L}_{\text{zero}} - d_H \qquad \text{as } \sigma \downarrow 0. \tag{862}$$

Now fix an initialization in

$$\mathcal{E}_H^{\text{joint}} := \mathcal{E}_H^{\text{org}} \cap \mathcal{E}_{\text{esc},H}. \tag{863}$$

For all sufficiently small σ, (862) gives

$$\mathcal{L}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\text{bot}}^{(\sigma,H)}\right)\right) < \mathcal{L}_{\text{zero}} - \frac{3 d_H}{4}, \tag{864}$$

whereas (855) gives

$$\mathcal{L}\left(\mathbf{M}^{(\sigma)}\left(\tau_{\mathrm{org}}^{(\sigma,H)}\right)\right) > \mathcal{L}_{\mathrm{zero}} - \frac{d_H}{4}. \tag{865}$$

Since gradient flow makes the loss nonincreasing, these inequalities force

$$0 < \tau_{\mathrm{org}}^{(\sigma,H)} < \tau_{\mathrm{bot}}^{(\sigma,H)} \tag{866}$$

for all sufficiently small σ on $\mathcal{E}_H^{\mathrm{joint}}$.

The uniform microscopicity and flat-loss conclusions on $[0, \tau_{\mathrm{bot}}^{(\sigma,H)}]$ follow directly from Theorem B.27. In particular, since $a_H \to 0$, for every fixed $\varepsilon > 0$ they hold for all sufficiently large H after taking the inner $\sigma \downarrow 0$ limit.

Finally,

$$\begin{aligned}\mathbb{P}\left(\mathcal{E}_H^{\mathrm{joint}}\right) &= \mathbb{P}\left(\mathcal{E}_H^{\mathrm{org}} \cap \mathcal{E}_{\mathrm{esc},H}\right) \\ &\geq \mathbb{P}\left(\mathcal{E}_{\mathrm{esc},H}\right) - \mathbb{P}\left((\mathcal{E}_H^{\mathrm{org}})^c\right) \\ &\geq \pi_L - o(1).\end{aligned} \tag{867}$$

On $\mathcal{E}_H^{\mathrm{joint}}$, the two accuracy statements, the time ordering, and the uniform microscopicity and flat-loss conclusions all hold for sufficiently small σ. Taking the sequential limit proves (849). □

Recovery: open generically, explicit under spectral alignment. Theorem B.28 establishes the first two stages of the accuracy dynamics for Gaussian initialization: early target-driven organization followed by the rank-one predictive bottleneck. It does not establish recovery of the same Gaussian-initialized trajectory. For $L \geq 3$, global convergence of deep matrix factorization from the multivariate vanishing-Gaussian initialization considered here is not known at the level of generality required for such a conclusion.

For gradient flow, Bah et al. (2022) prove convergence to a critical point and show that, for almost every initialization, the limiting end-to-end matrix is a global minimizer of the loss restricted to the manifold of matrices of some rank k. For $L = 2$, the limiting rank can be shown to be maximal, which yields global convergence. For $L \geq 3$, however, their argument does not determine the limiting rank, and therefore does not establish convergence to the unrestricted global minimizer. An analogous limitation remains for gradient descent (Nguegnang et al., 2024).

Global convergence is known under stronger assumptions. For example, Arora et al. (2019) obtain it under approximate balance together with an initial-loss condition excluding rank-deficient solutions, while the near-zero random-initialization result of Elkabetz & Cohen (2021) applies to scalar output. These results do not cover the multivariate vanishing-Gaussian regime considered here, which starts near the highly degenerate rank-zero saddle.

The missing recovery mechanism is closely related to the saddle-to-saddle picture proposed by Jacot et al. (2021). After the rigorously established first escape analyzed above, they conjecture that the vanishing-initialization trajectory successively visits critical points of increasing rank,

$$0 \longrightarrow 1 \longrightarrow 2 \longrightarrow \cdots, \tag{868}$$

eventually reaching a global minimizer. In the present interpretation, such subsequent escapes would correspond to recruitment of singular modes that are absent from the rank-one bottleneck, and would provide the missing recovery mechanism for the Gaussian trajectory. This increasing-rank sequence, however, is not currently established in the required generality.

The mechanism can nevertheless be followed exactly in the classical spectrally aligned setting of Saxe et al. (2014). Let

$$\mathbf{T} = \sum_{k=1}^{r} s_k \mathbf{p}_k \mathbf{q}_k^\top, \qquad s_k > 0, \tag{869}$$

and define

$$\mathbf{P} = (\mathbf{p}_1, \ldots, \mathbf{p}_r), \qquad \mathbf{Q} = (\mathbf{q}_1, \ldots, \mathbf{q}_r). \tag{870}$$

For $H \geq r$, choose $\mathbf{B} \in \mathbb{R}^{H \times r}$ with

$$\mathbf{B}^\top \mathbf{B} = \mathbf{I}_r, \tag{871}$$

and initialize the factors on the balanced decoupled manifold

$$\mathbf{W}_1(0) = \mathbf{B}\mathbf{A}(0)\mathbf{Q}^\top, \tag{872}$$

$$\mathbf{W}_\ell(0) = \mathbf{B}\mathbf{A}(0)\mathbf{B}^\top, \qquad 2 \le \ell \le L-1, \tag{873}$$

and

$$\mathbf{W}_L(0) = \mathbf{P}\mathbf{A}(0)\mathbf{B}^\top, \tag{874}$$

where

$$\mathbf{A}(0) = \operatorname{diag}\left(a_1(0), \dots, a_r(0)\right), \qquad a_k(0) > 0. \tag{875}$$

This aligned manifold is invariant under gradient flow (Saxe et al., 2014). The singular directions therefore remain fixed, and the end-to-end predictor has the form

$$\mathbf{M}(\tau) = \sum_{k=1}^{r} m_k(\tau)\mathbf{p}_k\mathbf{q}_k^\top, \qquad m_k(\tau) = a_k(\tau)^L. \tag{876}$$

In our normalization, the modes evolve independently according to

$$\dot{m}_k = L\left(s_k - m_k\right) m_k^{2-2/L}. \tag{877}$$

Equation (877) makes the mechanism of the bottleneck and subsequent recovery explicit. The relative growth rate of mode k is

$$\frac{d}{d\tau}\log m_k = L\left(s_k - m_k\right) m_k^{(L-2)/L}. \tag{878}$$

For $L \ge 3$, modes that are already larger receive a multiplicative growth advantage near the origin, while larger target singular values provide an additional advantage through the factor $s_k - m_k$. The leading mode can therefore separate strongly from the remaining modes and temporarily dominate the predictor, producing the aligned analogue of the rank-one degeneracy bottleneck.

This dominance cannot persist indefinitely. As m_1 approaches its target value s_1, its residual $s_1 - m_1$ vanishes and its relative growth rate tends to zero. A lagging mode that remains far below its target value still has a nonvanishing residual and continues to grow. Thus the leading mode saturates while the remaining modes are progressively recruited. In fact, every positive mode converges to its target value, so the full target is eventually reconstructed.

Corollary B.29 (Permanent classification recovery in the aligned regime)**.** *Consider the aligned balanced initialization* (872)–(874)*, with $a_k(0) > 0$ for every k. Suppose that both $\mathbf{T}$ and its leading singular component*

$$\mathbf{T}_1 = s_1\mathbf{p}_1\mathbf{q}_1^\top \tag{879}$$

have unique top-scoring classes on every training example. Define

$$A_{\mathbf{T}} = \mathcal{A}(\mathbf{T}), \qquad A_1 = \mathcal{A}(\mathbf{T}_1), \tag{880}$$

let $\delta_1 > 0$ be the minimum decision gap of $\mathbf{T}_1$, and set

$$R = \max_i \|\mathbf{x}_i\|_2. \tag{881}$$

If, at some time τ_{bot},

$$\max_{k\ge 2} m_k(\tau_{\mathrm{bot}}) < \frac{\delta_1}{2\sqrt{2}\,R\,s_1} m_1(\tau_{\mathrm{bot}}), \tag{882}$$

with the maximum interpreted as 0 when $r = 1$, then

$$\mathcal{A}\left(\mathbf{M}(\tau_{\mathrm{bot}})\right) = A_1. \tag{883}$$

Moreover,

$$\mathbf{M}(\tau) \longrightarrow \mathbf{T} \qquad \text{as } \tau \to \infty, \tag{884}$$

and consequently there exists a finite $\tau_{\mathrm{rec}} > \tau_{\mathrm{bot}}$ such that

$$\mathcal{A}\left(\mathbf{M}(\tau)\right) = A_{\mathbf{T}} \qquad \text{for every } \tau \ge \tau_{\mathrm{rec}}. \tag{885}$$

Thus, whenever

$$A_{\mathbf{T}} > A_1, \tag{886}$$

an aligned trajectory that enters the leading-mode predictive bottleneck subsequently recovers permanently to the target accuracy.

Proof. By invariance of the aligned manifold,

$$\mathbf{M}(\tau) = \sum_{k=1}^{r} m_k(\tau)\mathbf{p}_k\mathbf{q}_k^\top, \tag{887}$$

and the mode amplitudes satisfy (877).

At $\tau = \tau_{\mathrm{bot}}$, write

$$\mathbf{M}(\tau_{\mathrm{bot}}) = \frac{m_1(\tau_{\mathrm{bot}})}{s_1}\mathbf{T}_1 + \mathbf{E}, \tag{888}$$

where

$$\mathbf{E} = \sum_{k=2}^{r} m_k(\tau_{\mathrm{bot}})\mathbf{p}_k\mathbf{q}_k^\top. \tag{889}$$

Since the singular directions are orthonormal,

$$\|\mathbf{E}\|_{\mathrm{op}} = \max_{k\geq 2} m_k(\tau_{\mathrm{bot}}), \tag{890}$$

with the right-hand side equal to 0 when $r = 1$.

For each example, let

$$\widehat{c}_i^{(1)} = \arg\max_c \left(\mathbf{T}_1\mathbf{x}_i\right)_c. \tag{891}$$

For any $c \neq \widehat{c}_i^{(1)}$,

$$\begin{aligned}
&\left(\mathbf{e}_{\widehat{c}_i^{(1)}} - \mathbf{e}_c\right)^\top \mathbf{M}(\tau_{\mathrm{bot}})\mathbf{x}_i \\
&\quad= \frac{m_1(\tau_{\mathrm{bot}})}{s_1}\left(\mathbf{e}_{\widehat{c}_i^{(1)}} - \mathbf{e}_c\right)^\top \mathbf{T}_1\mathbf{x}_i + \left(\mathbf{e}_{\widehat{c}_i^{(1)}} - \mathbf{e}_c\right)^\top \mathbf{E}\mathbf{x}_i \\
&\quad\geq \frac{m_1(\tau_{\mathrm{bot}})}{s_1}\delta_1 - \sqrt{2}\,R\|\mathbf{E}\|_{\mathrm{op}} \\
&\quad> \frac{m_1(\tau_{\mathrm{bot}})}{2s_1}\delta_1 > 0,
\end{aligned} \tag{892}$$

where the penultimate inequality follows from (882). Thus the classifier at τ_{bot} agrees with that of $\mathbf{T}_1$, proving (883).

It remains to prove recovery. For each k, positivity of $m_k(0)$ is preserved by (877). Moreover,

$$\operatorname{sign}(\dot{m}_k) = \operatorname{sign}(s_k - m_k). \tag{893}$$

Hence each $m_k(\tau)$ moves monotonically toward s_k. It cannot cross s_k, because s_k is an equilibrium of the scalar autonomous ODE and uniqueness holds for positive m_k. Therefore $m_k(\tau)$ has a finite positive limit $m_{k,\infty}$.

If $m_{k,\infty} \neq s_k$, then continuity of the right-hand side of (877) would imply that $\dot{m}_k$ remains bounded away from zero for all sufficiently large τ, contradicting convergence of $m_k(\tau)$. Hence

$$m_{k,\infty} = s_k \qquad \text{for every } k. \tag{894}$$

It follows that

$$\begin{aligned}
\mathbf{M}(\tau) &= \sum_{k=1}^{r} m_k(\tau)\mathbf{p}_k\mathbf{q}_k^\top \\
&\longrightarrow \sum_{k=1}^{r} s_k\mathbf{p}_k\mathbf{q}_k^\top = \mathbf{T},
\end{aligned} \tag{895}$$

which proves (884).

Finally, since $\mathbf{T}$ has a unique top-scoring class on every training example and the training set is finite, its minimum decision gap

$$\delta_{\mathbf{T}} = \min_i \min_{c\neq\widehat{c}_i^{\,T}} \left[\left(\mathbf{T}\mathbf{x}_i\right)_{\widehat{c}_i^{\,T}} - \left(\mathbf{T}\mathbf{x}_i\right)_c\right] \tag{896}$$

is strictly positive, where

$$\widehat{c}_i^{\,T} = \arg\max_c \left(\mathbf{T}\mathbf{x}_i\right)_c . \tag{897}$$

Whenever

$$\|\mathbf{M}(\tau) - \mathbf{T}\|_{\mathrm{op}} < \frac{\delta_{\mathbf{T}}}{2\sqrt{2}\,R}, \tag{898}$$

every target score margin remains positive. Since $\mathbf{M}(\tau) \to \mathbf{T}$, this inequality holds for all sufficiently large τ. Therefore there exists a finite τ_{rec} after which the classifier agrees with that of $\mathbf{T}$ permanently, proving (885). □

The aligned result should not be interpreted as completing Theorem B.28 for the same Gaussian-initialized trajectory. Rather, it isolates the recruitment and recovery mechanism in a setting where the individual singular modes can be followed exactly. The resulting picture is consistent with the saddle-to-saddle scenario of Jacot et al. (2021): depth first amplifies differences between weak modes and produces a low-rank predictive bottleneck; once the leading mode approaches its target strength, its residual gradient collapses, while the missing modes remain driven and are progressively recruited. Their eventual growth restores the full target and, with it, the target classifier.

C Phenomenology and Experimental Details

This appendix provides the experimental details and additional controls underlying the phenomenological observations of Section 2. We organize the experiments from the most controlled setting to progressively more conventional learning problems. Appendix C.1 studies unary-operation learning, where the target structure and the factorization can be varied directly. Appendix C.2 then moves to MNIST and tests the dependence of the well on initialization scale, width, loss, biases, factorization depth, and optimization method. Finally, Appendix C.3 considers a convolutional network on CIFAR-10 and compares both training and test dynamics.

Figures 2 and 3 show representative examples in the main text. The figures below provide the corresponding broader sweeps and controls. Our aim here is therefore not to repeat the phenomenological discussion of Section 2, but to document the experimental setups and make clear which comparisons support the empirical conclusions used in the paper.

C.1 Controlled unary-operation experiments

The unary setting provides the cleanest controlled experiments because the target geometry and the factorization can be modified independently. We consider three settings. First, a shallow factorization trained on a permutation target gives a flat-spectrum control for which the rigorous theory predicts no well. Second, we keep the same permutation target but use a three-factor parameterization, testing whether a well can occur without target spectral heterogeneity. Third, we return to the shallow factorization but replace the permutation by an iid random unary target, restoring the heterogeneous target spectrum analyzed in Appendix B.2. For each setting we vary both initialization scale and hidden width.

Setup, optimization, and measurement. Unless stated otherwise, our unary experiments use $p = 512$ inputs and $p = 512$ output classes and train on the complete operation table. There is therefore no train–test split and no minibatch sampling. We optimize

$$\mathcal{L}_{\mathbf{T}}(\mathbf{M}) = \frac{1}{p}\|\mathbf{M} - \mathbf{T}\|_{\mathrm{F}}^2. \tag{899}$$

Full-batch gradient descent is implemented using `torch.optim.SGD` with zero momentum and zero weight decay.

The implementation stores examples as columns: for input i,

$$\mathbf{T}_{ci} = \mathbf{1}\{f(i) = c\}. \tag{900}$$

This is the same target convention as in Section 4. The prediction for input i is obtained by taking the largest entry of column i.

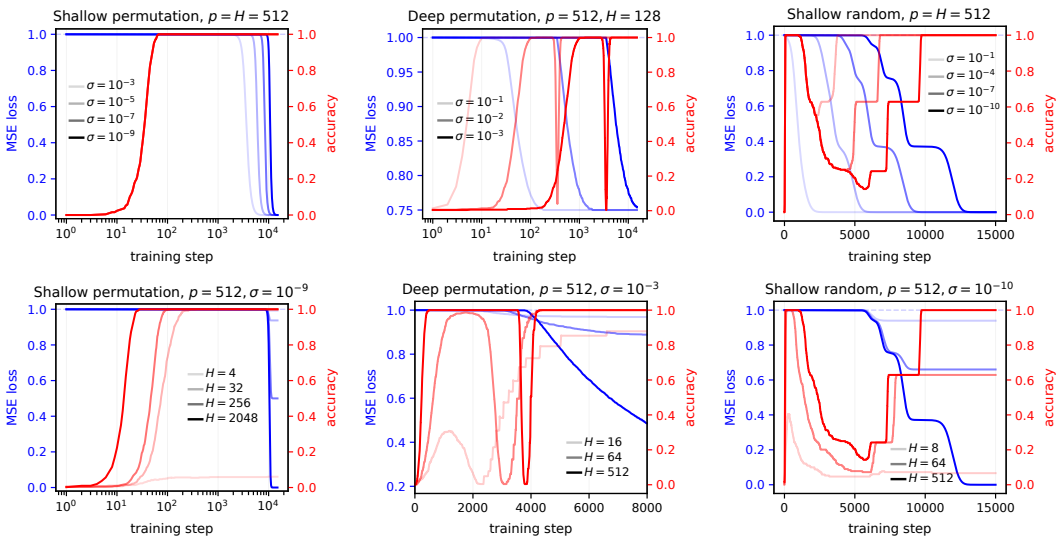

Figure 5: **Controlled unary-operation experiments.** Columns correspond to the shallow permutation model (left), the three-factor permutation model (middle), and the shallow model on an iid random unary target (right). The top row varies the initialization scale σ; the bottom row varies the hidden width H. Darker curves denote smaller σ in the top row and larger H in the bottom row. Blue curves show full-population MSE and red curves show training accuracy. The shallow permutation model reaches high accuracy while the loss remains near its initial value, without developing a well. Changing to the three-factor parameterization produces a pronounced well for the same flat-spectrum target, while a heterogeneous random unary target produces a well already in the shallow model, consistently with the heterogeneous-spectrum mechanism analyzed theoretically.

Every unary run contains $15{,}000$ parameter updates, with loss and top-one accuracy recorded throughout optimization.

The shallow permutation and shallow random-unary experiments use learning rate

$$\eta_{\text{shallow}} = 0.5, \tag{901}$$

whereas the three-factor permutation experiments use

$$\eta_{\text{deep}} = 50. \tag{902}$$

The learning rates were chosen separately for the two parameterizations. Accordingly, the shallow-versus-deep permutation comparison should be read as showing that the flat-spectrum target can exhibit a well under a deeper factorized optimization dynamics, rather than as a matched learning-rate depth ablation.

Initialization. We use fan-in-normalized Gaussian initialization multiplied by a global scale σ. For the shallow model

$$\mathbf{M} = \mathbf{U}^\top \mathbf{V}, \qquad \mathbf{U}, \mathbf{V} \in \mathbb{R}^{H \times p}, \tag{903}$$

the entries are initialized as

$$\mathbf{V}_{hi}(0) \sim \mathcal{N}\left(0, \frac{\sigma^2}{p}\right), \qquad \mathbf{U}_{hc}(0) \sim \mathcal{N}\left(0, \frac{\sigma^2}{H}\right). \tag{904}$$

For the deeper model

$$\mathbf{M} = \mathbf{U}^\top \mathbf{W} \mathbf{V}, \qquad \mathbf{W} \in \mathbb{R}^{H \times H}, \tag{905}$$

we additionally use

$$\mathbf{W}_{hk}(0) \sim \mathcal{N}\left(0, \frac{\sigma^2}{H}\right), \tag{906}$$

while $\mathbf{U}$ and $\mathbf{V}$ retain the scalings in (904). There is no additional factor of $1/H$ in the forward map.

This normalization is useful for the width sweeps because the leading scale of the initial predictor is independent of H:

$$\mathbb{E}[\mathbf{M}_{ci}(0)^2] = \frac{\sigma^4}{p} \qquad \text{for } \mathbf{M} = \mathbf{U}^\top \mathbf{V}, \tag{907}$$

and

$$\mathbb{E}[\mathbf{M}_{ci}(0)^2] = \frac{\sigma^6}{p} \qquad \text{for } \mathbf{M} = \mathbf{U}^\top \mathbf{W}\mathbf{V}. \tag{908}$$

Thus varying H changes the latent dimension without simultaneously introducing an explicit growth of the initial logit scale.

This initialization differs from the unnormalized Gaussian initialization used in the rigorous unary theory. The experiments should therefore be viewed as finite-width qualitative tests of the same phenomenology rather than direct finite-p evaluations of the asymptotic formulas.

Sweep parameters and random seeds. The same permutation target is used in all permutation experiments. It is generated once using PyTorch seed 0. For the shallow permutation width sweep we fix $\sigma = 10^{-9}$ and use $H \in \{4, 32, 256, 2048\}$, with initialization seed 1. For the corresponding initialization sweep we fix $H = 512$, use $\sigma \in \{10^{-3}, 10^{-5}, 10^{-7}, 10^{-9}\}$, and use initialization seed 100. The same seed is reset for every value of σ, so the same underlying Gaussian tensors are rescaled rather than independently resampled.

For the three-factor permutation model, the width sweep fixes $\sigma = 10^{-3}$, uses $H \in \{16, 64, 512\}$, and uses initialization seed 200. The initialization sweep fixes $H = 128$, uses $\sigma \in \{10^{-1}, 10^{-2}, 10^{-3}\}$, and uses initialization seed 300, again rescaling the same underlying Gaussian realization across the sweep. All of these runs contain $15{,}000$ updates; for visual clarity, the plotted three-factor width panel displays the first $8{,}000$ steps.

For the random-unary experiments, the target is sampled independently as

$$f(i) \overset{\text{iid}}{\sim} \mathrm{Unif}([p]) \tag{909}$$

using target seed 1000. The particular realization used in the figures has exactly 328 occupied output classes, and hence $\mathrm{rank}(T) = 328$. The same target is used in both random-unary sweeps. The initialization seed is 2000 throughout. The initialization sweep fixes $H = 512$ and uses $\sigma \in \{10^{-1}, 10^{-4}, 10^{-7}, 10^{-10}\}$; the width sweep fixes $\sigma = 10^{-10}$ and uses $H \in \{8, 64, 512\}$.

Shallow permutation: microscopic classification without a well. The left column of Figure 5 gives the flat-spectrum control. Decreasing σ strongly postpones the macroscopic decrease of the MSE, while the early classification transition changes much less. For sufficiently small initialization there is therefore a long interval in which accuracy is already perfect while the loss remains close to its zero-predictor value. Nevertheless, none of the displayed shallow permutation trajectories develops an accuracy well: once perfect classification is reached, it is retained. This is the finite-width analogue of the behavior proved in Theorem 4.4.

The width sweep also illustrates the distinction between matrix fitting and classification. The permutation target has rank 512, whereas the predictor $\mathbf{U}^\top \mathbf{V}$ has rank at most H. Hence the $H = 4, 32$, and 256 runs cannot fit the target matrix exactly, but can nevertheless attain perfect top-one accuracy (except for the $H = 4$ run). Increasing width makes the early classifier form more rapidly and eventually removes the rank obstruction to exact matrix fitting.

Three-factor permutation: a well without target spectral heterogeneity. The middle column keeps exactly the same permutation target but changes the parameterization to $\mathbf{M} = \mathbf{U}^\top \mathbf{W}\mathbf{V}$. The behavior is qualitatively different. At the larger initialization, accuracy evolves comparatively monotonically, whereas sufficiently small initialization reveals a pronounced high–low–high trajectory. The width sweep shows the complementary dependence on latent dimension: wider models develop a stronger early classifier and a more pronounced subsequent nonmonotonicity.

Because the permutation target has a flat positive spectrum, the heterogeneous-spectrum interference mechanism proved in Appendix B.2 cannot operate here in its original form. The experiment therefore demonstrates that target spectral heterogeneity is not necessary for an accuracy well once the parameterization is changed. This is qualitatively consistent with the broader bottleneck picture of Section 3.

Random unary target: heterogeneous spectra in the shallow model. The right column returns to $\mathbf{M} = \mathbf{U}^\top \mathbf{V}$ but replaces the permutation by the random target (909). Sufficiently small initialization produces a clear early accuracy hill followed by a substantial well. In the realizable $H = 512$ runs, accuracy later recovers to one. The stagewise form of the trajectory is qualitatively consistent with the occupancy-dependent spectral dynamics analyzed in Appendix B.2.

The width sweep also exposes behavior outside the realizable regime. Since the realized target has rank 328, the $H = 8$ and $H = 64$ runs cannot fit it exactly. These narrow networks reach very high accuracy during the early stage but settle at substantially lower accuracy after much longer MSE optimization. Thus reduction of the optimized Frobenius error can destroy class-ordering information that supported a better intermediate classifier. The $H = 512$ model is realizable and eventually recovers perfect accuracy.

Evolution of the score matrix. To visualize the mechanism underlying the random-unary well more directly, Figure 4 (left) repeats the shallow experiment in a small system with $p = 10$, $H = 50$, and $\sigma = 10^{-9}$, using full-batch gradient descent with learning rate 0.1. The output labels are permuted in decreasing order of target occupancy, and the inputs are grouped by their target output for visualization. At selected points along the trajectory we display the end-to-end score matrix

$$\mathbf{M}(t) = \mathbf{U}(t)^\top \mathbf{V}(t), \tag{910}$$

normalized separately at each time by

$$\widetilde{\mathbf{M}}(t) = \frac{\mathbf{M}(t)}{\max_{c,i} |\mathbf{M}_{ci}(t)|}. \tag{911}$$

Squares indicate the target output and dots the current prediction for each input. The snapshots include three states with the same accuracy $A = 0.6$, taken respectively before the early peak, during the subsequent decline, and during recovery. They show that equal classification accuracy can correspond to qualitatively different score-matrix structures at different stages of the well. The last four snapshots use a symmetric logarithmic color scale to make weaker score components visible.

Taken together, the unary experiments provide two complementary departures from the shallow permutation control. A well appears after changing the factorized dynamics while keeping the target spectrum flat, and it also appears after changing to a heterogeneous target while keeping the shallow factorization. Across these settings, decreasing initialization separates the microscopic accuracy dynamics from macroscopic loss decay, while increasing width generally strengthens the early classifier.

C.2 Accuracy wells in fully connected networks on MNIST

The unary experiments allow unusually direct control of the target geometry and factorization. We next ask which aspects of the accuracy-well phenomenology persist in a standard multiclass classification problem. We use MNIST to examine the effects of initialization scale, width, loss, architectural biases, factorization depth, optimization method, and random seed. We additionally use saved checkpoints from one fixed trajectory to probe how the participation of individual hidden neurons changes through the accuracy hill, well, and recovery.

Dataset and common training protocol. We use the standard MNIST training set of $60{,}000$ grayscale 28×28 images from ten classes. Pixel values are divided by 255 and each image is flattened to

$$x \in \mathbb{R}^{784}. \tag{912}$$

We apply no additional centering, normalization, or data augmentation. The MNIST test set is not used for optimization or model selection. It is used only for the post-hoc checkpoint diagnostics described below.

Unless stated otherwise, models are trained for $10{,}000$ minibatch updates with batch size

$$B = 1024, \tag{913}$$

using plain SGD with learning rate

$$\eta = 0.005, \tag{914}$$

zero momentum, zero weight decay, and no learning-rate schedule. Minibatches are obtained from random permutations of the full training set. Consecutive blocks of 1024 examples are used; when fewer than 1024 examples remain, the remainder is discarded and a new random permutation is generated.

Loss and top-one accuracy are evaluated on the current minibatch immediately before the corresponding parameter update. Unless explicitly identified as test accuracy, the curves therefore report minibatch training quantities.

The random streams are controlled across experimental conditions. For the main sweeps, the global PyTorch and NumPy seed is 1, the default weight-initialization seed is 1234, and the minibatch-permutation seed is 5678. The minibatch generator is reset for every run, so different experimental conditions see the same minibatch sequence. Within an initialization-scale sweep of a fixed architecture, the initialization generator is also reset for every value of σ, so that the same underlying Gaussian realization is merely rescaled. The experiments were run in single precision on an NVIDIA Tesla P100-PCIE-16GB GPU.

ReLU model, initialization, and losses. Our nonlinear model is a one-hidden-layer ReLU network

$$f_\theta(x) = W_2\ \mathrm{ReLU}(W_1 x + b_1) + b_2, \qquad W_1 \in \mathbb{R}^{H\times 784}, \qquad W_2 \in \mathbb{R}^{10\times H}. \tag{915}$$

Depending on the experiment, either bias, both biases, or neither bias is included. Whenever present, a bias is trainable and initialized at zero.

The weights use fan-in-normalized Gaussian initialization,

$$(W_1)_{hi}(0) \sim \mathcal{N}\left(0, \frac{\sigma^2}{784}\right), \qquad (W_2)_{ch}(0) \sim \mathcal{N}\left(0, \frac{\sigma^2}{H}\right). \tag{916}$$

For the squared-loss experiments, the target for an example of class y is the one-hot vector e_y, and we optimize

$$\mathcal{L}_{\mathrm{MSE}} = \frac{1}{B}\sum_{i=1}^{B} \|f_\theta(x_i) - e_{y_i}\|_2^2 . \tag{917}$$

With this normalization, the zero predictor has loss one.

For the cross-entropy experiments we use

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{B}\sum_{i=1}^{B} \log \frac{\exp f_{\theta,y_i}(x_i)}{\sum_{c=1}^{10} \exp f_{\theta,c}(x_i)}. \tag{918}$$

The corresponding zero-logit value is $\log 10$.

Architectural sensitivity: hidden and output biases. We first use squared loss to test how the location of a trainable bias affects the phenomenon. We compare four architectures:

Architecture	hidden bias b_1	output bias b_2
No biases	absent	absent
Hidden bias only	present	absent
Output bias only	absent	present
Hidden and output biases	present	present

Apart from the specified biases, the architectures use the same weight initialization, loss, minibatch sequence, and optimization protocol.

For each architecture we perform an initialization-scale sweep with

$$H = 8192, \qquad \sigma \in \{10^{-1}, 10^{-2}, 10^{-4}\}, \tag{919}$$

and a width sweep with

$$\sigma = 10^{-4}, \qquad H \in \{32, 512, 8192\}. \tag{920}$$

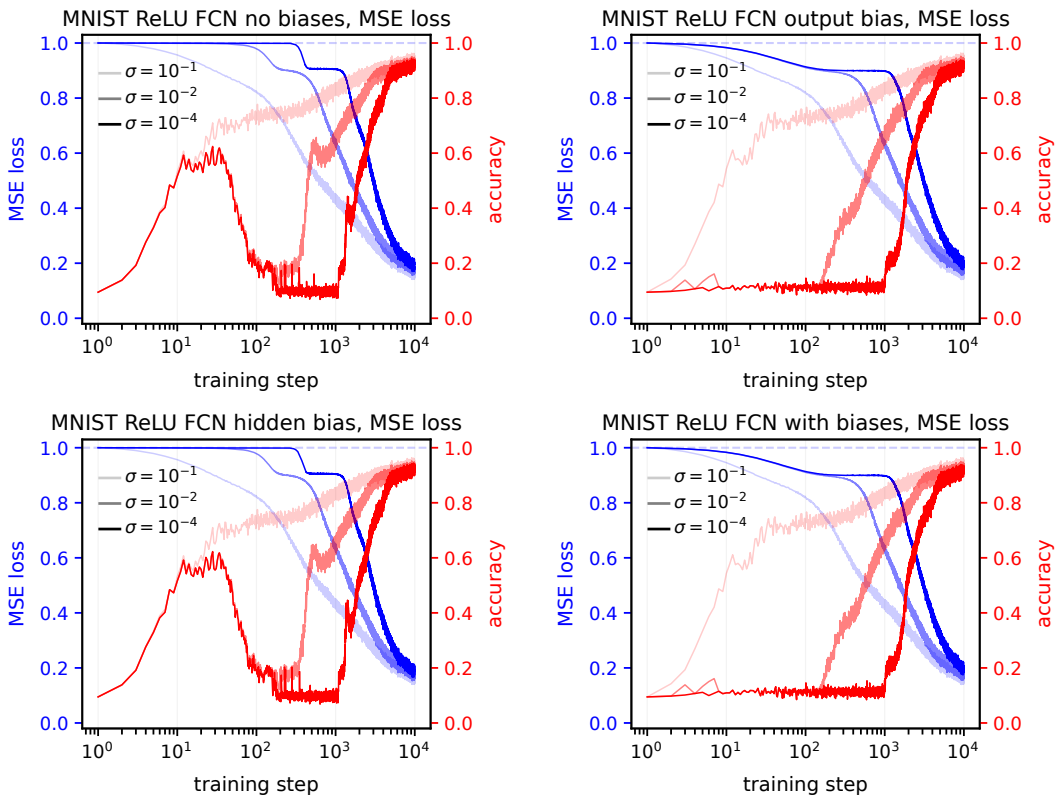

Figure 6: **Initialization-scale sweeps for the MNIST ReLU bias ablation under squared loss.** Columns distinguish the absence (left) and presence (right) of an output bias, while rows distinguish the absence (top) and presence (bottom) of a hidden bias. All panels use $H = 8192$ and $\sigma \in \{10^{-1}, 10^{-2}, 10^{-4}\}$, with darker curves denoting smaller initialization. Blue curves show minibatch MSE and red curves show minibatch top-one accuracy. Without an output bias, sufficiently small initialization produces a pronounced accuracy well regardless of whether a hidden bias is present. Introducing an output bias suppresses the pronounced well in both corresponding architectures.

The full sweeps are shown in Figures 6 and 7. A direct four-way comparison at $H = 8192$ and $\sigma = 10^{-4}$ is shown in the right panel of Figure 9.

The distinction between hidden and output biases is robust across the two sweeps. Adding a hidden bias leaves the characteristic early hill and subsequent drop largely intact. A directly trainable class-wise output bias, by contrast, suppresses the early hill and removes the pronounced well. Adding a hidden bias together with the output bias does not restore the effect. Thus the relevant architectural observation is not that biases in general remove the well; rather, the location at which a bias enters the predictor matters.

The initialization sweeps also reproduce the strong dependence on initialization scale seen in the controlled unary experiments. For the architectures that exhibit a well, decreasing σ makes substantial accuracy evolution visible while the MSE remains close to its zero-predictor baseline. The width sweeps show that the strength and timing of the early classifier also depend on capacity.

Changing the objective: cross-entropy. We next keep the bias-free ReLU architecture and replace squared loss by cross-entropy. The training protocol remains 10,000 SGD updates with $\eta = 0.005$ and batch size 1024. The initialization sweep fixes

$$H = 8192, \qquad \sigma \in \{10^{-1}, 10^{-2}, 10^{-4}\}, \tag{921}$$

while the width sweep fixes

$$\sigma = 10^{-4}, \qquad H \in \{32, 512, 8192\}. \tag{922}$$

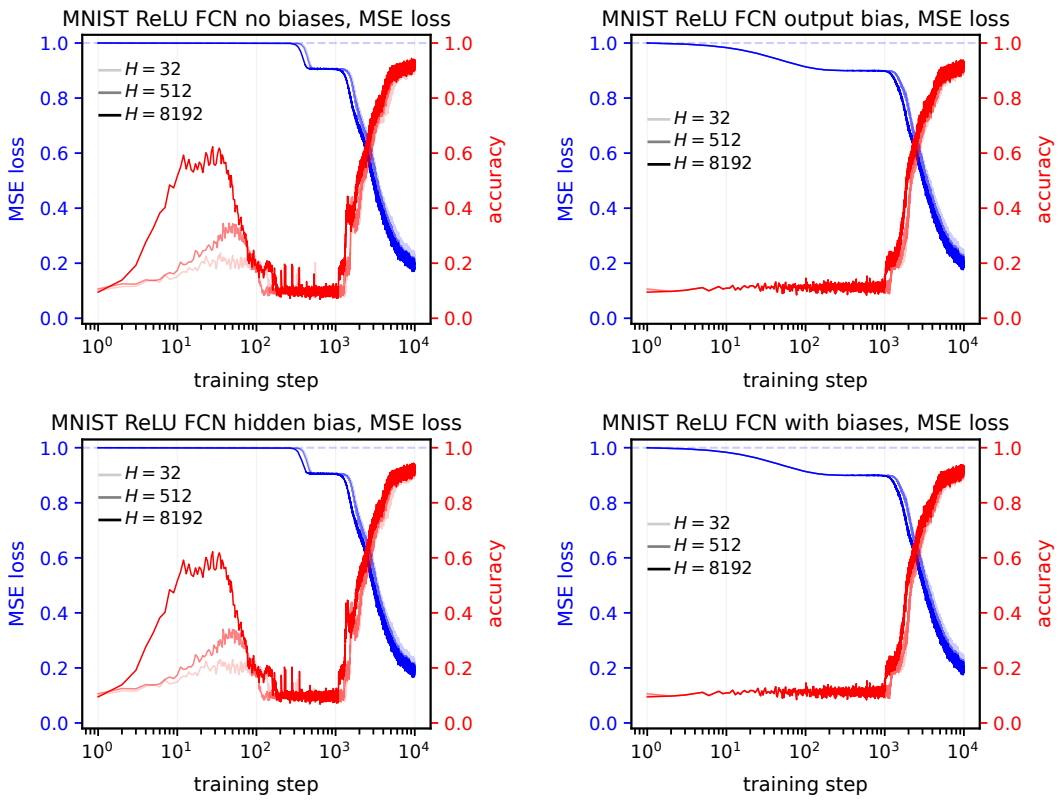

Figure 7: **Width sweeps for the MNIST ReLU bias ablation under squared loss.** The panel arrangement is the same as in Figure 6. All runs use $\sigma = 10^{-4}$ and $H \in \{32, 512, 8192\}$, with darker curves denoting larger width. The no-bias and hidden-bias-only networks exhibit pronounced nonmonotone accuracy trajectories, whereas architectures containing an output bias do not develop a comparably pronounced well.

The two sweeps are shown in the middle column of Figure 2.

The same qualitative accuracy well persists under cross-entropy. At smaller initialization, accuracy rises substantially before the cross-entropy has moved far from its zero-logit value, then falls before the later high-accuracy state is reached. Increasing width changes the timing and prominence of the early classifier while preserving the nonmonotone trajectory. Thus the phenomenon is not tied to the one-hot squared objective used in the unary experiments.

Robustness to random seed. To verify that the observed cross-entropy trajectory is not a peculiarity of one initialization or one minibatch ordering, we repeat the bias-free ReLU experiment for five independent random seeds. We fix

$$H = 8192, \qquad \sigma = 10^{-4}, \qquad \eta = 0.005, \qquad B = 1024, \tag{923}$$

and train for $10{,}000$ SGD updates with cross-entropy. The five runs use seeds $0, 1, 2, 3, 4$; for each run both the Gaussian initialization and the minibatch-permutation stream are changed. Thus this experiment probes robustness to the two principal stochastic ingredients of the training procedure rather than varying only the initial weights.

The resulting accuracy curves are shown in the left panel of Figure 9. All five runs display the same qualitative sequence: an early accuracy hill, a substantial subsequent drop, and a late sharp recovery. The location and height of the early maximum vary moderately across seeds, but the existence and approximate timing of the well are highly reproducible.

Parameterization versus small initialization: linear controls. The next experiments isolate the role of factorization from nonlinear activation functions. We consider depth-L linear networks

$$f_\theta(x) = W_L W_{L-1} \cdots W_1 x \tag{924}$$

without biases. The architectures are

$$784 \to 10 \quad (L=1), \qquad 784 \to H \to 10 \quad (L=2), \qquad 784 \to H \to H \to 10 \quad (L=3), \tag{925}$$

with $H = 512$ for $L \geq 2$. Every layer is initialized independently with

$$(W_k)_{ij}(0) \sim \mathcal{N}\left(0, \frac{\sigma^2}{d_{k-1}}\right), \tag{926}$$

where d_{k-1} is the input dimension of layer k. Independent layer-wise Gaussian draws are generated using successive seeds $1234 + k$.

The depth comparison uses

$$L \in \{1, 2, 3\}, \qquad \eta = 0.005, \qquad B = 1024, \qquad 10{,}000 \text{ updates}, \tag{927}$$

with depth-dependent initialization scales

$$\sigma_{L=1} = 10^{-9}, \qquad \sigma_{L=2} = 10^{-3}, \qquad \sigma_{L=3} = 10^{-1}. \tag{928}$$

The scales are chosen so that the dynamics of each parameterization is clearly resolved within the common optimization horizon. Since the initialization scale changes with depth, this comparison is intended as a qualitative parameterization comparison rather than as a depth-scaling law at fixed σ.

Importantly, all three architectures represent the same class of end-to-end functions: each defines a linear map from $\mathbb{R}^{784}$ to $\mathbb{R}^{10}$. Because $H = 512 > 10$, the factorized models do not suffer an end-to-end rank restriction relative to the direct $L = 1$ classifier. The principal difference is therefore the parameterization of the linear map and the gradient dynamics that this parameterization induces.

As a complementary control, we fix the direct $L = 1$ model and vary only its initialization scale:

$$W_{ci}(0) \sim \mathcal{N}\left(0, \frac{\sigma^2}{784}\right), \qquad \sigma \in \{10^{-1}, 10^{-5}, 10^{-9}\}. \tag{929}$$

These runs again use learning rate 0.005, batch size 1024, and $10{,}000$ updates.

The left panel of Figure 8 shows that changing the parameterization can qualitatively alter the accuracy dynamics. Both factorized models develop an early hill, subsequent deterioration, and later recovery, despite having exactly the same end-to-end function class as the direct linear classifier. The effect therefore does not require a nonlinear activation or additional representational power.

The right panel provides the complementary control. We keep the direct $L = 1$ parameterization fixed and reduce the initialization scale from 10^{-1} to 10^{-9}. Despite this eight-order-of-magnitude change, the direct classifier does not develop the pronounced hill–well–recovery trajectory seen in the factorized models. In particular, the two smallest initializations produce very similar dynamics. Thus small initialization alone is not sufficient to generate an accuracy well in this setting; the factorized parameterization qualitatively changes the optimization dynamics.

Optimizer robustness. We next test the sensitivity of the phenomenon to the optimization method. We use the bias-free ReLU network with cross-entropy,

$$H = 1024, \qquad \sigma = 10^{-4}, \qquad B = 1024, \tag{930}$$

and train for $10{,}000$ updates. The optimizer configurations are

Optimizer	Learning rate	Additional parameters
SGD	10^{-2}	momentum $= 0$
SGD with momentum	10^{-2}	momentum $= 0.9$
Adam	10^{-5}	$\beta_1 = 0.9,\ \beta_2 = 0.999,\ \epsilon = 10^{-8}$

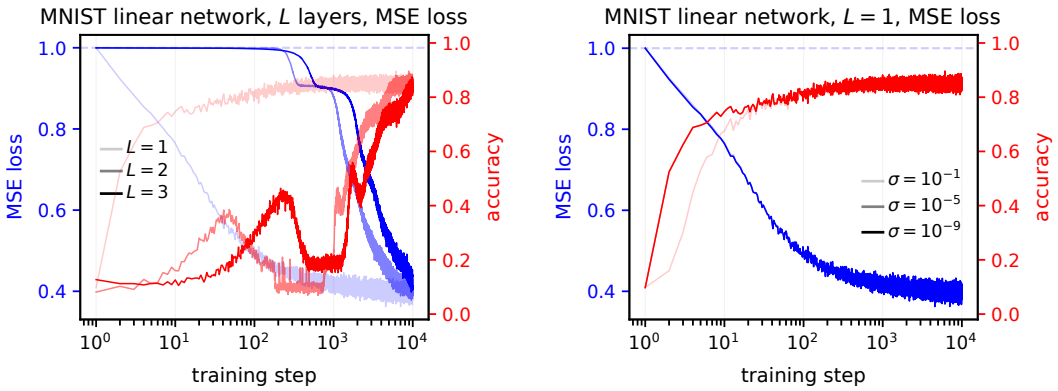

Figure 8: **Linear-network controls on MNIST under squared loss.** *Left*: comparison of the direct $L = 1$ model with two- and three-factor parameterizations $L = 2, 3$, using $\sigma_{L=1} = 10^{-9}$, $\sigma_{L=2} = 10^{-3}$, and $\sigma_{L=3} = 10^{-1}$. Although all three parameterizations represent the same class of end-to-end linear maps, the factorized models develop pronounced accuracy wells while the direct model has much simpler accuracy dynamics. *Right*: initialization-scale sweep for the direct classifier with $\sigma \in \{10^{-1}, 10^{-5}, 10^{-9}\}$. Even when the direct model is initialized extremely close to zero, its accuracy evolves comparatively monotonically and does not develop the pronounced well observed in the factorized models. Blue curves show minibatch MSE and red curves show minibatch top-one accuracy.

All three runs use zero weight decay and matched initialization and minibatch streams. The momentum run uses standard PyTorch SGD momentum, while the Adam run uses the standard PyTorch values of β_1, β_2, and ϵ. The resulting accuracy trajectories are shown in the top-right panel of Figure 2.

A pronounced well appears for both SGD variants, and nonmonotone accuracy is also visible under Adam, although the drop is substantially shallower for the Adam hyperparameters used here. Because the three optimizers use different learning rates, this experiment is best viewed as evidence that the phenomenon is not specific to plain SGD rather than as a quantitative optimizer comparison.

Neuron participation and output diversity through the well. The preceding experiments establish the phenomenology but do not by themselves probe the internal reorganization associated with the proposed target-driven organization–degeneracy bottleneck–recruitment picture. We therefore analyze two complementary diagnostics along one fixed bias-free ReLU trajectory trained with squared loss,

$$H = 8192, \qquad \sigma = 10^{-4}, \qquad \eta = 0.005, \qquad B = 1024. \tag{931}$$

The model is trained for 10,000 updates, and checkpoints are saved at 20 approximately logarithmically spaced times,

$$t \in \{0, 1, 2, 3, 5, 8, 13, 22, 36, 60, 100, 167, 278, 464, 774, \\ 1292, 2154, 3594, 5995, 10000\}. \tag{932}$$

For this diagnostic run we use initialization seed 0 and a separately seeded minibatch stream. All quantities below are evaluated post hoc on the full MNIST test set of 10,000 examples and do not affect training.

We first quantify how broadly the class-dependent score contribution is distributed across hidden neurons. Write the bias-free network as

$$f_\theta(x) = \sum_{h=1}^{H} v_h \, \phi(w_h^\top x), \qquad \phi(z) = \max\{z, 0\}, \tag{933}$$

where $v_h \in \mathbb{R}^C$, with $C = 10$, is the output-weight vector associated with hidden neuron h. For each neuron, define its class-conditional mean activation

$$a_h(c) = \frac{1}{N_c} \sum_{i:\, y_i = c} \phi(w_h^\top x_i), \qquad c = 1, \ldots, C, \tag{934}$$

where N_c is the number of test examples of class c, and let $a_h \in \mathbb{R}^C$ denote the vector with entries $a_h(c)$.

A component of v_h proportional to the all-ones vector shifts all output logits equally and therefore cannot affect their ordering. We remove this classification-irrelevant direction by defining

$$\widetilde{v}_h = v_h - \frac{\mathbf{1}^\top v_h}{C}\mathbf{1}. \tag{935}$$

The corresponding contribution of neuron h to the centered class-conditional score matrix is

$$\widetilde{B}_h = a_h \widetilde{v}_h^\top \in \mathbb{R}^{C\times C}. \tag{936}$$

We associate to neuron h the contribution energy

$$e_h = \|\widetilde{B}_h\|_{\mathrm{F}}^2 = \|a_h\|_2^2 \|\widetilde{v}_h\|_2^2. \tag{937}$$

This measures the magnitude of the neuron's contribution to class-dependent score differences; it does not assume that the contribution is individually aligned with the correct classifier.

We quantify the spread of this contribution over the hidden layer using the participation ratio

$$N_{\mathrm{eff}} = \frac{\left(\sum_{h=1}^H e_h\right)^2}{\sum_{h=1}^H e_h^2}, \tag{938}$$

and report the normalized *effective participation fraction*

$$\rho_{\mathrm{eff}} = \frac{N_{\mathrm{eff}}}{H}. \tag{939}$$

Whenever the total contribution energy is nonzero, $1/H \le \rho_{\mathrm{eff}} \le 1$. A larger value means that class-dependent score contribution is distributed over a larger effective fraction of the hidden layer, whereas a smaller value indicates concentration onto fewer neurons. Importantly, ρ_{eff} measures concentration of contribution magnitudes rather than their correctness or mutual coherence, and is therefore interpreted jointly with classification accuracy.

As a complementary function-space diagnostic, we measure how broadly the network uses the available output classes. At checkpoint t, let

$$\widehat{y}_i(t) = \arg\max_c f_{\theta(t),c}(x_i) \tag{940}$$

be the predicted class of test example i, and define the marginal predicted-class frequencies

$$p_c(t) = \frac{1}{N_{\mathrm{test}}} \sum_{i=1}^{N_{\mathrm{test}}} \mathbf{1}\{\widehat{y}_i(t) = c\}, \qquad c = 1, \dots, C. \tag{941}$$

We define the effective number of used output classes by the participation ratio

$$C_{\mathrm{eff}}(t) = \frac{1}{\sum_{c=1}^C p_c(t)^2}, \tag{942}$$

and normalize it by the total number of classes,

$$\rho_{\mathrm{out}}(t) = \frac{C_{\mathrm{eff}}(t)}{C} = \frac{1}{C\sum_{c=1}^C p_c(t)^2}. \tag{943}$$

We refer to ρ_{out} as the *output diversity*. It lies between $1/C$ and 1. Uniform use of all C output classes gives $\rho_{\mathrm{out}} = 1$, while complete collapse onto a single predicted class gives $\rho_{\mathrm{out}} = 1/C$. More generally, approximately uniform use of only k classes gives $\rho_{\mathrm{out}} \approx k/C$. Unlike accuracy, output diversity does not measure whether predictions are correct: a classifier may distribute its predictions evenly across classes while assigning the wrong label to every example. It therefore supplies information complementary to accuracy.

The right panel of Figure 4 shows full-test accuracy, the hidden-neuron participation fraction ρ_{eff}, and the output diversity ρ_{out} at the saved checkpoints. The three quantities reveal a structured

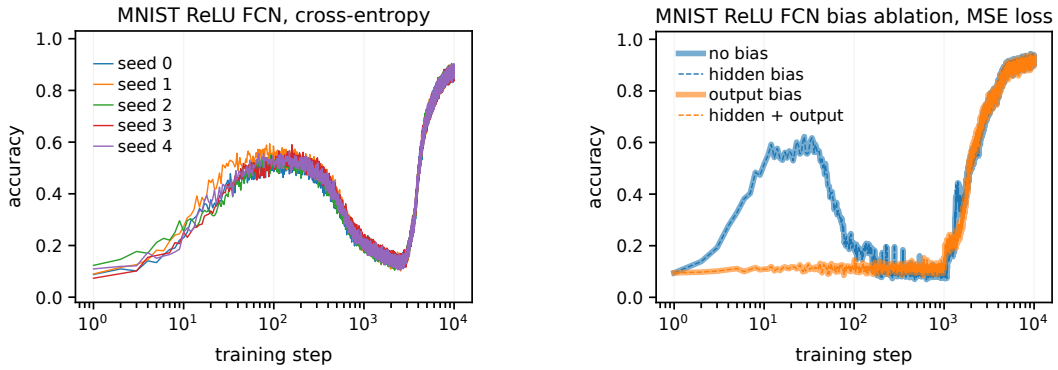

Figure 9: **Robustness and bias ablation for the MNIST accuracy well.** *Left*: seed robustness for the bias-free ReLU network trained with cross-entropy at $H = 8192$, $\sigma = 10^{-4}$, showing raw minibatch accuracy trajectories for five independent initializations and minibatch streams. All five runs exhibit the same hill–well–recovery pattern. *Right*: four-way bias ablation under squared loss at $H = 8192$, $\sigma = 10^{-4}$. The pronounced well is preserved with no bias or a hidden bias alone, but is suppressed when an output bias is present.

sequence consistent with the heuristic picture of Section 3. During the initial target-driven organization, accuracy rises while output predictions remain distributed across several classes and the class-dependent score contribution is spread over a relatively large effective fraction of the hidden layer. As the trajectory enters the accuracy well, both diagnostics collapse: ρ_{eff} falls sharply, showing that the centered class-dependent score contribution becomes concentrated on a much smaller effective subset of hidden neurons, while ρ_{out} approaches its minimum value $1/C$, indicating that predictions themselves become concentrated onto very few output classes. Thus the accuracy minimum coincides with a bottleneck both in hidden-unit participation and in the diversity of the resulting classifier outputs.

The recovery stage reverses both forms of degeneracy. Output diversity rises rapidly toward one as the classifier begins to use the full set of output classes again. Hidden-neuron participation also rebounds, indicating renewed involvement of additional components of the hidden layer, although it does not return to its initial value. Accuracy rises concurrently to its final high level. This simultaneous restoration of output diversity and partial increase in neuronal participation is consistent with the recruitment stage of the heuristic picture.

The two diagnostics probe different aspects of the same transition. ρ_{eff} is a parameterization-dependent measure of how class-dependent score contribution is distributed across hidden neurons, whereas ρ_{out} is a purely functional statistic describing the marginal diversity of the classifier's predictions. Their simultaneous collapse near the bottom of the well therefore gives more specific evidence for a transient degeneracy bottleneck than accuracy alone.

Taken together, the MNIST experiments show that the accuracy well persists across two losses, a wide range of widths, multiple random seeds, and several optimization methods, while remaining sensitive to parameterization. A class-wise output bias suppresses the pronounced well, whereas a hidden bias does not; factorizing a linear classifier can create a strong well without changing its end-to-end function class, while simply shrinking the initialization of the direct classifier does not.

The checkpoint diagnostics provide a complementary mechanistic observation. As the early classifier deteriorates, class-dependent score contribution becomes sharply concentrated onto a smaller effective subset of hidden neurons, while the marginal distribution of predicted labels simultaneously collapses onto very few output classes. Both forms of degeneracy are reversed during recovery: output diversity rapidly approaches its maximum and hidden participation partially rebounds as accuracy rises. This gives quantitative support for the heuristic picture developed in Section 3, while leaving open the precise nonlinear dynamics responsible for the bottleneck.

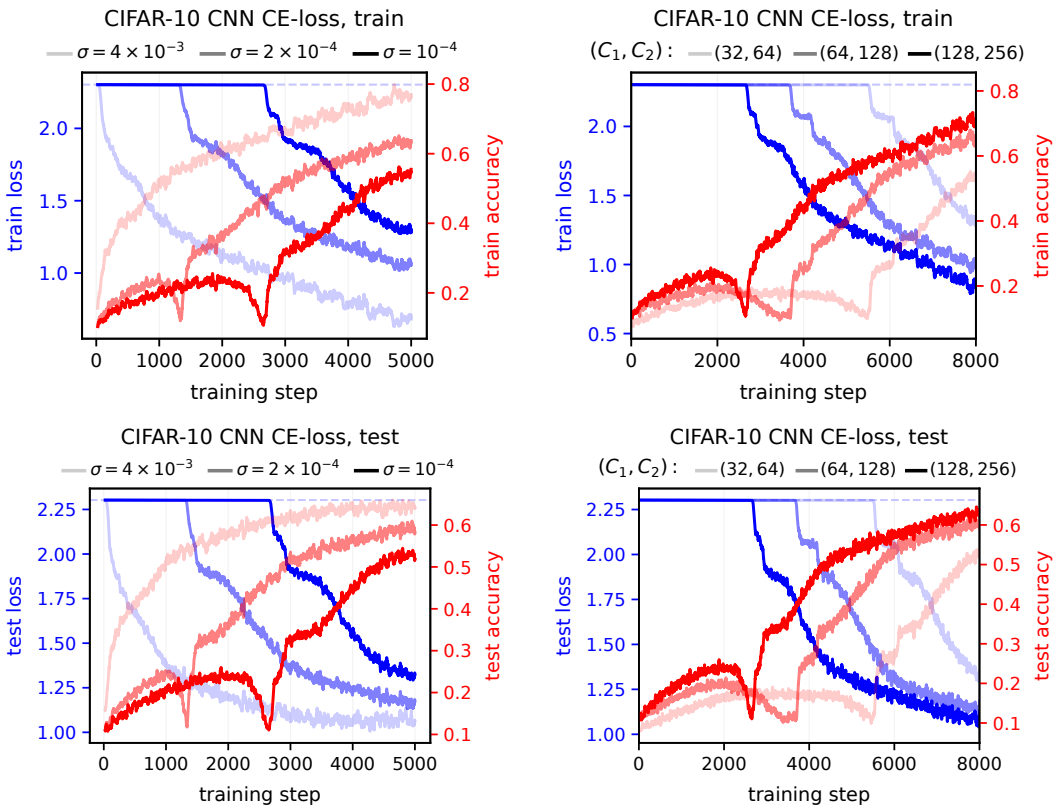

Figure 10: **Accuracy-well dynamics in a two-hidden-layer CNN on CIFAR-10.** Columns correspond to the initialization-scale sweep (left) and the model-capacity sweep (right), while rows show training dynamics (top) and held-out test dynamics (bottom). In the left column, the channel widths are fixed to $(C_1, C_2) = (128, 256)$, and darker curves correspond to smaller initialization scale σ. In the right column, $\sigma = 10^{-4}$ is fixed, and darker curves correspond to larger channel widths. Blue curves show cross-entropy loss and red curves show top-one accuracy. The horizontal dashed blue line denotes the zero-logit cross-entropy $\log 10$. Each displayed trajectory is a 25-step moving average of the corresponding per-minibatch measurements. Small initialization exposes a pronounced early accuracy hill followed by an accuracy well, with substantial accuracy dynamics occurring while the cross-entropy remains close to its zero-logit value. In the capacity sweep, larger models form a stronger early classifier and undergo an earlier and more pronounced hill–well–recovery trajectory. The corresponding held-out test curves closely reproduce the qualitative training dynamics.

C.3 Accuracy wells in a convolutional network on CIFAR-10

The controlled experiments above establish the accuracy-well phenomenon in analytically tractable matrix-factorization models and in fully connected networks on MNIST. We next ask whether the same qualitative dynamics remain visible in a more conventional nonlinear architecture trained on natural images. To this end, we train a simple two-hidden-layer convolutional neural network on CIFAR-10 (Krizhevsky, 2009). Our objective is not to optimize benchmark performance, but rather to test for the phenomenon in a minimal genuinely convolutional setting. We therefore omit pooling, normalization layers, residual connections, dropout, and data augmentation, and study separately the dependence on initialization scale and model capacity. We record both training and held-out test dynamics.

Architecture and data preprocessing. We use the standard CIFAR-10 train–test split, consisting of 50,000 training and 10,000 test RGB images of spatial resolution 32×32 from ten classes (Krizhevsky, 2009). The dataset is loaded through the Hugging Face `uoft-cs/cifar10` inter-

face. Pixel intensities are first divided by 255 and subsequently normalized channelwise using

$$\mu = (0.4914,\ 0.4822,\ 0.4465), \qquad s = (0.2470,\ 0.2435,\ 0.2616). \tag{944}$$

No data augmentation is applied.

The network consists of two convolutional hidden layers followed by a fully-connected classifier:

$$h_1(x) = \mathrm{ReLU}\Big(\mathrm{Conv}_{3\times3}^{3\to C_1}(x)\Big), \tag{945}$$

$$h_2(x) = \mathrm{ReLU}\Big(\mathrm{Conv}_{3\times3}^{C_1\to C_2}(h_1(x))\Big), \tag{946}$$

and

$$f_\theta(x) = W_{\mathrm{out}}\,\mathrm{vec}(h_2(x)), \qquad W_{\mathrm{out}} \in \mathbb{R}^{10\times(32\cdot32\cdot C_2)}. \tag{947}$$

Both convolutions have kernel size 3×3, stride one, and padding one, so the spatial resolution remains 32×32 through both convolutional layers. All convolutional and fully-connected layers are bias-free. The model contains no pooling, batch normalization, dropout, or residual connections.

Optimization and initialization. We minimize the mean cross-entropy loss using plain minibatch SGD with fixed learning rate

$$\eta = 0.025, \tag{948}$$

batch size 128, zero momentum, and zero weight decay. No learning-rate schedule is used.

Every scalar entry of each convolutional and fully-connected weight tensor is initialized independently according to

$$\theta_j(0) \sim \mathcal{N}(0, \sigma^2). \tag{949}$$

Thus, in these experiments, σ denotes a common *entrywise* standard deviation and is not divided by the fan-in of the corresponding layer. This choice deliberately differs from the parameterizations used in the preceding unary and MNIST experiments and therefore provides an additional robustness test: the CIFAR-10 experiments ask whether the accuracy-well phenomenology persists under a substantially different initialization convention rather than being tied to one particular normalization.

Measurement protocol and reproducibility. At every optimization step, the implementation draws one training minibatch and one test minibatch, each containing 128 examples. Separate shuffled streams are used for the two splits. Training loss and accuracy and test loss and accuracy are all evaluated at the same parameter state, immediately before the SGD update. Only the training cross-entropy is differentiated and used to update the parameters; the test data therefore play no role in optimization.

Within each epoch, a fresh random permutation of the corresponding split is generated and divided into complete minibatches of size 128; an incomplete remainder is omitted before a new permutation is generated. Loss and accuracy are stored at every optimization step. The curves in Figure 10 show 25-step moving averages of these per-minibatch measurements.

We match the pseudorandom streams across experimental conditions. The global seed is 1, the dedicated initialization generator uses seed 123, and the training and test permutation generators use seeds 456 and 789, respectively. The initialization generator is reset at the beginning of each run. Consequently, within the fixed-architecture initialization-scale sweep, the same underlying Gaussian realization is multiplied by the different values of σ. The training and test data streams are likewise reset for every run, so corresponding conditions encounter the same sequences of minibatch indices. Each curve displayed in Figure 10 corresponds to one run with these fixed seeds.

The complete implementation of experiments, including data loading, preprocessing, model construction, optimization, metric logging, and storage of the trajectories, is provided in the supplementary Jupyter notebook `cifarcnnaccwell.ipynb`.

Initialization-scale sweep. We first fix the architecture to

$$(C_1, C_2) = (128, 256) \tag{950}$$

and train for 5,000 SGD updates using

$$\sigma \in \left\{4 \times 10^{-3},\ 2 \times 10^{-4},\ 10^{-4}\right\}. \tag{951}$$

The left column of Figure 10 shows that decreasing the initialization scale progressively separates the early accuracy dynamics from the macroscopic decay of the cross-entropy.

For the largest initialization, $\sigma = 4 \times 10^{-3}$, the cross-entropy begins decreasing almost immediately and the accuracy increases without a pronounced accuracy well. At the two smaller initialization scales, the behavior is qualitatively different. The loss remains close to the zero-logit value

$$\mathcal{L}_0 = \log 10, \tag{952}$$

for an extended interval, even though accuracy has already risen appreciably above chance.

The separation is clearest for $\sigma = 10^{-4}$. While the loss is still essentially on its initial plateau, the accuracy first rises from chance and forms a clear early hill. It then drops sharply back toward chance before recovering again as the cross-entropy enters its macroscopic descent. Thus, the nonmonotonicity is already present in a regime in which the continuous training objective changes very little: useful class-ordering information is formed and subsequently degraded before conventional loss-based learning becomes macroscopically visible.

Reducing σ also delays the characteristic transitions. The accuracy well occurs earlier for $\sigma = 2 \times 10^{-4}$ and later for $\sigma = 10^{-4}$, consistently with the general role of small initialization in opening a longer microscopic dynamical regime.

The held-out trajectories in the bottom-left panel show the same qualitative behavior at closely corresponding training times. In particular, for small initialization, test accuracy also develops an early hill, subsequently collapses, and later recovers. Hence the transient early predictor is not merely fitting the minibatches used for the SGD updates: it already captures class-dependent structure that generalizes to unseen CIFAR-10 examples.

Model-capacity sweep. We next fix

$$\sigma = 10^{-4} \tag{953}$$

and vary the convolutional channel widths according to

$$(C_1, C_2) \in \{(32, 64),\ (64, 128),\ (128, 256)\}\,. \tag{954}$$

Each condition is trained for $10{,}000$ SGD updates. To focus on the transient regime of interest, the right column of Figure 10 displays the first $8{,}000$ updates.

The capacity sweep shows a strong systematic dependence of the transient dynamics on model size. The largest network, $(C_1, C_2) = (128, 256)$, forms the strongest early classifier and undergoes the earliest pronounced accuracy well. The intermediate network develops a lower early hill and makes the corresponding transition later, while the smallest network forms the weakest early classifier and remains in the microscopic regime for considerably longer. Nonmonotone structure is nevertheless visible throughout the family. The same ordering and approximately the same transition times appear in the held-out test trajectories.

The capacity sweep is not intended as a controlled infinite-width scaling limit. Because the same entrywise initialization variance σ^2 is used while C_1 and C_2 vary, changing the channel widths also changes aggregate activation and initialization statistics. Accordingly, the experiment does not isolate width while holding all other initialization statistics fixed. Its empirical conclusion is instead that, within this family of increasingly large CNNs under a fixed entrywise initialization convention, greater model capacity is associated with a stronger early classifier, an earlier transition, and a more pronounced hill–well–recovery trajectory.

This parametrization difference is also useful from the broader phenomenological perspective. The preceding unary and MNIST experiments and the present CNN experiment use different initialization conventions, yet closely related accuracy-well dynamics are observed across these settings. The CIFAR-10 results therefore provide evidence that the phenomenon persists across substantially different parameterizations; we do not, of course, claim invariance under arbitrary choices of parametrization.

Additional deep convolutional experiment. As an additional architectural stress test, we also consider a substantially deeper convolutional network with five hidden convolutional layers followed by a fully connected output layer. We use

$$(C_1, C_2, C_3, C_4, C_5) = (256, 256, 256, 256, 256), \tag{955}$$

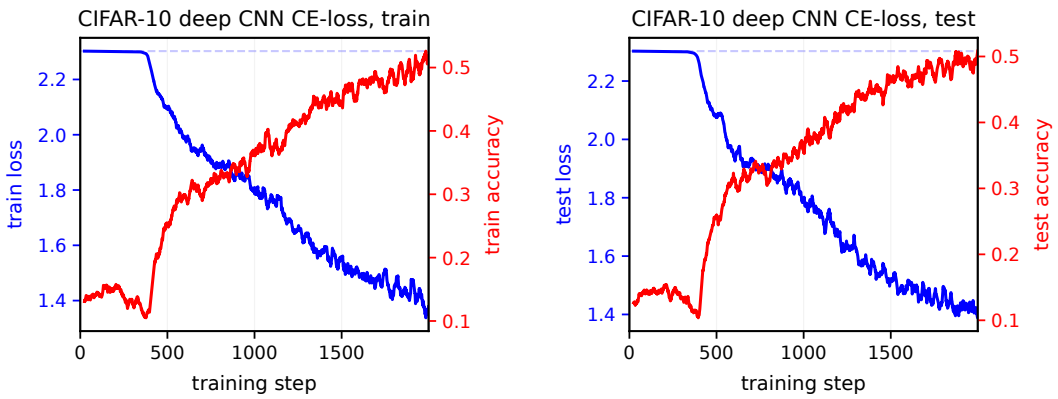

Figure 11: **Dynamics of a deeper convolutional network on CIFAR-10.** Training trajectories are shown on the left and held-out test trajectories on the right. The network has five convolutional hidden layers with $(C_1, \dots, C_5) = (256, 256, 256, 256, 256)$, followed by a fully connected ten-class output layer. All weights use entrywise Gaussian initialization with $\sigma = 8\times10^{-3}$; optimization uses minibatch SGD with learning rate 0.025 and batch size 128. Blue curves show cross-entropy loss and red curves show top-one accuracy; the horizontal dashed blue line denotes the zero-logit cross-entropy $\log 10$. Curves are 25-step moving averages of the per-minibatch measurements.

with a ReLU nonlinearity after every convolution. As in the preceding CNN experiments, all convolutions have 3×3 kernels, stride one, padding one, and no biases, pooling, normalization layers, dropout, or residual connections. We retain the same CIFAR-10 preprocessing, minibatch size 128, SGD learning rate $\eta = 0.025$, and train–test measurement protocol described above. All weight entries are initialized independently from

$$\mathcal{N}(0, \sigma^2), \qquad \sigma = 8 \times 10^{-3}, \tag{956}$$

and the network is trained for $2{,}000$ SGD updates. Figure 11 shows the resulting training and held-out test trajectories. Despite the considerable increase in depth relative to the two-convolution architecture considered above, the dynamics remain strongly structured at small initialization: accuracy begins to improve while the network is still close to the zero-logit regime, and the corresponding training and test trajectories evolve on closely related timescales. This additional experiment therefore shows that the qualitative separation between early predictive dynamics and macroscopic loss evolution is not restricted to the shallow convolutional architecture used in the parameter sweeps above.

Summary. The CIFAR-10 experiments extend the observed accuracy-well phenomenology to a nonlinear convolutional architecture trained on natural images. At sufficiently small initialization, the CNN exhibits the characteristic sequence

$$\text{chance} \longrightarrow \text{early accuracy hill} \longrightarrow \text{accuracy well} \longrightarrow \text{recovery}, \tag{957}$$

with substantial hill–well dynamics occurring while the cross-entropy remains close to its zero-logit value. Smaller initialization separates the microscopic accuracy dynamics more clearly from macroscopic loss decay, while the capacity sweep shows that larger models in the tested family form a stronger early classifier and traverse the subsequent transition earlier. The close correspondence between training and held-out trajectories further shows that the transient classifier already generalizes beyond the optimization minibatches.

We do not infer from these experiments that the internal mechanism producing the CNN accuracy well is identical to the spectral mechanism proved for the unary targets. Rather, the results show that the same qualitative nonmonotone accuracy dynamics survive substantial changes in architecture, dataset, loss function, stochastic optimization, and initialization parametrization, supporting the broader phenomenological picture developed in this work.