How are linear representations learned? Exact solutions to the dynamics of abstraction

William W. Yang[1] **Andrew M. Saxe**[*,1,2] **Peter E. Latham**[*,1]

Abstract

In neural networks, concepts often appear to be encoded as consistent linear directions in representation space. This idea is known as the linear representation hypothesis and underpins many methods for interpretability and control, from linear probing to activation steering. However, the way these representations emerge during training remains poorly understood. Here, we develop a framework to study how linear representations form during training via the alignment of concept directions – a process we call "abstraction". In a linear network setting, we obtain exact solutions for the full trajectory of abstraction. These solutions reveal key analytic principles governing abstraction: (i) data and target geometry jointly determine abstraction at the end-of-learning, (ii) initialization scale controls the maximum abstraction reached during training, and (iii) abstraction improves with network depth. We extend our theory to nonlinear networks and find that ReLU makes abstraction dynamics depend more on input geometry than target geometry. More generally, we prove that abstraction is typically weaker in activations compared to preactivations for a broad class of activation functions. We find evidence for this in Gemma 4 and DINOv3 and apply our theory to improve linear probe generalization in LLMs. Together, our results provide a dynamical theory of representation learning with implications for interpretability.

1 Introduction

A recurring empirical observation in deep neural networks is that high-level concepts are often represented as approximately linear directions in activation space [1]. This idea has recently been formalized as the linear representation hypothesis (LRH) [2, 3, 4]. To use a classical example [5], a linear representation of the "gender" concept implies that the concept vectors $\boldsymbol{v}_{\text{king}} - \boldsymbol{v}_{\text{queen}}$ and $\boldsymbol{v}_{\text{man}} - \boldsymbol{v}_{\text{woman}}$ are approximately parallel in model activations. A related idea exists in neuroscience, where a concept is said to be represented in an "abstract" format when concept vectors are highly aligned across contexts [6, 7]. Borrowing that neuroscience terminology, we use the term "abstraction" to refer to the alignment of concept vectors, i.e. $\cos(\boldsymbol{v}_{\text{king}} - \boldsymbol{v}_{\text{queen}}, \boldsymbol{v}_{\text{man}} - \boldsymbol{v}_{\text{woman}})$, during training.

Understanding how and when abstraction happens in training is central to modern deep learning. Numerous methods in interpretability rely on the existence of linear representations by extracting concept directions for linear probing [8, 9] or activation steering [10, 11]. Despite this, our understanding of how such directions arise remains incredibly limited.

Previous work has studied abstraction in the limit of infinite training [3], or at the global minima of the loss landscape [12]. This leaves open a simple question: what are the dynamics of abstraction during finite training? More importantly, these previous theories considered settings in which representations become perfectly abstract (i.e. fully linear), whereas this is far from the case in real models.

[*]Co-senior authors; equal contribution.
[1]Gatsby Computational Neuroscience Unit, University College London, London W1T 4JG, United Kingdom.
[2]Sainsbury Wellcome Centre, University College London, London W1T 4JG, United Kingdom.
Correspondence to: <william.yang.25@ucl.ac.uk>.

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

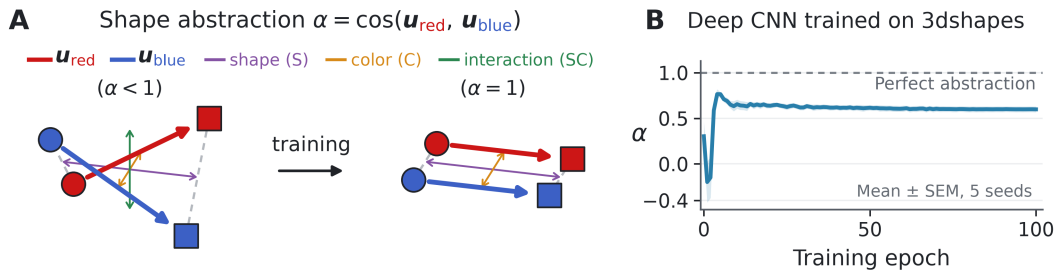

Figure 1: **Problem setting.** We consider neural networks trained on inputs $\boldsymbol{x} \sim p(\boldsymbol{x} \mid s, c)$ and targets $\boldsymbol{y}(s, c)$ where (s, c) are binary latents for the concepts "shape" and "color". We ask: how does the network learn a linear representation of the shape concept? We measure this using the *abstraction score* α. **(A)** Illustrative cartoon of abstraction. Left: At initialization, α is low and increases during training. Right: When $\alpha = 1$, the shape concept vectors $\boldsymbol{u}_{\text{red}}$ and $\boldsymbol{u}_{\text{blue}}$ are parallel, and the shape concept is represented linearly in the activations. We call this *perfect abstraction*. In §3 we develop a fully solvable setting where α is determined by the relative magnitude of the direction encoding the shape-color interaction (**green**), and we are able to derive exact solutions for the entire training-time trajectory of $\alpha(t)$ (Theorem 1). **(B)** We plot α for a deep ReLU CNN training on the 3dshapes dataset [13] (see §E.3.2 for experiment details). This reveals non-trivial, non-monotonic dynamics which prior work cannot explain. Open questions include: Why does α plateau below 1, unlike in previous theories [3, 12]? What causes the early peak, and what determines its height? We use our theory to address these and more.

In real-world settings, we often see nontrivial, non-monotonic dynamics of abstraction during training, and representations are often far from perfectly abstract. For example, see Fig. 1B. This motivates the central questions of this paper: *What is the trajectory of abstraction during training? How is it affected by properties of the dataset? What are the roles of depth and nonlinearities?* We aim to address these in an exactly solvable setting enabling complete insight into the dynamics of abstraction.

Key contributions. In §3-§4 we study a minimal linear network setting to obtain exact solutions for the training-time trajectory of abstraction and then use these to derive laws for the final and maximum abstraction reached during training (§4.1). We then extend the theory to arbitrarily deep networks and derive laws for the effect of depth on abstraction (§4.2). In §5 we extend our theory to infinite-width nonlinear networks. In this setting we prove that a broad class of activation functions generally weaken abstraction (§5.1) and show that ReLU reshapes dynamics to depend more on inputs than targets (§5.2). In §6, we apply our theory to deep learning and neuroscience, by showing that ablating nonlinearities improves abstraction and linear probe generalization in DINOv3 and Gemma 4 (Fig. 6), and that abstraction improves along the macaque ventral-stream hierarchy (Fig. 7).

2 Related work

AI safety and interpretability. Linear concept directions play a prominent role in empirical approaches to LLM interpretability. Examples include detecting deception [9], eliciting truthful answers [11], and identifying "persona" vectors underlying character traits [14]. Recently, [15] provided evidence for the formation of persona directions early in pretraining. Our work complements this by characterizing how dataset, initialization, and architecture shape the emergence of such directions in training. A common empirical strategy for identifying concept directions is to use a contrastive dataset and compute the difference-of-means in activations (e.g. evil vs harmless). In §3.1, we show the practical relevance of our work: greater abstraction improves how well the identified direction generalizes and reduces the number of samples needed to reliably estimate a steering vector.

Abstract representations. Recent experimental work has shown that high-level brain regions develop a so-called "abstract" representational geometry supporting generalization [6, 7, 16, 17, 18, 19, 20, 21, 22]. In neuroscience, this geometry is quantified by *parallelism* [6]. In this paper, our abstraction score is a special case of the parallelism for two contexts. Recent empirical work showed that representations in ReLU networks preserved input geometry more than sigmoid-like networks

[23]. In §5.2, we provide a theoretical account of these findings. Recently, [12] proved that network representations become perfectly abstract in their final layer at the global optimum of the loss. Our work studies the gradient flow of abstraction (rather than assuming optimality) as well as its evolution across layers and nonlinearities, and extends their setting to nonlinear inputs/targets.

Linear networks. We build on a line of previous work on linear network learning dynamics, e.g. [24, 25, 26, 27], which have been studied as analytically solvable yet insightful settings [28, 29]. Recent works have characterized how initialization shapes learning dynamics and hidden representational similarity in two-layer linear networks [30, 31]. We complement these results by deriving explicit and interpretable expressions for the behavior of a specific representation observable (abstraction score) for linear networks of arbitrary depth and in two-layer nonlinear networks.

3 Preliminaries

Here, we propose a minimal model which is simple enough to admit exact solutions for dynamics, yet complex enough to exhibit the phenomenon of abstraction. Consider the two-layer linear network

$$f_{\boldsymbol{W},\boldsymbol{W}_r}(\boldsymbol{x}) = \boldsymbol{W}_r\boldsymbol{W}\boldsymbol{x}, \quad \boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}, \quad \boldsymbol{W} \in \mathbb{R}^{D\times D_x}, \quad \boldsymbol{W}_r \in \mathbb{R}^{D_y\times D}, \tag{1}$$

where $\boldsymbol{z}$ is the hidden representation. We study the population loss with ridge penalty γ^{-1} on $\boldsymbol{W}_r$

$$\mathcal{L}(\boldsymbol{W}_r, \boldsymbol{W}) = \frac{1}{2}\,\mathbb{E}\left[\|f_{\boldsymbol{W},\boldsymbol{W}_r}(\boldsymbol{x}) - \boldsymbol{y}\|^2\right] + \frac{1}{2\gamma}\|\boldsymbol{W}_r\|_F^2, \tag{2}$$

Following [2, 32, 12], we model "concepts" as binary latent variables, and consider the case of two latents $(s, c) \in \{+1, -1\}^2$. As mnemonics, we refer to s as "**s**hape" ($+1$: square, -1: circle) and c as "**c**olor" ($+1$: red, -1: blue). We model the conditional distribution $p(\boldsymbol{x} \mid s, c)$ as a mixture:

$$\boldsymbol{x} = \boldsymbol{\mu}(s, c) + \sigma_x\boldsymbol{\epsilon}, \qquad \mathbb{E}[\boldsymbol{\epsilon}] = 0, \qquad \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \boldsymbol{I}, \qquad \boldsymbol{\epsilon} \perp (s, c), \tag{3}$$

and we model the target $\boldsymbol{y} \equiv \boldsymbol{y}(s, c)$ as a deterministic function of the latents.[3] The latents have the uniform marginal distribution $p(s, c) = 1/4$, giving four balanced classes $\{\square, \square, \circ, \circ\}$. The dataset enters the learning dynamics via the input and input-target correlation matrices, defined as

$$\boldsymbol{\Sigma}_{XX} \coloneqq \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top], \qquad \boldsymbol{\Sigma}_{XY} \coloneqq \mathbb{E}[\boldsymbol{x}\boldsymbol{y}^\top]. \tag{4}$$

Our main quantity of interest is the ***abstraction score*** of shape. Let $\bar{\boldsymbol{z}}_a = \mathbb{E}[\boldsymbol{z} \mid a]$ be the mean representation of class $a \in \{\square, \square, \circ, \circ\}$. The abstraction score for shape is the cosine similarity:

$$\alpha \coloneqq \cos(\boldsymbol{u}_{\text{red}}, \boldsymbol{u}_{\text{blue}}), \qquad \boldsymbol{u}_{\text{red}} = \bar{\boldsymbol{z}}_{\square} - \bar{\boldsymbol{z}}_{\circ}, \qquad \boldsymbol{u}_{\text{blue}} = \bar{\boldsymbol{z}}_{\square} - \bar{\boldsymbol{z}}_{\circ}. \tag{5}$$

For brevity, this paper will focus only on shape abstraction, but all analysis can be extended to color by symmetry. The case of $\alpha = 1$ corresponds to the definition of "linearity" in linear representation hypothesis literature [2, 3] (see §A.2).

Our aim is to obtain exact solutions for the trajectory of $\alpha(t)$ during gradient flow. This is generically intractable, so we make two key simplifying assumptions in our analysis. The first assumption concerns the regularized readout $\boldsymbol{W}_r$.[4] We assume the following:

Assumption 1 (Optimal Readout)**.** *The readout is the unique minimizer of $\mathcal{L}(\boldsymbol{W}_r, \boldsymbol{W})$ given $\boldsymbol{W}$:*

$$\boldsymbol{W}_r(t) = \boldsymbol{W}_r^*(\boldsymbol{W}(t)) = \gamma\boldsymbol{\Sigma}_{XY}^\top\boldsymbol{W}(t)^\top\left(\boldsymbol{I} + \gamma\boldsymbol{W}(t)\boldsymbol{\Sigma}_{XX}\boldsymbol{W}(t)^\top\right)^{-1}, \quad \forall t$$

This corresponds to the limit in which the learning rate for $\boldsymbol{W}_r$ is much larger than for $\boldsymbol{W}$ [33, 34]. We argue this is more realistic than freezing $\boldsymbol{W}_r$, which is commonly assumed in prior work on learning dynamics, e.g. [35, 36]. In §A.4 we empirically study this assumption and show that it still captures the key qualitative behavior of the naturalistic case where $\boldsymbol{W}_r$ has a finite learning rate.

For our second assumption, we select a natural basis for studying the abstraction α. By leveraging the standard generalization of Fourier transforms to finite groups [37], Theorem A.2 states that any

[3] Our setting relaxes assumptions in prior work [12] by allowing both $\boldsymbol{y}$ and $\boldsymbol{\mu}$ to be nonlinear in the latents.

[4] Note that in (2), we deliberately regularize only $\boldsymbol{W}_r$ since our quantity of interest α is invariant to the scale of $\boldsymbol{W}$. For intuition, notice in Fig. 1A, the angle between the red and blue vectors is invariant to their norms.

function $\{+1,-1\}^K \mapsto \mathbb{R}^d$ has a unique multilinear polynomial expansion in the Fourier basis. Without loss of generality, we can thus expand in the Fourier *modes* $m \in \mathcal{M} := \{\emptyset, S, C, SC\}$,

$$\boldsymbol{\mu}(s,c) = \sum_m \hat{\boldsymbol{x}}_m \chi_m(s,c), \quad \boldsymbol{y}(s,c) = \sum_m \hat{\boldsymbol{y}}_m \chi_m(s,c), \quad \mathbb{E}[\boldsymbol{z}(t) \mid s,c] = \sum_m \hat{\boldsymbol{z}}_m(t)\chi_m(s,c), \tag{6}$$

where $\chi_m = 1, s, c, sc$ for $m = \emptyset, S, C, SC$ respectively. We call $\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}, \hat{\boldsymbol{z}}$ the Fourier *components* and χ_m are *characters* of the group $(\mathbb{Z}_2)^2$ associated with (s,c).[5] In this Fourier basis we assume,

Assumption 2 (Fourier Orthogonality)**.** *The Fourier components are mutually orthogonal:*

$$\langle \hat{\boldsymbol{x}}_m, \hat{\boldsymbol{x}}_{m'} \rangle = d_m^{(x)} \delta_{m,m'}, \quad \langle \hat{\boldsymbol{y}}_m, \hat{\boldsymbol{y}}_{m'} \rangle = d_m^{(y)} \delta_{m,m'}, \quad \langle \hat{\boldsymbol{z}}_m(0), \hat{\boldsymbol{z}}_{m'}(0) \rangle = q_{m,0}\, \delta_{m,m'},$$

where $m, m' \in \mathcal{M}$, $d_m^{(x)} := \|\hat{\boldsymbol{x}}_m\|^2, d_m^{(y)} := \|\hat{\boldsymbol{y}}_m\|^2$ *are Fourier energies of the input and target, and* $q_{m,0} := \|\hat{\boldsymbol{z}}_m(0)\|^2$ *is the initial Fourier energy of the representation at training time* $t = 0$.

Under this assumption, $\boldsymbol{\Sigma}_{XX}, \boldsymbol{\Sigma}_{XY}$ diagonalize in the Fourier components $\hat{\boldsymbol{x}}_m$, and abstraction has a very natural, tractable expression in terms of the representation's Fourier energies q_m:

$$\alpha = \frac{q_S - q_{SC}}{q_S + q_{SC}} = \frac{1-\nu_z}{1+\nu_z}, \qquad \nu_z := \frac{q_{SC}}{q_S}, \qquad q_m := \|\hat{\boldsymbol{z}}_m\|^2, \qquad m \in \mathcal{M}. \tag{7}$$

We call ν_z the representation's ***entanglement ratio***, since the mode SC encodes shape-color interaction. When ν_z is large, shape and color are strongly entangled and α is low (see Fig. 1A). Similarly, we define $\nu_x := d_{SC}^{(x)}/d_S^{(x)}$ and $\nu_y := d_{SC}^{(y)}/d_S^{(y)}$ as the entanglement ratios in the input and targets.

3.1 Why study the abstraction score?

Here we use our setup to motivate why the abstraction score α is a useful quantity to understand in practice. We provide simple examples to illustrate how the α of hidden activations influences the effectiveness of two common interpretability methods: (i) linear probes and (ii) activation steering.

(i) Generalization of linear probes. It is common practice to detect concepts, e.g. truthfulness [11], using linear probes trained on a limited dataset. What determines whether these probes generalize to contexts beyond the dataset? In our setting, consider a probe trained to classify "shape" from only red activations. For intuition, assume $p(\boldsymbol{z} \mid s,c)$ is Gaussian with covariance $\sigma^2 \boldsymbol{I}$ and the probe is Bayes-optimal. We then ask: how well does this probe generalize to unseen blue samples?[6] Writing Φ for the standard normal CDF, the generalization error is monotonically decreasing in α (see §A.5):

$$P_{\text{error}}^{\text{red}\to\text{blue}} = \Phi\left(-\frac{\|\boldsymbol{u}_{\text{blue}}\|}{2\sigma} \cdot \alpha\right), \qquad \text{and in particular,} \quad \frac{\partial P_{\text{error}}^{\text{red}\to\text{blue}}}{\partial \alpha} < 0, \tag{8}$$

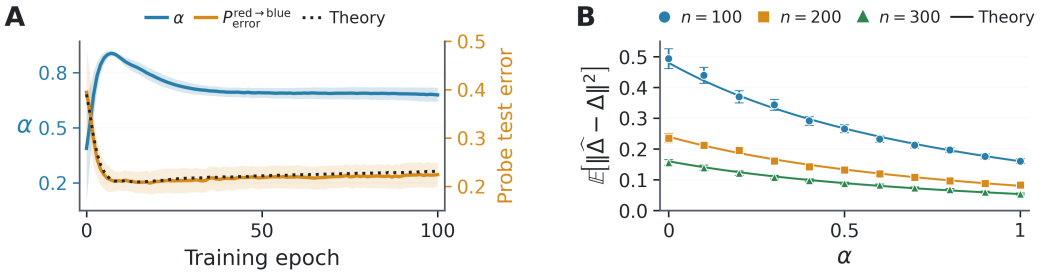

Figure 2: **Motivating the abstraction score. (A)** We test the predictions of 3.1.(i) by training a deep ReLU network on synthetic data (see §E.1.1). The empirical test error of a shape probe transferred from red samples to blue samples (orange) appears well-approximated by (8) (black dotted), which illustrates how α influences probe generalization error. **(B)** We test the predictions of 3.1.(ii) using synthetic Gaussian data (see §E.1.2). As α increases, the empirical error of estimating the steering vector (markers) decreases, and this appears well-approximated by (9) (lines).

[5] This is the Klein four-group: the smallest non-cyclic group, and smallest in the family $(\mathbb{Z}_2)^K$ where α can be defined.

[6] In neuroscience, this is also known as the cross-condition generalization performance [6].

(ii) Estimation of steering vectors from finite samples. In activation steering, one often estimates a "steering vector" by computing the difference-of-means from a small contrastive dataset. What determines the accuracy of these estimates? In our setting, consider estimating a shape steering vector for $\circ \to \square$ using n independent samples per shape. For intuition, assume $p(\mathbf{z} \mid s, c)$ is Gaussian with mean $\mathbb{E}[\mathbf{z} \mid s, c]$ in (6) and covariance $\sigma^2 \mathbf{I}_D$. Denote the "true" population steering vector Δ, whereas the estimated steering vector $\widehat{\Delta}$ is the empirical difference-of-means between sampled squares and circles. Then the expected error of the estimate is monotonically decreasing in α (see §A.6):

$$\mathbb{E}\left[\left\|\widehat{\Delta} - \Delta\right\|^2\right] = \frac{2}{n}\left[q_S \frac{1-\alpha}{1+\alpha} + q_C + D\sigma^2\right]. \tag{9}$$

4 Solutions to the minimal model

Using Assum. 1, we can rewrite the population loss $\mathcal{L}(\boldsymbol{W}_r^*, \boldsymbol{W})$ in (2) as a reduced objective $\mathcal{F}(\boldsymbol{W})$, and the representation's gradient flow is $\dot{\boldsymbol{z}} = -\nabla_{\boldsymbol{W}} \mathcal{F}\, \boldsymbol{x}$. Under Assum. 2, the dynamics of the Fourier energies q_m then decouple across modes with instantaneous growth rates $\kappa_m(q_m)$ (see §B.1),

$$\dot{q}_m = 2q_m \kappa_m(q_m), \qquad \kappa_m(q_m) := \gamma d_m^{(x)} d_m^{(y)} \Big/ \left[1 + \gamma\left(1 + \sigma_x^2 / d_m^{(x)}\right) q_m\right]^2 \tag{10}$$

Differentiating $\alpha = (q_S - q_{SC})/(q_S + q_{SC})$ then gives the following ODE for the abstraction score:

$$\dot{\alpha} = (1-\alpha^2)\big[\kappa_S(q_S) - \kappa_{SC}(q_{SC})\big]. \tag{11}$$

So α grows when the energy of the shape mode q_S grows faster relative to the interaction mode q_{SC}. We can then solve (11) to obtain the exact trajectory for α (see §B.2):

Theorem 1 (Exact solution for abstraction trajectory)**.** *Define $\tilde{q}_m := \gamma(1 + \sigma_x^2/d_m^{(x)})q_m$ and the invertible function $F(a) := \frac{1}{2}a^2 + 2a + \log a$ for $a > 0$. Then we have the following exact explicit solution for $\alpha(\tau)$ in terms of the rescaled learning time $\tau := 2\gamma d_S^{(x)} d_S^{(y)} t$:*

$$\alpha(\tau) = \tanh\left[\frac{1}{2}\log \frac{F^{-1}(F(\tilde{q}_{S,0}) + \tau)}{B F^{-1}(F(\tilde{q}_{SC,0}) + \nu_x \nu_y \tau)}\right], \qquad B = \frac{1 + \sigma_x^2/d_S^{(x)}}{1 + \sigma_x^2/d_{SC}^{(x)}} \tag{12}$$

Equipped with our core result Theorem 1, we can next derive analytic laws for the trajectory of α. For simplicity, in the rest of this paper we take $\sigma_x \to 0$, but all results can be generalized for $\sigma_x > 0$.

4.1 What determines the endpoint and trajectory of abstraction?

Proposition 2 (Terminal abstraction law)**.** *Define the terminal abstraction $\alpha_\infty := \lim_{t\to\infty} \alpha(t)$. For non-zero initializations, the terminal abstraction follows the law:*

$$\alpha_\infty = \frac{1 - \sqrt{\nu_x \nu_y}}{1 + \sqrt{\nu_x \nu_y}}. \tag{13}$$

This result shows that α_∞ is determined solely by the geometric mean of the input and target entanglement ratios. In particular, terminal abstraction will be perfect when *either* the inputs *or* targets are fully disentangled. Note that choosing targets $\boldsymbol{y}$ to be linear functions of the latents is equivalent to choosing $\nu_y = 0$, giving $\alpha_\infty = 1$. Thus Proposition 2 can recover both previous known results for perfect abstraction at the end of learning [3, 12]. To understand the *trajectory* rather than just the endpoint, we restrict our attention to a common setting of the dataset geometry:

Setting 1 (Low complexity regime)**.** *Denote α_x, α_y as the abstraction score of the inputs and targets respectively. A dataset is in this regime when $\alpha_\infty > 0 \iff \nu_x \nu_y < 1 \iff \alpha_x + \alpha_y > 0$.*

We use "low complexity" in the Boolean analysis sense of low-frequency or low-degree Fourier structure in the input-target correlations [38]. We conjecture this is a mild condition in standard supervised learning settings where targets can be thought of as handcrafted to be low complexity functions of latents of interest. It is less clear whether this holds in self-supervised settings, but recent work has shown transformers are biased towards learning functions with low complexity [39, 40].

In this regime, for $\alpha(0) < \alpha_\infty$, we can show that $\alpha(t)$ overshoots α_∞ (see §B.4). Unlike terminal abstraction, the *maximum abstraction* reached during training (i.e. the size of the peak) depends on the initialization scale of weights. Next, we analyze how initialization affects the maximum abstraction and link it to rich and lazy learning [31, 26] (see §B.5).

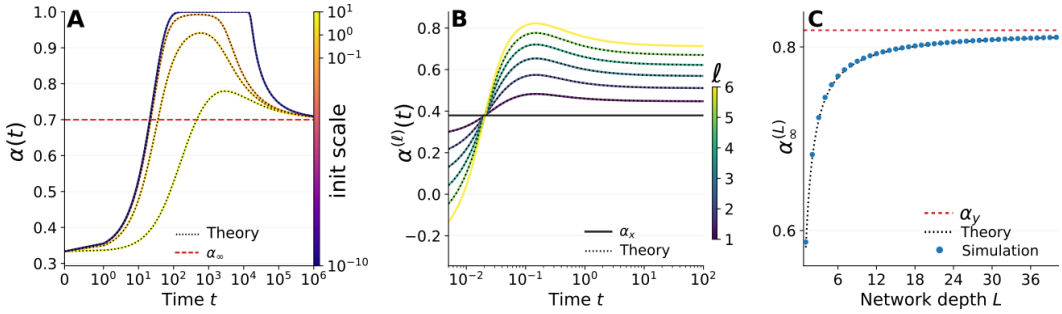

Figure 3: **Analytic laws governing abstraction dynamics in the minimal model**. All solid colored lines correspond to numerical simulations. **(A)** shows trajectory of $\alpha(t)$ in an $L = 1$ hidden layer network for varying weight initialization. For all initializations, $\alpha(t)$ converges to the red terminal value Proposition 2. Meanwhile, the maximum abstraction reached over the trajectory decreases with initialization scale (Proposition 3). For very small initialization, $\alpha(t)$ reaches and stays near 1 for a long time. **(B)** Abstraction interpolates layerwise in $L = 6$ layer network, where ℓ is the layer index; lighter colors correspond to deeper layers. **(C)** Terminal abstraction in final layer increases with network depth L, where the red limit is given by α_y.

Proposition 3 (Initialization scale controls maximum abstraction)**.** *Assume $0 < \nu_x\nu_y < 1$ and $\alpha_0 < \alpha_\infty$. Define the maximum abstraction $\alpha_{\max}(\zeta) := \sup_{t>0} \alpha(t)$. Fix $\nu_{z,0} := q_{SC,0}/q_{S,0}$ and consider the family of initializations $\tilde{q}_{S,0} = \zeta$, $\tilde{q}_{SC,0} = \zeta\,\nu_{z,0}$ for $\zeta > 0$. Then $\alpha_{\max}(\zeta)$ is **strictly decreasing in the initialization scale** ζ and in particular,* $\lim_{\zeta\downarrow 0} \alpha_{\max}(\zeta) = 1$.

Although the dataset controls terminal abstraction, the initialization scale controls the maximum reached over its trajectory (see Fig. 3A). In the rich limit, the model comes arbitrarily close to perfect abstraction even if α_∞ is low. In the lazy limit, trajectory monotonically approaches α_∞. The time spent near perfect abstraction diverges as $\zeta \downarrow 0$, as we prove in §B.6:

Proposition 4 (Rich-limit metastability)**.** *Assume $0 < \nu_x\nu_y < 1$ and $\alpha_0 < \alpha_\infty$. Consider the initialization family from Proposition 3. Fix any near-perfect abstraction threshold $1 - \varepsilon$ strictly above the terminal value, i.e. $0 < \varepsilon < 1 - \alpha_\infty$. Let $T_\varepsilon^{\text{in}}(\zeta)$ be the first time at which $\alpha(t;\zeta) \geq 1 - \varepsilon$, and let $R_\varepsilon(\zeta)$ be the total time spent in this band. Then, as $\zeta \downarrow 0$, we have:*

$$T_\varepsilon^{\text{in}}(\zeta) = O(1), \quad R_\varepsilon(\zeta) = \Theta\left(\log \frac{1}{\zeta}\right), \quad \textit{and in particular,}\ \ R_\varepsilon(\zeta) \to \infty \tag{14}$$

Near-perfect abstraction is ***metastable*** in the rich limit: after entry into a fixed near-perfect band, it persists over any fixed remaining training horizon as $\zeta \downarrow 0$, even when the terminal value is lower. Note the following limits do not commute: $\lim_{\zeta\downarrow 0} \lim_{t\to\infty} \alpha(t;\zeta) = \alpha_\infty$, whereas $\lim_{t\to\infty} \lim_{\zeta\downarrow 0} \alpha(t;\zeta) = 1$.

4.2 How does depth impact abstraction?

We now extend our model to a deep linear network with L hidden layers. Let the superscript $\ell \in \{1, \ldots, L\}$ denote the layer, so we have $q_m^{(\ell)}(t)$ at each layer and $\alpha^{(\ell)}(t)$ is the layer's abstraction. To derive the dynamics for arbitrary depth, we assume a mild setting similar to [41, 26]:

Setting 2 (Balanced initialization)**.** *For each Fourier mode m, all hidden layers initially multiply its norm by the same positive gain $g_m(0)$. Thus $q_m^{(\ell)}(0) = d_m^{(x)} g_m(0)^{2\ell}$.*

Under Setting 2 we derive ODE for the representation energies of the final hidden layer (see §B.7):

$$\dot{q}_m = 2L\gamma d_m^{(y)} \left(d_m^{(x)}\right)^{1/L} \frac{(q_m)^{2-\frac{1}{L}}}{(1+\gamma q_m)^2} \tag{15}$$

We then ask: for a deep network with L hidden layers in this setting, how does abstraction differ across layers? In §B.7, we derive the following answer:

Proposition 5 (Layerwise interpolation and depth)**.** *Suppose Setting 2 holds, with $d_S^{(x)}, d_{SC}^{(x)}, d_S^{(y)}$, $d_{SC}^{(y)} > 0$ and positive initial gains. The abstraction at layer $\ell \in \{1, \ldots, L\}$ obeys the following*

interpolations in arctanh *space (also see Fig. 3B):*

$$\operatorname{arctanh}\alpha^{(\ell)}(t) = (1-\ell/L)\operatorname{arctanh}\alpha_x + (\ell/L)\operatorname{arctanh}\alpha^{(L)}(t) \tag{16}$$

$$\operatorname{arctanh}\alpha_\infty^{(L)} = \frac{1}{L+1}\operatorname{arctanh}\alpha_x + \frac{L}{L+1}\operatorname{arctanh}\alpha_y \tag{17}$$

Next we ask: how does L affect the final layer's terminal abstraction? By analytically continuing $L \in \mathbb{R}_+$ and differentiating (17), we obtain Corollary B.4 which shows that the final layer's terminal abstraction is monotonically increasing in depth if $\alpha_y > \alpha_x$ (see Fig. 3C).

4.3 Experiments

To evaluate whether our theory's predictions hold in practice, we train small ResNets on a task derived from the 3dshapes dataset [13]. Despite major departures from our assumptions, the results are qualitatively consistent with key predictions of our theory (see Fig. 4).

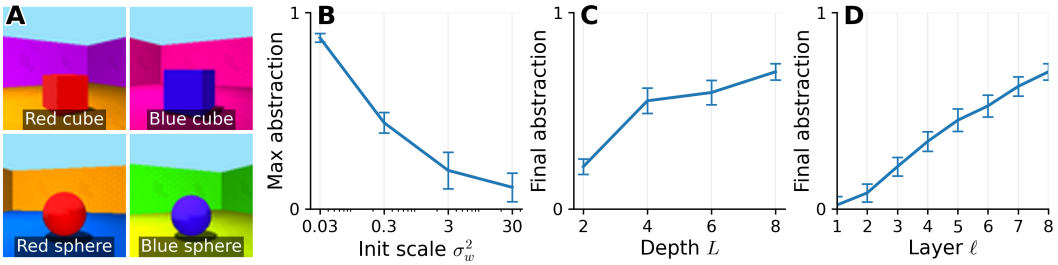

Figure 4: **Empirical evaluation of the theory in convolutional networks.** See §E.3 for experiment details. **(A)** We train small ResNets on a task constructed from the 3dshapes dataset [13] using binary shape and color latents, with $n = 1000$ examples per class while nuisance variables (floor color, wall color, size, orientation) vary randomly. For panels B-D, error bars show SEM across 20 seeds. **(B)** Maximum abstraction in last layer decreases with weight initialization scale σ_w^2, in agreement with Proposition 3. **(C)** Final abstraction after training in last layer increases with network depth L, in agreement with Proposition 5. **(D)** Final abstraction after training increases layerwise in a depth $L = 8$ network, in agreement with Proposition 5 (deeper layers are more abstract than early layers).

5 Extending the theory to nonlinear networks

In this section we extend our theory to two-layer nonlinear networks in the infinite-width limit and study which aspects of our theory qualitatively survive or change. We now distinguish between the preactivations $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$ and the activations $\boldsymbol{h} = \phi(\boldsymbol{z})$, where ϕ is a nonlinear activation function.

Define $q_m = D^{-1}\|\hat{\boldsymbol{z}}_m\|^2$ and $k_m = D^{-1}\|\hat{\boldsymbol{h}}_m\|^2$ as the width-normalized Fourier energies of the preactivations and features. We take $D \to \infty$ and assume $\hat{z}_{m,i} \overset{\text{iid}}{\sim} \mathcal{N}(0, q_m)$ for each Fourier mode m and neuron i. For a pair of classes write $g = (ss', cc')$. Their scalar preactivations have covariance r_g and common variance v. Define the *covariance map* $\psi_\phi(r, v) = \mathbb{E}[\phi(U)\phi(V)]$ for a centered Gaussian pair (U, V) with covariance r and common variance v. By law of large numbers, we have:

$$k_m = \frac{1}{4}\sum_{g\in\{\pm1\}^2}\chi_m(g)\psi_\phi(r_g, v), \quad r_g = \sum_{m\in\mathcal{M}} q_m\chi_m(g), \quad v = \sum_m q_m. \tag{18}$$

Thus the nonlinear feature geometry is still described by four Fourier energies. Using this, we next study the effect of the nonlinearity ϕ pointwise at each time, and on dynamics.

5.1 Instantaneous effect of nonlinearities on abstraction

The major difference to the linear network case is that we now have two abstraction scores: one for preactivations $\alpha_{\text{pre}} = (q_S - q_{SC})/(q_S + q_{SC})$ and one for activations $\alpha_{\text{post}} = (k_S - k_{SC})/(k_S + k_{SC})$. Here, we ask a simple question: can applying a nonlinearity ϕ to a representation improve abstraction? ***Strikingly, the answer is no for virtually all nonlinearities.*** Leveraging classical results from maximal correlation theory [42, 43], we obtain the following result:

Theorem 6 (Attenuation law, informal)**.** *Assume $\alpha_{\text{pre}} > 0$. Then, for any square-integrable $\phi : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E}[\phi(\sqrt{v}Z)^2] < \infty$, $Z \sim \mathcal{N}(0, 1)$ we have,*

$$\alpha_{\text{post}} \leq \alpha_{\text{pre}}. \tag{19}$$

See §C.6, Theorem C.4 for the formal statement and proof. In other words, applying a nonlinear activation function to a representation cannot improve, and generally weakens, abstraction. This holds for a broad class of functions, including standard choices such as ReLU, GELU, SiLU, tanh. Given §3.1, this result has implications for finding linear concept directions in interpretability, and in §6 we demonstrate one potential application of this in practice.

5.2 Dynamical effect of nonlinearities on abstraction

We next seek to understand how nonlinearities affect the dynamics of abstraction during learning. In our setting, we can generalize our results (10)-(11) to the preactivation dynamics as follows:

$$\dot{q}_m = 2q_m\kappa^{\phi}_m, \qquad \dot{\alpha}_{\text{pre}} = (1 - \alpha^2_{\text{pre}})(\kappa^{\phi}_S - \kappa^{\phi}_{SC}), \tag{20}$$

$$\kappa^{\phi}_m(\boldsymbol{q}) = \gamma d^{(x)}_m \sum_{m' \in \mathcal{M}} \frac{d^{(y)}_{m'} J_{m',m}(\boldsymbol{q})}{[1 + \gamma k_{m'}(\boldsymbol{q})]^2} \qquad J_{m',m}(\boldsymbol{q}) \coloneqq \frac{\partial k_{m'}}{\partial q_m}, \tag{21}$$

where $k_{m'}$ depends on ψ_ϕ in (18). For a linear network, $k_m = q_m$ and $J_{m',m} = \delta_{m',m}$, so each mode responds only to its corresponding target component. Introducing a nonlinearity can couple these responses through the Jacobian J such that even a Fourier mode absent from the targets can grow.

We next analytically study terminal abstraction by specializing to two nonlinearities: the error function $\phi_\beta(z) = \text{erf}(\beta z)$ for $\beta > 0$ and leaky ReLU (L-ReLU) $\phi_\omega(z) = \max(z, \omega z)$ for $\omega \in [0, 1]$. Using results on NNGP kernels [44, 45], we can obtain closed forms for the covariance maps in (18):

$$\psi_\phi(r, v) = \begin{cases} \dfrac{2}{\pi} \arcsin\left(\dfrac{2\beta^2 r}{1 + 2\beta^2 v}\right), & \phi = \phi_\beta \text{ (erf)}, \\ \dfrac{1 + \omega^2}{2} r + \dfrac{(1 - \omega)^2}{2\pi}\left[\sqrt{v^2 - r^2} - r \arccos(r/v)\right], & \phi = \phi_\omega \text{ (L-ReLU)}. \end{cases} \tag{22}$$

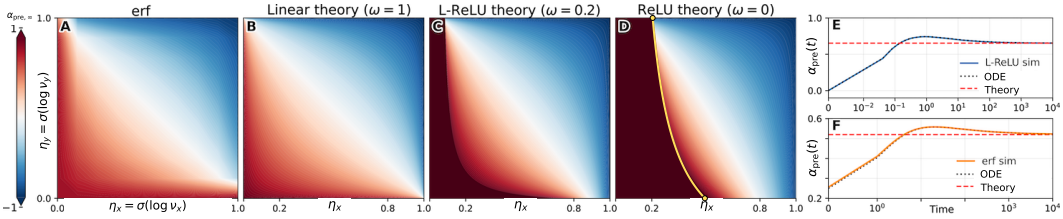

Figure 5: **Effect of nonlinearities on abstraction**. Panels A-D show phase diagrams of $\alpha_{\text{pre},\infty}$ as a function of the entanglement fractions $\eta_x \coloneqq \sigma(\log \nu_x)$ and $\eta_y \coloneqq \sigma(\log \nu_y)$ for inputs and targets. **(A)** Erf abstraction at the end of the simulated training horizon ($\beta = 1, \gamma = 1000$) resembles the linear reference. **(B)** Linear theory from Proposition 2. **(C-D)** L-ReLU terminal abstraction under Setting 3 evaluated by solving (25). Decreasing ω warps the shape to create a region of perfect abstraction which is more sensitive to inputs than targets. For ReLU ($\omega = 0$), in yellow we plot the analytical solution for the boundary of this region. **(E)** Abstraction trajectory of L-ReLU network ($\omega = 0.5$), with red line evaluated by solving (25). **(F)** Abstraction trajectory of erf network ($\beta = 0.1, \gamma = 1000$) where red line is from the linear network theory (Proposition 2).

(a) Erf. For erf, we can obtain good intuition by Taylor expanding (22) in $\beta^2 v$, which gives:

$$\psi_{\phi_\beta}(r, v) = \frac{4\beta^2}{\pi} r + \mathcal{O}(\beta^4 v^2), \qquad J_{m',m}(\boldsymbol{q}) = \frac{4\beta^2}{\pi} \delta_{m',m} + \mathcal{O}(\beta^4 v). \tag{23}$$

Substituting this into (21) shows that the erf network dynamics are, to leading order, those of a linear network with rescaled regularization $\gamma_{\text{eff}} = 4\beta^2\gamma/\pi$ (see §C.3). We plot a phase diagram of the simulated $\alpha_{\text{pre},\infty}$ for erf in Fig. 5A, and find that it closely resembles the linear theory in Fig. 5B. In Fig. 5F we also show that the linear theory prediction describes the late-time erf ODE trajectory.

(b) L-ReLU. For L-ReLU, we are able to obtain an implicit solution for $\alpha_{\mathrm{pre},\infty}$ by assuming:

Setting 3. *Assume* $d_S^{(x)} = d_C^{(x)}$, $d_S^{(y)} = d_C^{(y)}$, $q_S(0) = q_C(0)$ *are all positive, and* $d_\emptyset^{(y)} = q_\emptyset(0) = 0$.

Under this simplified setting, we can make a useful change of variables to the angle θ,

$$\alpha_{\mathrm{pre},\infty}(\theta) = -\frac{1+3\cos\theta}{3+\cos\theta}, \qquad \theta = \arccos\left(\frac{\nu_z - 2}{\nu_z + 2}\right). \tag{24}$$

We then obtain an implicit solution for $\alpha_{\mathrm{pre},\infty}$ by solving the following scalar equation for θ:

$$\nu_x \nu_R(\theta; \omega, \nu_y) = 1, \tag{25}$$

where ν_R has an explicit expression provided in §C.4. We plot this implicit solution in the phase diagrams Fig. 5C–D, which reveal a striking difference from the linear theory: as the leak ω decreases (i.e. we approach full ReLU), the phase diagram shifts to create a region of perfect abstraction. In §C.4 we derive the analytical solution for the boundary of this region in the ReLU case ($\omega = 0$), which we plot in yellow in Fig. 5D:

$$\nu_y(\nu_x) = \frac{2r^2(1-\nu_x)}{(1+r)\nu_x - r}, \qquad r = 1 - \frac{2}{\pi}. \tag{26}$$

This boundary is more sensitive to inputs than targets since $d\eta_y/d\eta_x$ is always strictly below -1, which is consistent with previous empirical observations that ReLU networks are more sensitive to input structure than target structure compared to sigmoid-like networks [23].

6 Applications

Application to transformer interpretability. Theorem 6 and (8) motivate a practical prediction: concept/steering vectors extracted via linear probes should be more abstract and generalize better after ablating nonlinearities. To operationalize this, we introduce the **local GELU ablation** procedure: (1) Measure abstraction in residual stream activations at layer ℓ to get a baseline; (2) in the MLP block immediately preceding the activations, replace GELU with the identity; (3) rerun the forward pass and measure abstraction in the modified activations. We test local GELU ablation in both visual and language concepts in frontier open transformer models (see Fig. 6). Refer to §E.4 for details.

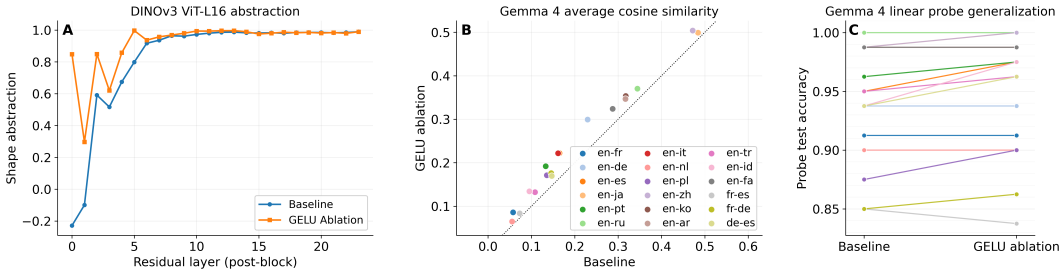

Figure 6: **Local GELU ablation improves concept abstraction and linear probe generalization in transformers**. We apply Theorem 6 to transformers by using the local GELU ablation technique described in §6. **(A)** For visual concepts, we evaluate DINOv3 ViT-L/16 [46] on the 3dshapes task described in §4.3. Local GELU ablation improves or preserves shape abstraction in ***every layer*** compared to baseline. Also, baseline abstraction improves layerwise which agrees with Proposition 5. **(B)** For language concepts, we analyze the final residual stream of Gemma 4 E2B [47] evaluated on 18 bilingual concepts (e.g. `English-Spanish`), each instantiated by 80 word pairs (e.g. `Man, Hombre`). Local GELU ablation ***improves abstraction for all 18 bilingual concepts*** (i.e. all points lie above diagonal). **(C)** The same intervention improves or preserves linear probe generalization for 17/18 concepts, with an average accuracy improvement of $\sim$1 percentage point.

Application to neuroscience. Although our theory was derived for artificial networks, Proposition 5 suggests an intriguing prediction: representations may be more abstract in deeper / higher-level brain regions which are closer to downstream behaviour and learning signals. As a proof-of-concept, we perform a simple experiment using publicly available data from the macaque visual cortex [48] and find qualitative agreement with this prediction (see Fig. 7). See §E.5 for discussion and details.

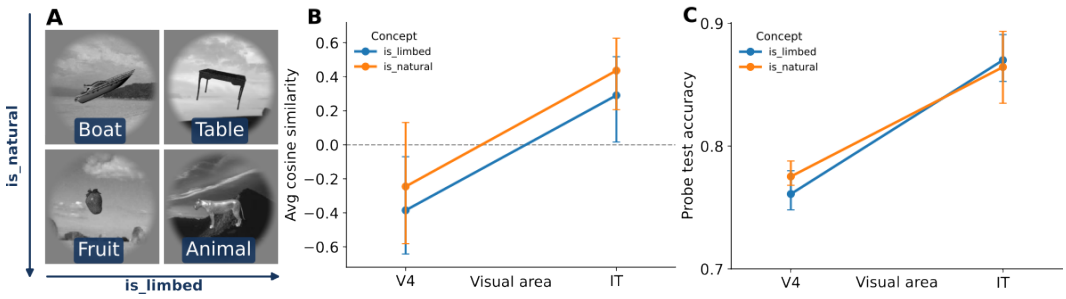

Figure 7: **Abstraction increases along macaque ventral visual stream.** We test a speculative prediction of Proposition 5 that deeper cortical layers exhibit greater abstraction scores using a subset of recordings in macaque V4 and IT from the public Majaj-Hong dataset [48]. **(A)** We construct a 2×2 factor design around the concepts of *naturalness* and *limbedness*. **(B)** In a pooled site-matched analysis, abstraction scores for both concepts are significantly higher in IT than in V4 (error bars are 95% bootstrap intervals). **(C)** Linear probes for both concepts generalize significantly better in IT.

7 Generalizing beyond two latent variables

In this paper we focused on the case of $K = 2$ latent variables because it is the simplest setting in which the abstraction score α can be defined, making the dynamics particularly instructive. Here, we end our paper by briefly illustrating how our theory can be generalized to representations of $K > 2$ latents denoted $\boldsymbol{s} = (s_1, \ldots, s_K) \in \{-1, +1\}^K$. For any nonempty proper subset $i \subsetneq [K]$, using Theorem A.2 we can define a concept of interest using the Fourier character $\chi_i(\boldsymbol{s}) = \prod_{a \in i} s_a$. Denote the set of interaction modes containing i with $\mathcal{A}_i = \{j \subseteq [K] : i \subsetneq j\}$, where the number of such interaction modes is $|\mathcal{A}_i| = 2^{K-|i|} - 1$. Then, by extending Assum. 2 to all 2^K Fourier modes, the generalized expression for the abstraction score of concept i is (see Appendix D):

$$\alpha_i = \frac{q_i - \frac{1}{|\mathcal{A}_i|} \sum_{j \in \mathcal{A}_i} q_j}{q_i + \sum_{j \in \mathcal{A}_i} q_j}, \qquad q_j = \|\hat{\boldsymbol{z}}_j\|^2. \tag{27}$$

As we can see, the basic structure of the abstraction score remains the same for $K > 2$: it compares the Fourier energy of the concept i with the total energy of all its higher-order interactions.

An interesting consequence of this generalization is that abstraction acquires a *nested structure*: An interaction mode that entangles lower-order concepts can itself have a well-defined abstraction score which is determined by its relationship with even higher-order interactions. Thus an interaction mode can be perfectly abstract even if its constituent latents are not.

8 Discussion

In this paper we developed a dynamical theory of abstraction in deep networks. We started in a minimal linear network setting which we fully solved and extended to nonlinear networks. Throughout, we provided evidence that our theory captures nontrivial phenomena in realistic settings and demonstrated applications of our theory to both deep learning and neuroscience.

Future work. Promising future directions include (i) relaxation of our theoretical assumptions; (ii) broader testing and application of the local GELU ablation method; and (iii) in-depth study of the theory's implications for interpretability and control of LLMs.

Limitations. Key limitations of our theory include (i) the simplifying assumptions made in §3 and (ii) the difficulty of extending our theory to deep nonlinear networks. Despite these, experimental results in realistic settings qualitatively support our theoretical predictions.

Acknowledgments and Disclosures

Funding: The authors received no external funding or third-party support for this work.
Competing Interests: The authors declare no competing interests.

References

[1] Neel Nanda, Andrew Lee, and Martin Wattenberg. Emergent Linear Representations in World Models of Self-Supervised Sequence Models, September 2023. arXiv preprint arXiv:2309.00941.

[2] Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models, July 2024. arXiv preprint arXiv:2311.03658.

[3] Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the Origins of Linear Representations in Large Language Models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 21879–21911. PMLR, July 2024.

[4] Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The Geometry of Categorical and Hierarchical Concepts in Large Language Models, February 2025. arXiv preprint arXiv:2406.01506.

[5] Tomas Mikolov, Wen-tau Yih, and Geoffrey Zweig. Linguistic Regularities in Continuous Space Word Representations. In Lucy Vanderwende, Hal Daumé III, and Katrin Kirchhoff, editors, *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, Georgia, June 2013. Association for Computational Linguistics.

[6] Silvia Bernardi, Marcus K. Benna, Mattia Rigotti, Jérôme Munuera, Stefano Fusi, and C. Daniel Salzman. The Geometry of Abstraction in the Hippocampus and Prefrontal Cortex. *Cell*, 183(4):954–967.e21, November 2020.

[7] Ramon Nogueira, Chris C. Rodgers, Randy M. Bruno, and Stefano Fusi. The geometry of cortical representations of touch in rodents. *Nature Neuroscience*, 26(2):239–250, February 2023.

[8] Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. Finding Neurons in a Haystack: Case Studies with Sparse Probing, June 2023. arXiv preprint arXiv:2305.01610.

[9] Nicholas Goldowsky-Dill, Bilal Chughtai, Stefan Heimersheim, and Marius Hobbhahn. Detecting Strategic Deception with Linear Probes. In *Forty-Second International Conference on Machine Learning*, June 2025.

[10] Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering Language Models With Activation Engineering, October 2024. arXiv preprint arXiv:2308.10248.

[11] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-Time Intervention: Eliciting Truthful Answers from a Language Model, June 2024. arXiv preprint arXiv:2306.03341.

[12] Bin Wang, W. Jeffrey Johnston, and Stefano Fusi. A mathematical theory for understanding when abstract representations emerge in neural networks, March 2026. arXiv preprint arXiv:2510.09816.

[13] Hyunjik Kim and Andriy Mnih. Disentangling by Factorising. In *Proceedings of the 35th International Conference on Machine Learning*, pages 2649–2658. PMLR, July 2018.

[14] Runjin Chen, Andy Arditi, Henry Sleight, Owain Evans, and Jack Lindsey. Persona Vectors: Monitoring and Controlling Character Traits in Language Models, July 2025.

[15] Viktor Moskvoretskii, Dominik Glandorf, Jorge Medina Moreira, Tanja Käser, and Robert West. Tracing persona vectors through LLM pretraining, May 2026.

[16] Justin D. Shin, Wenbo Tang, and Shantanu P. Jadhav. Protocol for geometric transformation of cognitive maps for generalization across hippocampal-prefrontal circuits. *STAR Protocols*, 4(3):102513, September 2023.

[17] Valeria Fascianelli, Aldo Battista, Fabio Stefanini, Satoshi Tsujimoto, Aldo Genovesio, and Stefano Fusi. Neural representational geometries reflect behavioral differences in monkeys and recurrent neural networks. *Nature Communications*, 15(1):6479, August 2024.

[18] Hristos S. Courellis, Juri Minxha, Araceli R. Cardenas, Daniel L. Kimmel, Chrystal M. Reed, Taufik A. Valiante, C. Daniel Salzman, Adam N. Mamelak, Stefano Fusi, and Ueli Rutishauser. Abstract representations emerge in human hippocampal neurons during inference. *Nature*, 632(8026):841–849, August 2024.

[19] Lara M. Boyle, Lorenzo Posani, Sarah Irfan, Steven A. Siegelbaum, and Stefano Fusi. Tuned geometries of hippocampal representations meet the computational demands of social memory. *Neuron*, 112(8):1358–1371.e9, April 2024.

[20] Karyna Mishchanchuk, Gabrielle Gregoriou, Albert Qü, Alizée Kastler, Quentin J. M. Huys, Linda Wilbrecht, and Andrew F. MacAskill. Hidden state inference requires abstract contextual representations in the ventral hippocampus. *Science*, 386(6724):926–932, November 2024.

[21] Pia-Kelsey O'Neill, Lorenzo Posani, Jozsef Meszaros, Phebe Warren, Carl E. Schoonover, Andrew J. P. Fink, Stefano Fusi, and C. Daniel Salzman. The representational geometry of emotional states in basolateral amygdala, April 2024. bioRxiv preprint 2023.09.23.558668.

[22] Takuya Ito, Tim Klinger, Douglas H. Schultz, John D. Murray, Michael W. Cole, and Mattia Rigotti. Compositional generalization through abstract representations in human and artificial neural networks, September 2022. arXiv preprint arXiv:2209.07431.

[23] Matteo Alleman, Jack W. Lindsey, and Stefano Fusi. Task structure and nonlinearity jointly determine learned representational geometry, January 2024. arXiv preprint arXiv:2401.13558.

[24] Andrew M. Saxe, James L. McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks, February 2014. arXiv preprint arXiv:1312.6120.

[25] Andrew K. Lampinen and Surya Ganguli. An analytic theory of generalization dynamics and transfer learning in deep linear networks, January 2019. arXiv preprint arXiv:1809.10374.

[26] Daniel Kunin, Allan Raventós, Clémentine Dominé, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: Exact solutions reveal how unbalanced initializations promote rapid feature learning. *Advances in Neural Information Processing Systems*, 37:81157–81203, December 2024.

[27] Alexandra Maria Proca, Clémentine Carla Juliette Dominé, Murray Shanahan, and Pedro A. M. Mediano. Learning dynamics in linear recurrent neural networks. In *Forty-Second International Conference on Machine Learning*, June 2025.

[28] Yoonsoo Nam, Seok Hyeong Lee, Clementine C. J. Domine, Yeachan Park, Charles London, Wonyl Choi, Niclas Goring, and Seungjai Lee. Position: Solve Layerwise Linear Models First to Understand Neural Dynamical Phenomena (Neural Collapse, Emergence, Lazy/Rich Regime, and Grokking), May 2025. arXiv preprint arXiv:2502.21009.

[29] Jamie Simon, Daniel Kunin, Alexander Atanasov, Enric Boix-Adserà, Blake Bordelon, Jeremy Cohen, Nikhil Ghosh, Florentin Guth, Arthur Jacot, Mason Kamb, Dhruva Karkada, Eric J. Michaud, Berkan Ottlik, and Joseph Turnbull. There Will Be a Scientific Theory of Deep Learning, April 2026. arXiv preprint arXiv:2604.21691.

[30] Lukas Braun, Clémentine Dominé, James Fitzgerald, and Andrew Saxe. Exact learning dynamics of deep linear networks with prior knowledge. *Advances in Neural Information Processing Systems*, 35:6615–6629, December 2022.

[31] Clémentine C. J. Dominé, Nicolas Anguita, Alexandra M. Proca, Lukas Braun, Daniel Kunin, Pedro A. M. Mediano, and Andrew M. Saxe. From Lazy to Rich: Exact Learning Dynamics in Deep Linear Networks, March 2025. arXiv preprint arXiv:2409.14623.

[32] Daniel J. Korchinski, Dhruva Karkada, Yasaman Bahri, and Matthieu Wyart. On the Emergence of Linear Analogies in Word Embeddings, October 2025. arXiv preprint arXiv:2505.18651.

[33] Raphaël Barboni, Gabriel Peyré, and François-Xavier Vialard. Ultra-fast feature learning for the training of two-layer neural networks in the two-timescale regime, July 2025. arXiv preprint arXiv:2504.18208.

[34] Pierre Marion and Raphaël Berthier. Leveraging the two timescale regime to demonstrate convergence of neural networks, October 2023. arXiv preprint arXiv:2304.09576.

[35] David Saad and Sara A. Solla. On-line learning in soft committee machines. *Physical Review E*, 52(4):4225–4243, October 1995.

[36] Sanjeev Arora, Simon S. Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-Grained Analysis of Optimization and Generalization for Overparameterized Two-Layer Neural Networks, May 2019. arXiv preprint arXiv:1901.08584.

[37] Ryan O'Donnell. Analysis of Boolean Functions, May 2021.

[38] Greg Yang and Hadi Salman. A fine-grained spectral perspective on neural networks. *arXiv preprint arXiv:1907.10599*, 2019.

[39] Bhavya Vasudeva, Deqing Fu, Tianyi Zhou, Elliott Kau, Youqi Huang, and Vatsal Sharan. Transformers learn low sensitivity functions: Investigations and implications. In *The Thirteenth International Conference on Learning Representations*, 2025.

[40] Satwik Bhattamishra, Arkil Patel, Varun Kanade, and Phil Blunsom. Simplicity bias in Transformers and their ability to learn sparse Boolean functions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5767–5791, Toronto, Canada, July 2023. Association for Computational Linguistics.

[41] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A Convergence Analysis of Gradient Descent for Deep Linear Neural Networks, October 2019. arXiv preprint arXiv:1810.02281.

[42] Yaming Yu. On the maximal correlation coefficient. *Statistics & Probability Letters*, 78(9):1072–1075, 2008.

[43] Mark Veraar. Correlation inequalities and applications to vector-valued Gaussian random variables and fractional Brownian motion. *Potential Analysis*, 30(4):341–370, May 2009.

[44] Youngmin Cho and Lawrence Saul. Kernel Methods for Deep Learning. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc., 2009.

[45] Christopher Williams. Computing with Infinite Networks. In *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996.

[46] Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3, August 2025. arXiv preprint arXiv:2508.10104.

[47] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, Pier Giuseppe Sessa, Aakanksha Chowdhery, Adam Roberts, Aditya Barua, Alex Botev, Alex Castro-Ros, Ambrose Slone, Amélie Héliou, Andrea Tacchetti, Anna Bulanova, Antonia Paterson, Beth Tsai, Bobak Shahriari, Charline Le Lan, Christopher A. Choquette-Choo, Clément Crepy, Daniel Cer, Daphne Ippolito, David Reid, Elena Buchatskaya, Eric Ni, Eric Noland, Geng Yan, George Tucker, George-Christian Muraru, Grigory Rozhdestvenskiy, Henryk Michalewski, Ian Tenney, Ivan Grishchenko, Jacob Austin, James Keeling, Jane Labanowski, Jean-Baptiste Lespiau, Jeff Stanway, Jenny Brennan, Jeremy Chen, Johan Ferret, Justin Chiu, Justin Mao-Jones, Katherine Lee, Kathy Yu, Katie Millican, Lars Lowe Sjoesund, Lisa Lee, Lucas Dixon, Machel Reid, Maciej Mikuła, Mateo Wirth, Michael Sharman, Nikolai Chinaev, Nithum Thain, Olivier Bachem, Oscar Chang, Oscar Wahltinez, Paige Bailey, Paul Michel, Petko Yotov, Rahma Chaabouni, Ramona Comanescu, Reena Jana, Rohan Anil, Ross McIlroy, Ruibo Liu, Ryan Mullins, Samuel L. Smith, Sebastian Borgeaud, Sertan Girgin, Sholto Douglas, Shree Pandya, Siamak Shakeri, Soham De, Ted Klimenko, Tom Hennigan, Vlad Feinberg, Wojciech Stokowiec, Yu-hui Chen, Zafarali Ahmed, Zhitao Gong, Tris Warkentin, Ludovic Peran, Minh Giang, Clément Farabet, Oriol Vinyals, Jeff Dean, Koray Kavukcuoglu, Demis Hassabis, Zoubin Ghahramani, Douglas Eck, Joelle Barral, Fernando Pereira, Eli Collins, Armand Joulin, Noah Fiedel, Evan Senter, Alek Andreev, and Kathleen Kenealy. Gemma: Open Models Based on Gemini Research and Technology, April 2024. arXiv preprint arXiv:2403.08295.

[48] Najib J. Majaj, Ha Hong, Ethan A. Solomon, and James J. DiCarlo. Simple Learned Weighted Sums of Inferior Temporal Neuronal Firing Rates Accurately Predict Human Core Object Recognition Performance. *Journal of Neuroscience*, 35(39):13402–13418, September 2015.

[49] Carl Allen and Timothy Hospedales. Analogies Explained: Towards Understanding Word Embeddings, May 2019. arXiv preprint arXiv:1901.09813.

[50] Dhruva Karkada, James B. Simon, Yasaman Bahri, and Michael R. DeWeese. Closed-Form Training Dynamics Reveal Learned Features and Linear Structure in Word2Vec-like Models, February 2025.

[51] Dhruva Karkada, Daniel J. Korchinski, Andres Nava, Matthieu Wyart, and Yasaman Bahri. Symmetry in language statistics shapes the geometry of model representations, February 2026.

[52] Gene Golub and Victor Pereyra. Separable nonlinear least squares: The variable projection method and its applications. *Inverse Problems*, 19(2):R1, February 2003.

[53] Cheol-Taek Kim and Ju-Jang Lee. Training Two-Layered Feedforward Networks With Variable Projection Method. *IEEE Transactions on Neural Networks*, 19(2):371–375, February 2008.

[54] Jack W Lindsey and Elias B Issa. Factorized visual representations in the primate visual system and deep neural networks. *eLife*, 13:RP91685, July 2024.

Appendix

Appendix Contents

A Supplementary discussion **18**

A.1 Further related work . . . 18

A.2 Relationship to the linear representation hypothesis . . . 18

A.3 Fourier representation and orthogonality . . . 20

A.4 Optimal readout and variable projection . . . 24

A.5 Abstraction and generalization of linear probes . . . 26

A.6 Estimation of steering vectors from finite samples . . . 27

B Population derivations for linear networks **29**

B.1 Population gradient flow and decoupled Fourier modes . . . 29

B.2 Exact solutions for the Fourier energies and abstraction . . . 31

B.3 Early-time growth, terminal abstraction, and boundary cases . . . 32

B.4 Non-monotonicity and overshoot . . . 33

B.5 Initialization scale . . . 35

B.6 Rich-limit residence-time metastability . . . 36

B.7 Depth dynamics . . . 37

C Nonlinear networks: Fourier geometry and learning dynamics **42**

C.1 Fourier coordinates and nonlinear feature geometry . . . 42

C.2 Population learning dynamics under Gaussian closure . . . 43

C.3 Erf in the weakly nonlinear regime . . . 45

C.4 Leaky-ReLU dynamics and terminal abstraction . . . 46

C.5 Magnitude attenuation for erf and leaky ReLU . . . 49

C.6 Magnitude attenuation for square-integrable activations . . . 51

C.7 Normalization and prescribed offsets in the simulations . . . 52

D Generalization to multiple latent variables **53**

D.1 Fourier modes and context-dependent concept directions . . . 53

D.2 Abstraction as a ratio of Fourier energies . . . 53

E Experimental details **55**

E.1 Experimental details for motivating examples . . . 55

E.2 Experimental details for theory-simulation agreement figures . . . 55

E.3 Experimental details for 3dshapes experiments . . . 57

E.4 Experimental details for Gemma 4 ablation experiments . . . 58

E.5 Experimental details for macaque ventral-stream analysis . . . 60

F Other **61**

F.1 Compute resources 61
F.2 Assets and licenses 61

A Supplementary discussion

A.1 Further related work

Below is a further list of related work.

Linear analogies in word embeddings. Abstraction is closely related to the linear analogy structures found in word embeddings such as Word2Vec [5]. Several theories investigate how such structures arise [49, 50, 32, 51]. Recently, [50] took a learning-dynamics approach, deriving gradient-flow solutions for a Word2Vec-like embedding model in terms of modes of an approximate pointwise mutual information matrix. Our work builds on this by explicitly studying the dynamics of the observable responsible for the linear analogy structure.

Linear representation hypothesis. Recent work [2, 3, 4] formalized the definition of linearity for a binary concept as when the representation difference induced by changing that concept lies in a positive cone. Our abstraction score is a continuous relaxation of this condition (see §A.2). [3] is closest in spirit to our work, as they prove that gradient descent asymptotically results in perfect abstraction. We extend this by studying the *dynamics* of abstraction during training (rather than just end-of-training), and also the evolution of abstraction across layers and through nonlinearities.

A.2 Relationship to the linear representation hypothesis

The linear representation hypothesis (LRH) motivates a geometric question: does changing a concept move the representation in the same direction across contexts? We first make this question precise for the two class-mean differences used in our abstraction score, then state the scope of its connection to LRH.

Two contexts and a common positive direction. Write the representation mean of a class as

$$\boldsymbol{r}_{s,c} \coloneqq \bar{\boldsymbol{z}}_{s,c} = \mathbb{E}[\boldsymbol{z} \mid s, c], \qquad (s, c) \in \{-1, +1\}^2. \tag{28}$$

The square-minus-circle differences in the red and blue contexts are

$$\boldsymbol{\delta}_+^S = \boldsymbol{r}_{+1,+1} - \boldsymbol{r}_{-1,+1} = \boldsymbol{u}_{\text{red}}, \qquad \boldsymbol{\delta}_-^S = \boldsymbol{r}_{+1,-1} - \boldsymbol{r}_{-1,-1} = \boldsymbol{u}_{\text{blue}}. \tag{29}$$

When both differences are nonzero, their abstraction score is

$$\alpha = \frac{\langle \boldsymbol{\delta}_+^S, \boldsymbol{\delta}_-^S \rangle}{\|\boldsymbol{\delta}_+^S\| \, \|\boldsymbol{\delta}_-^S\|}. \tag{30}$$

For $\boldsymbol{v} \neq 0$, define its positive ray by $\mathrm{Cone}_+(\boldsymbol{v}) = \{a\boldsymbol{v} : a > 0\}$. We use "positive cone" in this one-dimensional sense: both differences must have the same direction and orientation, while their lengths may differ.

Proposition A.1 (Perfect abstraction and a common positive ray)**.** *For nonzero $\boldsymbol{\delta}_+^S$ and $\boldsymbol{\delta}_-^S$, the following conditions are equivalent: $\alpha = 1$; $\boldsymbol{\delta}_-^S = a\boldsymbol{\delta}_+^S$ for some $a > 0$; and both differences belong to $\mathrm{Cone}_+(\boldsymbol{v})$ for some nonzero $\boldsymbol{v}$.*

Proof. Expanding a squared distance gives

$$\left\| \frac{\boldsymbol{\delta}_+^S}{\|\boldsymbol{\delta}_+^S\|} - \frac{\boldsymbol{\delta}_-^S}{\|\boldsymbol{\delta}_-^S\|} \right\|^2 = 1 + 1 - 2\alpha = 2(1 - \alpha). \tag{31}$$

Thus $\alpha = 1$ exactly when the two unit vectors agree, or $\boldsymbol{\delta}_-^S = (\|\boldsymbol{\delta}_-^S\|/\|\boldsymbol{\delta}_+^S\|)\boldsymbol{\delta}_+^S$. The multiplier is positive. Taking $\boldsymbol{v} = \boldsymbol{\delta}_+^S$ gives the common ray; conversely, positive multiples of the same nonzero vector have equal unit vectors and hence cosine one. □

Equation (31) also gives a quantitative interpretation: the distance between the normalized concept directions is $\sqrt{2(1-\alpha)}$. This makes abstraction a continuous measure of angular agreement. At $\alpha = -1$, the two differences lie on the same line with opposite orientations, so the meaning of a positive shape shift reverses between contexts. If either difference vanishes, its direction and the cosine score are undefined.

What this establishes about LRH. The common-ray condition is a geometric ingredient of the binary-concept linearity definitions in [2, 3]. Their language-model formulations specify which context or token pairs are valid counterfactuals and impose alignment over the relevant collection. In particular, the embedding-space definition of [2] also places conditions on the induced distributions of target and off-target concepts. Our proposition concerns the four representation means $\boldsymbol{r}_{s,c}$. Alignment of these means alone does not establish alignment of every individual example or token pair, or a relationship between embedding and unembedding directions. It establishes the corresponding two-context condition at the level measured by our score. The exact positive-ray condition is independent of the choice of positive-definite inner product; the graded cosine values depend on that choice. Here we use the Euclidean geometry of the representation.

Fourier interpretation. Expand the class means as

$$\bar{\boldsymbol{z}}_{s,c} = \hat{\boldsymbol{z}}_{\emptyset} + s\hat{\boldsymbol{z}}_S + c\hat{\boldsymbol{z}}_C + sc\hat{\boldsymbol{z}}_{SC}. \tag{32}$$

Then $\boldsymbol{\delta}^S_{\pm} = 2(\hat{\boldsymbol{z}}_S \pm \hat{\boldsymbol{z}}_{SC})$. When the representation Fourier components are orthogonal (§A.3), the two differences have equal squared length $4(q_S + q_{SC})$, where $q_m = \|\hat{\boldsymbol{z}}_m\|^2$. Their score is therefore

$$\alpha = \frac{q_S - q_{SC}}{q_S + q_{SC}}, \qquad \alpha = 1 \iff q_{SC} = 0 \quad \text{when } q_S > 0. \tag{33}$$

In this orthogonal setting, perfect abstraction means that the shape shift is identical in both contexts. The general positive-ray condition allows unequal shift lengths; orthogonality is what makes the two lengths equal here.

Terminal abstraction and deterministic interactions. For the noiseless linear training setting and initialization conditions of Proposition 2, the terminal score is

$$\alpha_\infty = \frac{1 - \sqrt{\nu_x \nu_y}}{1 + \sqrt{\nu_x \nu_y}}, \qquad \nu_x = \frac{d^{(x)}_{SC}}{d^{(x)}_S}, \qquad \nu_y = \frac{d^{(y)}_{SC}}{d^{(y)}_S}. \tag{34}$$

Here $d^{(x)}_S d^{(y)}_S > 0$, and $d^{(x)}_m$ and $d^{(y)}_m$ are squared Fourier norms. Thus $\alpha_\infty = 1$ exactly when $\nu_x \nu_y = 0$: the input means or the targets have no interaction component. The SC component describes deterministic dependence on the product sc, and can be present even though $\sigma_x = 0$. Its absence is a statement about concept geometry, separate from the absence of observation noise. The terminal law therefore identifies when the two-context mean directions become perfectly aligned and when their limiting alignment remains partial. The transient results in §B.4 and the depth results in §B.7 describe how this geometry is reached.

A.3 Fourier representation and orthogonality

This section explains the coordinates and geometric assumption used in §3. There are two separate ingredients: every function of binary latents has an exact Fourier expansion, and Assum. 2 additionally requires its vector-valued components to be orthogonal. We first distinguish these statements, then explain abstraction, input noise, and initialization in these coordinates.

Theorem A.2 (Fourier–Walsh expansion)**.** *For any integers $k, d \geq 1$, every function $\boldsymbol{f} : \{-1,+1\}^k \mapsto \mathbb{R}^d$ has a unique multilinear polynomial expansion. Write $[k] = \{1, \ldots, k\}$ and define $\chi_m(\boldsymbol{z}) = \prod_{j \in m} z_j$ for $m \subseteq [k]$, with $\chi_\emptyset = 1$. Then:*

$$\boldsymbol{f}(\boldsymbol{z}) = \sum_{m \subseteq [k]} \hat{\boldsymbol{f}}_m \chi_m(\boldsymbol{z}), \qquad \hat{\boldsymbol{f}}_m = \frac{1}{2^k} \sum_{\boldsymbol{z} \in \{-1,+1\}^k} \boldsymbol{f}(\boldsymbol{z}) \chi_m(\boldsymbol{z}). \tag{35}$$

The vectors $\hat{\boldsymbol{f}}_m \in \mathbb{R}^d$ are the Fourier components*, and the subsets m label the Fourier* modes*. Our setting specializes to $k = 2$, with $z_1 = s$, $z_2 = c$, and mode labels $\mathcal{M} = \{\emptyset, S, C, SC\}$:*

$$\chi_\emptyset(s,c) = 1, \qquad \chi_S(s,c) = s, \qquad \chi_C(s,c) = c, \qquad \chi_{SC}(s,c) = sc. \tag{36}$$

Proof. Under the uniform distribution on $\{-1,+1\}^k$, the coordinates are independent and each has mean zero. Since $z_j^2 = 1$, $\chi_A \chi_B = \prod_{j \in A \triangle B} z_j$, where $A \triangle B$ contains the indices belonging to exactly one of the two sets. Hence

$$\frac{1}{2^k} \sum_{\boldsymbol{z}} \chi_A(\boldsymbol{z}) \chi_B(\boldsymbol{z}) = \begin{cases} 1, & A = B, \\ 0, & A \neq B. \end{cases} \tag{37}$$

There are 2^k orthonormal characters and 2^k points in the domain, so these characters form a basis for all scalar functions on the domain. Applying this basis expansion to each coordinate of $\boldsymbol{f}$ gives existence. Multiplying the expansion by χ_A and averaging gives the displayed coefficient formula and proves uniqueness. Every character is multilinear because no coordinate appears more than once. □

What the four components describe. For a generic class-mean map $\boldsymbol{r}(s,c)$, the theorem gives

$$\boldsymbol{r}(s,c) = \hat{\boldsymbol{r}}_\emptyset + s\hat{\boldsymbol{r}}_S + c\hat{\boldsymbol{r}}_C + sc\hat{\boldsymbol{r}}_{SC}. \tag{38}$$

The constant component $\hat{\boldsymbol{r}}_\emptyset = \mathbb{E}_{s,c}\boldsymbol{r}$ is the overall mean. The shape component is half the shape difference averaged over colors, and the color component has the analogous meaning. The interaction component is one quarter of a difference of differences:

$$\begin{aligned} \hat{\boldsymbol{r}}_S &= \frac{1}{4}\big[\boldsymbol{r}(+,+) - \boldsymbol{r}(-,+) + \boldsymbol{r}(+,-) - \boldsymbol{r}(-,-)\big], \\ \hat{\boldsymbol{r}}_{SC} &= \frac{1}{4}\big[\boldsymbol{r}(+,+) - \boldsymbol{r}(-,+) - \boldsymbol{r}(+,-) + \boldsymbol{r}(-,-)\big]. \end{aligned} \tag{39}$$

Thus the interaction records how the effect of changing shape depends on color. It is a deterministic property of the class means. It need not represent random noise, measurement error, or variation within a class. The same decomposition applies separately to input means $\boldsymbol{\mu}$, targets $\boldsymbol{y}$, and representation means $\bar{\boldsymbol{z}}$.

Two notions of orthogonality. Under the uniform latent distribution, $\mathbb{E}_{s,c}[\chi_m \chi_n] = \delta_{mn}$. This is orthogonality of scalar *functions of the latents*; it does not say that the vectors $\hat{\boldsymbol{r}}_m$ point in different directions. For example, $\boldsymbol{r}(s,c) = (s+c)\boldsymbol{e}$, with $\|\boldsymbol{e}\| = 1$, has $\hat{\boldsymbol{r}}_S = \hat{\boldsymbol{r}}_C = \boldsymbol{e}$. The Fourier expansion is exact, but these two components coincide. Even in this example, character orthogonality gives the energy identity

$$\mathbb{E}_{s,c}\|\boldsymbol{r}(s,c)\|^2 = \sum_{m,n} \langle \hat{\boldsymbol{r}}_m, \hat{\boldsymbol{r}}_n \rangle \mathbb{E}_{s,c}[\chi_m \chi_n] = \sum_m \|\hat{\boldsymbol{r}}_m\|^2. \tag{40}$$

Subtracting the overall mean removes only the $\emptyset$ term. This identity requires no Euclidean orthogonality between the components.

Assum. 2 imposes that additional Euclidean orthogonality on the input and target components:

$$\langle \hat{\boldsymbol{x}}_m, \hat{\boldsymbol{x}}_n \rangle = d_m^{(x)} \delta_{mn}, \qquad \langle \hat{\boldsymbol{y}}_m, \hat{\boldsymbol{y}}_n \rangle = d_m^{(y)} \delta_{mn}, \qquad d_m^{(x)} = \|\hat{\boldsymbol{x}}_m\|^2, \quad d_m^{(y)} = \|\hat{\boldsymbol{y}}_m\|^2. \tag{41}$$

In particular, $d_m^{(x)}$ and $d_m^{(y)}$ are energies, rather than vector lengths. Zero-energy components are allowed. Each ambient space must have enough dimensions for its nonzero orthogonal components, but input and target dimensions need not coincide. Orthogonality within the input space and within the target space does not impose any identification between those two spaces.

Abstraction before and after imposing orthogonality. Changing shape at fixed color gives

$$\begin{aligned} \boldsymbol{u}_+ &= \boldsymbol{r}(+,+) - \boldsymbol{r}(-,+) = 2(\hat{\boldsymbol{r}}_S + \hat{\boldsymbol{r}}_{SC}), \\ \boldsymbol{u}_- &= \boldsymbol{r}(+,-) - \boldsymbol{r}(-,-) = 2(\hat{\boldsymbol{r}}_S - \hat{\boldsymbol{r}}_{SC}). \end{aligned} \tag{42}$$

The constant and color components cancel. Write $A = \|\hat{\boldsymbol{r}}_S\|^2$, $B = \|\hat{\boldsymbol{r}}_{SC}\|^2$, and $h = \langle \hat{\boldsymbol{r}}_S, \hat{\boldsymbol{r}}_{SC} \rangle$. Direct expansion gives

$$\begin{aligned} \boldsymbol{u}_+^\top \boldsymbol{u}_- &= 4(A - B), & \|\boldsymbol{u}_\pm\|^2 &= 4(A + B \pm 2h), \\ \cos(\boldsymbol{u}_+, \boldsymbol{u}_-) &= \frac{A - B}{\sqrt{(A+B)^2 - 4h^2}}. \end{aligned} \tag{43}$$

The cosine is defined when both concept directions are nonzero. If the shape and interaction components are orthogonal, then $h = 0$, their two context-specific directions have equal length, and this reduces to

$$\alpha_r = \frac{A - B}{A + B} = \frac{1 - \nu_r}{1 + \nu_r}, \qquad \nu_r = \frac{B}{A}. \tag{44}$$

The first expression requires $A + B > 0$; the ratio expression also requires $A > 0$. At fixed positive A, increasing the orthogonal interaction energy lowers abstraction. In this family, $B = 0 < A$ gives $\alpha_r = 1$, $A = B > 0$ gives $\alpha_r = 0$, and $A = 0 < B$ gives $\alpha_r = -1$.

Outside the orthogonal family, a nonzero interaction can change only the length of a concept direction. For example, $\hat{\boldsymbol{r}}_{SC} = a\hat{\boldsymbol{r}}_S$ with $\hat{\boldsymbol{r}}_S \neq 0$ and $|a| < 1$ gives two positively parallel directions and hence $\alpha_r = 1$. Thus the interpretation of an energy ratio as an angular entanglement score uses orthogonality; the Fourier expansion alone does not suffice. For the input, target, and aligned hidden geometries, the relevant ratios are precisely

$$\nu_x = \frac{d_{SC}^{(x)}}{d_S^{(x)}}, \qquad \nu_y = \frac{d_{SC}^{(y)}}{d_S^{(y)}}, \qquad \nu_z = \frac{q_{SC}}{q_S}, \qquad q_m = \|\hat{\boldsymbol{z}}_m\|^2. \tag{45}$$

Color abstraction follows by exchanging S and C. Although the constant and color modes cancel from shape differences, they can still affect nonlinear learning through the total preactivation variance; their dynamical role is treated in Appendix C.

A population construction satisfying the assumption. Choose orthonormal input directions $\boldsymbol{e}_m$ and target directions $\boldsymbol{v}_m$ for their respective nonzero modes. Any prescribed nonnegative energies can be realized by

$$\begin{aligned} \boldsymbol{\mu}(s,c) &= \sum_{m: d_m^{(x)} > 0} \sqrt{d_m^{(x)}}\, \boldsymbol{e}_m \chi_m(s,c), \quad \boldsymbol{y}(s,c) = \sum_{m: d_m^{(y)} > 0} \sqrt{d_m^{(y)}}\, \boldsymbol{v}_m \chi_m(s,c), \\ \boldsymbol{x} &= \boldsymbol{\mu}(s,c) + \sigma_x \boldsymbol{\epsilon}, \qquad \mathbb{E}\boldsymbol{\epsilon} = 0, \qquad \mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \boldsymbol{I}_{D_x}, \qquad \boldsymbol{\epsilon} \perp (s,c). \end{aligned} \tag{46}$$

The coefficient formula immediately gives $\hat{\boldsymbol{x}}_m = \mathbb{E}[\boldsymbol{x}\chi_m] = \sqrt{d_m^{(x)}}\boldsymbol{e}_m$ and $\hat{\boldsymbol{y}}_m = \sqrt{d_m^{(y)}}\boldsymbol{v}_m$ for present modes. The random term has zero conditional mean, so it does not add a Fourier component to the class-mean map. Its within-class covariance is $\sigma_x^2 \boldsymbol{I}_{D_x}$, and its total expected squared magnitude is $D_x \sigma_x^2$. Thus increasing $d_{SC}^{(x)}$ changes the deterministic interaction, whereas increasing σ_x changes within-class variation. These are distinct interventions even if both worsen a particular prediction task.

For a linear representation, $\bar{\boldsymbol{z}}(s,c) = \boldsymbol{W}\boldsymbol{\mu}(s,c)$, so $\hat{\boldsymbol{z}}_m = \boldsymbol{W}\hat{\boldsymbol{x}}_m$; its within-class covariance is $\sigma_x^2 \boldsymbol{W}\boldsymbol{W}^\top$. Consequently the cosine of class-mean directions and the noise seen by a probe are different quantities. Their relationship is derived in §A.5. All nonlinear theory in Appendix C sets $\sigma_x = 0$; an interaction component may remain nonzero in that noiseless setting.

The four-class Gram matrix. A small matrix provides a useful visualization of the same population geometry. Index its rows and columns by the four latent classes $u = (s, c)$, and let $G_{uv} = \boldsymbol{\mu}(u)^\top \boldsymbol{\mu}(v)$. Put $H_{um} = \chi_m(u)/2$, so $\boldsymbol{H}^\top \boldsymbol{H} = \boldsymbol{I}_4$, and let $\widehat{\boldsymbol{X}}$ have columns $\hat{\boldsymbol{x}}_m$. The matrix of four class means is $2\widehat{\boldsymbol{X}}\boldsymbol{H}^\top$, which gives

$$\boldsymbol{H}^\top \boldsymbol{G} \boldsymbol{H} = 4\widehat{\boldsymbol{X}}^\top \widehat{\boldsymbol{X}}. \tag{47}$$

Thus the character basis diagonalizes the class Gram matrix exactly when the input Fourier vectors are orthogonal. Under the assumption, $\boldsymbol{G} = 4\boldsymbol{H}\,\mathrm{diag}(d_m^{(x)})\boldsymbol{H}^\top$. The eigenvalues of $\boldsymbol{G}$ are $4d_m^{(x)}$; the averaging operator $\boldsymbol{G}/4$ has eigenvalues $d_m^{(x)}$. The factor four counts classes, not training examples.

More directly, orthogonality gives

$$G_{uv} = d_\emptyset^{(x)} + d_S^{(x)} ss' + d_C^{(x)} cc' + d_{SC}^{(x)} ss' cc', \qquad u = (s, c), \quad v = (s', c'). \tag{48}$$

There are four possible values, according to the two class agreements:

Shape	Color	Class-mean inner product
same	same	$d_\emptyset^{(x)} + d_S^{(x)} + d_C^{(x)} + d_{SC}^{(x)}$
same	different	$d_\emptyset^{(x)} + d_S^{(x)} - d_C^{(x)} - d_{SC}^{(x)}$
different	same	$d_\emptyset^{(x)} - d_S^{(x)} + d_C^{(x)} - d_{SC}^{(x)}$
different	different	$d_\emptyset^{(x)} - d_S^{(x)} - d_C^{(x)} + d_{SC}^{(x)}$

In particular all four class means have the same squared length. This is a consequence of the orthogonality assumption, not a property of an arbitrary four-class dataset.

For independently drawn examples $\boldsymbol{x}_i$ with fixed class labels u_i, the population construction gives

$$\mathbb{E}[\boldsymbol{x}_i^\top \boldsymbol{x}_j \mid u_i, u_j] = \boldsymbol{\mu}(u_i)^\top \boldsymbol{\mu}(u_j) + D_x \sigma_x^2 \mathbf{1}_{\{i=j\}}. \tag{49}$$

There can therefore be an additional diagonal value in the expected sample Gram matrix, beyond the four class-agreement values. The extra term vanishes when $\sigma_x = 0$. A finite set of noisy samples usually has fluctuations around these values: balancing the class counts does not make the displayed expectation an exact sample identity. Our analysis uses population moments and class means, so it does not require such an identity for a realized sample Gram matrix.

Gaussian initialization: expectation and concentration. Let the D rows of $\boldsymbol{W}_0$ be independent with $\boldsymbol{w}_i \sim \mathcal{N}(0, \sigma_w^2 \boldsymbol{I}_{D_x}/D)$, where $\sigma_w > 0$. Then

$$\mathbb{E}_{\boldsymbol{W}_0}[\boldsymbol{W}_0^\top \boldsymbol{W}_0] = \sum_{i=1}^{D} \mathbb{E}[\boldsymbol{w}_i \boldsymbol{w}_i^\top] = \sigma_w^2 \boldsymbol{I}_{D_x}.$$

At any width, the Gram matrix is thus a scaled identity in expectation, so orthogonal input modes have zero expected hidden cross inner products. Large width additionally controls deviations from these expectations in a realized initialization. For each present mode set

$$P_{mn} = \langle \boldsymbol{W}_0 \hat{\boldsymbol{x}}_m, \boldsymbol{W}_0 \hat{\boldsymbol{x}}_n \rangle, \qquad T_{mn} = \frac{P_{mn}}{\sigma_w^2 \sqrt{d_m^{(x)} d_n^{(x)}}}. \tag{50}$$

For an orthonormal set of input directions, the scalar projections $\xi_{i,m} = \boldsymbol{w}_i^\top \boldsymbol{e}_m$ are independent Gaussians with variance σ_w^2/D, and $P_{mn} = \sqrt{d_m^{(x)} d_n^{(x)}} \sum_i \xi_{i,m} \xi_{i,n}$. If $m \neq n$, each summand has zero mean and variance σ_w^4/D^2. If $m = n$, its mean is σ_w^2/D and its variance is $2\sigma_w^4/D^2$, using $\mathbb{E}\xi^4 = 3(\mathbb{E}\xi^2)^2$. Independence across rows therefore gives

$$\begin{aligned} \mathbb{E}P_{mn} &= \sigma_w^2 d_m^{(x)} \delta_{mn}, & \mathrm{Var}(P_{mn}) &= \frac{\sigma_w^4 d_m^{(x)} d_n^{(x)}}{D}(1 + \delta_{mn}), \\ \mathbb{E}T_{mn} &= \delta_{mn}, & \mathrm{Var}(T_{mn}) &= \frac{1 + \delta_{mn}}{D}. \end{aligned} \tag{51}$$

Chebyshev's inequality and a union bound over the at most four diagonal and six distinct off-diagonal entries yield, for $\varepsilon > 0$,

$$\mathbb{P}\left(\max_{m,n} |T_{mn} - \delta_{mn}| \geq \varepsilon\right) \leq \min\left\{1, \frac{14}{D\varepsilon^2}\right\}. \tag{52}$$

On the complementary event with $0 < \varepsilon < 1$,

$$\left|\frac{q_m(0)}{\sigma_w^2 d_m^{(x)}} - 1\right| < \varepsilon, \qquad |\cos(\hat{\boldsymbol{z}}_m(0), \hat{\boldsymbol{z}}_n(0))| \leq \frac{\varepsilon}{1-\varepsilon} \quad (m \neq n). \tag{53}$$

In particular, for positive shape and interaction input energies, $\nu_z(0) \to \nu_x$ in expectation, and also in probability as $D \to \infty$. This argument controls only the four signal directions; it does not require $\boldsymbol{W}_0^\top \boldsymbol{W}_0$ to approximate a scalar identity on an input space whose dimension may also be large. It describes initialization, and does not by itself prove exact mode decoupling or uniform approximation of the training trajectory at finite width.

The additional assumption in the nonlinear extension. §C.1 uses the convention $q_m = D^{-1}\|\hat{\boldsymbol{z}}_m\|^2$. With row covariance $\sigma_w^2 \boldsymbol{I}_{D_x}$ instead of $\sigma_w^2 \boldsymbol{I}_{D_x}/D$, the preceding calculation applies to $D^{-1}P_{mn}$ and gives the same limiting energies $q_m(0) = \sigma_w^2 d_m^{(x)}$. The preactivation Fourier coefficients across neurons are then initially independent centered Gaussians. Relabeling symmetry supplies orthogonality of the population Fourier components, while a Gaussian closure assumption is used to express the nonlinear learning dynamics in terms of their energies.

A.4 Optimal readout and variable projection

Assum. 1 sets the readout to its ridge optimum for the current representation. This is a variable-projection reduction: the parameters entering the prediction linearly are minimized out before updating the feature weights [52, 53]. It is also motivated by training regimes in which the readout adapts faster than the features [33, 34]. The readout still changes as the representation changes. Here we explain the approximation and compare it with joint training at equal learning rates.

The finite-data version of the population loss. For N training pairs, define normalized data matrices

$$\boldsymbol{X}_N = N^{-1/2}[\boldsymbol{x}_1, \dots, \boldsymbol{x}_N], \qquad \boldsymbol{Y}_N = N^{-1/2}[\boldsymbol{y}_1, \dots, \boldsymbol{y}_N], \qquad \boldsymbol{Z}_N = \boldsymbol{W}\boldsymbol{X}_N. \tag{54}$$

Then the empirical counterpart of (2) is

$$\begin{aligned}\mathcal{L}_N(\boldsymbol{W}_r, \boldsymbol{W}) &= \frac{1}{2}\|\boldsymbol{W}_r\boldsymbol{Z}_N - \boldsymbol{Y}_N\|_F^2 + \frac{1}{2\gamma}\|\boldsymbol{W}_r\|_F^2 \\ &= \frac{1}{2N}\sum_{i=1}^{N}\|\boldsymbol{W}_r\boldsymbol{W}\boldsymbol{x}_i - \boldsymbol{y}_i\|^2 + \frac{1}{2\gamma}\|\boldsymbol{W}_r\|_F^2.\end{aligned} \tag{55}$$

The $N^{-1/2}$ factors therefore implement the sample average; they are not an additional change of regularization. Put $\boldsymbol{C}_N = \boldsymbol{Z}_N\boldsymbol{Z}_N^\top + \gamma^{-1}\boldsymbol{I}_D$. Differentiating with respect to the readout gives

$$\nabla_{\boldsymbol{W}_r}\mathcal{L}_N = \boldsymbol{W}_r\boldsymbol{C}_N - \boldsymbol{Y}_N\boldsymbol{Z}_N^\top, \qquad \boldsymbol{W}_r^* = \boldsymbol{Y}_N\boldsymbol{Z}_N^\top\boldsymbol{C}_N^{-1}. \tag{56}$$

The matrix $\boldsymbol{C}_N$ is positive definite for $\gamma > 0$, so the optimum is unique. Substitution yields

$$\mathcal{L}_N^*(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{Y}_N\|_F^2 - \frac{1}{2}\operatorname{Tr}\left[\boldsymbol{Y}_N\boldsymbol{Z}_N^\top\boldsymbol{C}_N^{-1}\boldsymbol{Z}_N\boldsymbol{Y}_N^\top\right]. \tag{57}$$

Replacing the empirical second moments by population expectations gives the main-text readout formula and the derivation in §B.1. These identities do not require Fourier orthogonality or an aligned initialization.

Readout relaxation and the loss gap. For fixed $\boldsymbol{W}$, readout gradient flow with learning rate η_r gives

$$\dot{\boldsymbol{E}} = -\eta_r\boldsymbol{E}\boldsymbol{C}_N, \qquad \boldsymbol{E} = \boldsymbol{W}_r - \boldsymbol{W}_r^*, \qquad \|\boldsymbol{E}(t)\|_F \le e^{-\eta_r t/\gamma}\|\boldsymbol{E}(0)\|_F. \tag{58}$$

This follows by diagonalizing the symmetric matrix $\boldsymbol{C}_N$, whose eigenvalues are at least γ^{-1}. When $\boldsymbol{W}$ also moves, the equation contains the additional term $-\dot{\boldsymbol{W}}_r^*$. The approximation is consequently useful when readout relaxation is fast relative to movement of the optimum. It need not capture the initial transient of a slowly adapting readout.

Completing the square gives an exact measure of the discrepancy:

$$\mathcal{L}_N(\boldsymbol{W}_r, \boldsymbol{W}) - \mathcal{L}_N^*(\boldsymbol{W}) = \frac{1}{2}\operatorname{Tr}(\boldsymbol{E}\boldsymbol{C}_N\boldsymbol{E}^\top) \ge 0. \tag{59}$$

The gap is zero precisely at the optimal readout. In variable-projected training, the feature gradient is evaluated with $\boldsymbol{W}_r = \boldsymbol{W}_r^*$; the derivative through this optimum vanishes by the readout normal equation, as shown in §B.1.

Simulation setup. We use a balanced synthetic dataset with two binary latents $(s, c) \in \{\pm 1\}^2$. Inputs and deterministic targets are

$$\boldsymbol{x} = a_s s\boldsymbol{e}_S + a_c c\boldsymbol{e}_C + a_{sc}sc\boldsymbol{e}_{SC} + \frac{\sigma}{\sqrt{D_x}}\boldsymbol{\epsilon}, \qquad \boldsymbol{y} = [s, c, \rho sc]^\top, \tag{60}$$

where $\boldsymbol{e}_S, \boldsymbol{e}_C, \boldsymbol{e}_{SC}$ are orthonormal directions in $\mathbb{R}^{D_x}$, drawn independently for each seed, and the noise has zero mean and identity covariance. The parameters are $D_x = 20$, $a_s = 1.0$, $a_c = 0.7$, $a_{sc} = 0.6$, $\sigma = 0.6$, and $\rho = 0.3$. In the notation of Assum. 2, this gives

$$(d_S^{(x)}, d_C^{(x)}, d_{SC}^{(x)}) = (1, 0.49, 0.36), \qquad (d_S^{(y)}, d_C^{(y)}, d_{SC}^{(y)}) = (1, 1, 0.09), \qquad \sigma_x^2 = \frac{0.36}{20}. \tag{61}$$

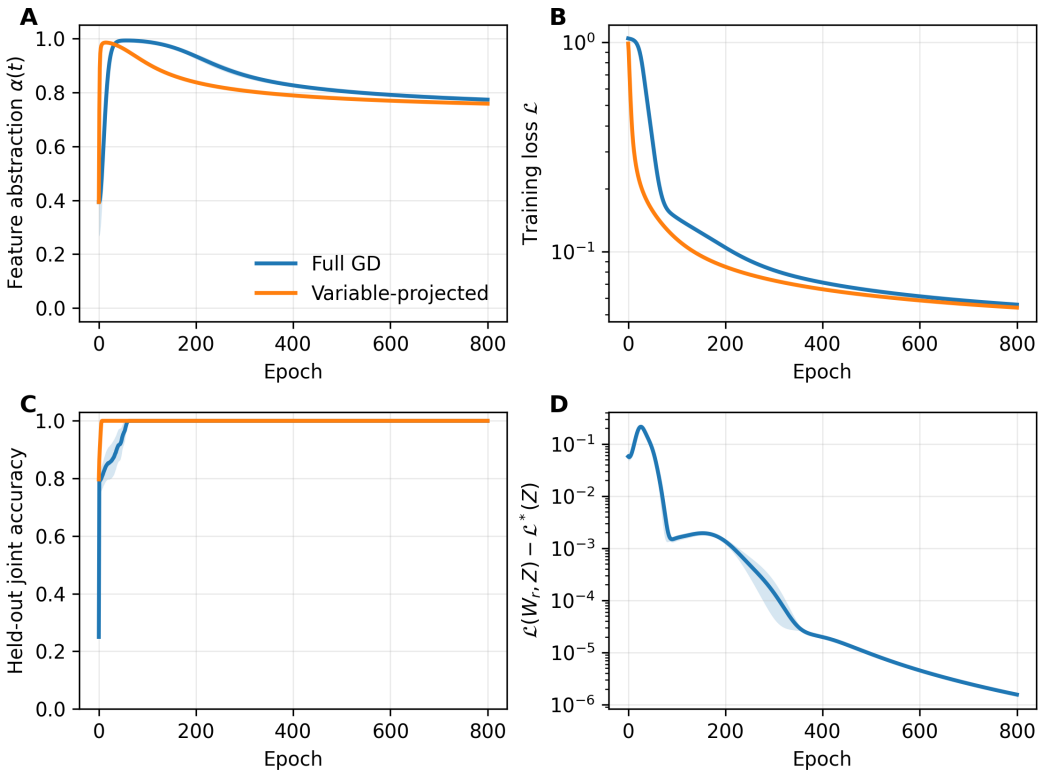

Figure 8: **Optimal readout versus joint training at equal learning rates.** Full-batch gradient descent updates both $\boldsymbol{W}$ and $\boldsymbol{W}_r$ on $\mathcal{L}_N$; variable projection updates $\boldsymbol{W}$ on $\mathcal{L}_N^*$, setting the readout to its ridge optimum at each step (Assum. 1). Curves show mean $\pm$ SEM over 12 seeds. **(A)** Shape abstraction. **(B)** Regularized training loss. **(C)** Held-out joint shape/color accuracy. **(D)** Loss gap (59) for joint training; it is zero by construction for variable projection.

Both constant Fourier components vanish. The interaction energies are properties of the class means and targets; σ_x^2 separately controls within-class input variation. There are 64 training examples per class ($N = 256$) and 256 held-out examples per class ($N_{\text{test}} = 1024$). The test set uses the same latent directions as the training set, with fresh noise. The matrices are normalized as in (55).

The network has hidden width $D = 4$, with $\boldsymbol{W} \in \mathbb{R}^{4\times 20}$ and $\boldsymbol{W}_r \in \mathbb{R}^{3\times 4}$. Entries of $\boldsymbol{W}(0)$ are Gaussian with standard deviation $0.2/\sqrt{20}$. Joint training initializes $\boldsymbol{W}_r = 0$, whereas variable projection immediately sets it to its optimum. Both procedures use the same data and initial feature weights for each seed, ridge parameter $\gamma = 10$, learning rate 0.1, and 800 full-batch updates. Results are averaged over 12 seeds; shaded regions show the standard error of the mean. We measure abstraction from the empirical class means using the definition in §3. Joint test accuracy is the fraction of examples for which both shape and color are correctly classified by thresholding their predictions at zero.

What the comparison shows. In Fig. 8, joint training first learns a useful readout, while variable projection uses the best readout for its current features immediately. Variable projection therefore reduces loss and reaches high test accuracy earlier; its abstraction peak also occurs earlier. At later times the two loss and abstraction curves become close, and the readout loss gap decreases substantially. The comparison supports the optimal-readout approximation for these later dynamics, while showing its limitations during the initial readout transient. This is a finite-width experiment with observation noise and random feature initialization; it tests the approximation beyond the exactly aligned linear family and does not assume the nonlinear Gaussian closure.

A.5 Abstraction and generalization of linear probes

Assumptions and centering. We make explicit the population assumptions behind paragraph (i) of Section 3.1. In each color context, the two shapes have equal prior probability, and $\boldsymbol{z} \mid s, c \sim \mathcal{N}(\bar{\boldsymbol{z}}_{s,c}, \sigma^2 \boldsymbol{I}_D)$ with $\sigma > 0$. Both shape directions below are nonzero. The representation is fixed, and we use the population class means; errors from estimating the probe or the centering offsets are not included. Let $\bar{\boldsymbol{z}}_a = \mathbb{E}[\boldsymbol{z} \mid a]$ denote a representation mean, and define the two shape directions

$$\boldsymbol{u}_{\text{red}} = \bar{\boldsymbol{z}}_{\square} - \bar{\boldsymbol{z}}_{\bigcirc}, \qquad \boldsymbol{u}_{\text{blue}} = \bar{\boldsymbol{z}}_{\square} - \bar{\boldsymbol{z}}_{\bigcirc}, \qquad \alpha = \frac{\boldsymbol{u}_{\text{red}}^\top \boldsymbol{u}_{\text{blue}}}{\|\boldsymbol{u}_{\text{red}}\| \|\boldsymbol{u}_{\text{blue}}\|}. \tag{62}$$

Center the red and blue examples separately at their respective population midpoints:

$$\begin{aligned} \boldsymbol{m}_{\text{red}} &= \tfrac{1}{2}(\bar{\boldsymbol{z}}_{\square} + \bar{\boldsymbol{z}}_{\bigcirc}), & \boldsymbol{m}_{\text{blue}} &= \tfrac{1}{2}(\bar{\boldsymbol{z}}_{\square} + \bar{\boldsymbol{z}}_{\bigcirc}), \\ \tilde{\boldsymbol{z}}_c &= \boldsymbol{z} - \boldsymbol{m}_c, & \tilde{\boldsymbol{z}}_c \mid s, c &\sim \mathcal{N}\left(\tfrac{s}{2}\boldsymbol{u}_c, \sigma^2 \boldsymbol{I}_D\right), \qquad c \in \{\text{red}, \text{blue}\}. \end{aligned} \tag{63}$$

Here the color names stand for $c = +1$ and $c = -1$, respectively. Equal shape priors make each midpoint equal to the corresponding color-conditional mean $\mathbb{E}[\boldsymbol{z} \mid c]$.

Bayes-optimal probe on red examples. With equal priors, the posterior log-odds on centered red examples are

$$\log \frac{p(s = +1 \mid \tilde{\boldsymbol{z}}_{\text{red}}, c = +1)}{p(s = -1 \mid \tilde{\boldsymbol{z}}_{\text{red}}, c = +1)} = \frac{\boldsymbol{u}_{\text{red}}^\top \tilde{\boldsymbol{z}}_{\text{red}}}{\sigma^2}. \tag{64}$$

Thus the Bayes-optimal red classifier predicts the sign of $\hat{\boldsymbol{w}}^\top \tilde{\boldsymbol{z}}_{\text{red}}$, where $\hat{\boldsymbol{w}} = \boldsymbol{u}_{\text{red}}/\|\boldsymbol{u}_{\text{red}}\|$. It has zero intercept in these centered coordinates. We transfer this same direction and zero threshold to the separately centered blue examples, giving $T = \hat{\boldsymbol{w}}^\top \tilde{\boldsymbol{z}}_{\text{blue}}$. The test-score distribution is exactly

$$T \mid s, c = -1 \sim \mathcal{N}\left(s\frac{\|\boldsymbol{u}_{\text{blue}}\|}{2}\alpha, \sigma^2\right), \tag{65}$$

since $\hat{\boldsymbol{w}}^\top \boldsymbol{u}_{\text{blue}} = \|\boldsymbol{u}_{\text{blue}}\|\alpha$ and $\|\hat{\boldsymbol{w}}\| = 1$. The two shapes have the same misclassification probability by symmetry. Writing Φ and φ for the standard normal distribution function and density gives

$$P_{\text{error}}^{\text{red}\to\text{blue}} = \Phi\left(-\frac{\|\boldsymbol{u}_{\text{blue}}\|}{2\sigma}\alpha\right), \tag{66}$$

$$\left.\frac{\partial P_{\text{error}}}{\partial \alpha}\right|_{\|\boldsymbol{u}_{\text{blue}}\|/\sigma} = -\frac{\|\boldsymbol{u}_{\text{blue}}\|}{2\sigma}\varphi\left(\frac{\|\boldsymbol{u}_{\text{blue}}\|}{2\sigma}\alpha\right) < 0. \tag{67}$$

This proves (8) exactly under the stated assumptions. Positive abstraction gives error below chance, zero abstraction gives chance error, and negative abstraction gives above-chance error. The monotonicity holds at fixed $\|\boldsymbol{u}_{\text{blue}}\|/\sigma$; changes in class separation or noise during training also affect error. The result concerns the population probe, rather than an arbitrary classifier fitted from finitely many red samples.

Why the centering convention matters. If blue examples instead retain the red midpoint, their score acquires the offset $b = \hat{\boldsymbol{w}}^\top(\boldsymbol{m}_{\text{blue}} - \boldsymbol{m}_{\text{red}})$. Writing $a = \|\boldsymbol{u}_{\text{blue}}\|\alpha/2$, the balanced test error becomes

$$P_{\text{error}} = \tfrac{1}{2}\left[\Phi\left(-\frac{a+b}{\sigma}\right) + \Phi\left(\frac{b-a}{\sigma}\right)\right]. \tag{68}$$

Separate centering sets $b = 0$ and recovers (66). It is sufficient, but not necessary: retaining the red midpoint also works when its displacement from the blue midpoint is orthogonal to the probe. In the orthogonal Fourier geometry, $\boldsymbol{m}_{\text{blue}} - \boldsymbol{m}_{\text{red}} = -2\hat{\boldsymbol{z}}_C$ is orthogonal to $\boldsymbol{u}_{\text{red}} = 2(\hat{\boldsymbol{z}}_S + \hat{\boldsymbol{z}}_{SC})$, so this condition holds automatically. Unequal shape priors or estimated midpoints generally introduce additional threshold terms.

Beyond isotropic representation noise. For a fixed mean-difference probe, the same error calculation applies with σ replaced by the projected standard deviation $\sigma_{\text{blue}} = \sqrt{\hat{\boldsymbol{w}}^\top \boldsymbol{\Sigma}_{\text{blue}} \hat{\boldsymbol{w}}}$, provided the blue score distributions are Gaussian with the same within-class covariance $\boldsymbol{\Sigma}_{\text{blue}}$. For non-Gaussian scores it is an approximation, as in the main-text experiment. With anisotropic red covariance, the mean-difference direction is generally no longer Bayes-optimal.

For the linear population model with Gaussian input noise, the score is exactly Gaussian and each class has covariance $\sigma_x^2 \boldsymbol{W}\boldsymbol{W}^\top$. In the mode-aligned family derived below, $\boldsymbol{u}_{\text{red}} = 2(\hat{\boldsymbol{z}}_S + \hat{\boldsymbol{z}}_{SC})$ and $\boldsymbol{u}_{\text{blue}} = 2(\hat{\boldsymbol{z}}_S - \hat{\boldsymbol{z}}_{SC})$. For positive input energies and $\sigma_x > 0$, orthogonality gives

$$\|\boldsymbol{u}_{\text{blue}}\| = 2\sqrt{q_S + q_{SC}}, \qquad \sigma_{\text{blue}}^2 = \sigma_x^2 \frac{q_S^2/d_S^{(x)} + q_{SC}^2/d_{SC}^{(x)}}{q_S + q_{SC}}, \tag{69}$$

so the exact error in this specialization is

$$P_{\text{error}}^{\text{red}\to\text{blue}} = \Phi\left(-\frac{q_S - q_{SC}}{\sigma_x \sqrt{q_S^2/d_S^{(x)} + q_{SC}^2/d_{SC}^{(x)}}} \right). \tag{70}$$

This also illustrates why abstraction and class separability are distinct: the score measures directional alignment, whereas transfer error depends on alignment and projected noise. If a concept direction vanishes, its cosine and normalized probe are undefined. In the zero-variance limit, the sign of a nonzero projected mean determines the error directly; the Gaussian formula is understood by continuity away from ties.

Projected-variance prediction. Keeping the equal shape priors and separate midpoint centering above, write $\boldsymbol{w}_{\text{red}} = \hat{\boldsymbol{w}} = \boldsymbol{u}_{\text{red}}/\|\boldsymbol{u}_{\text{red}}\|$. For a common blue within-class covariance $\boldsymbol{\Sigma}_{\text{blue}}$, which need not be isotropic, the projected scores have means $\pm\|\boldsymbol{u}_{\text{blue}}\|\alpha/2$ and variance $v_{\text{blue}} = \boldsymbol{w}_{\text{red}}^\top \boldsymbol{\Sigma}_{\text{blue}} \boldsymbol{w}_{\text{red}}$. If these scores are Gaussian and $v_{\text{blue}} > 0$, their balanced misclassification probability is exactly

$$P_{\text{error}}^{\text{red}\to\text{blue}} = \Phi\left(-\frac{\|\boldsymbol{u}_{\text{blue}}\|\alpha}{2\sqrt{\boldsymbol{w}_{\text{red}}^\top \boldsymbol{\Sigma}_{\text{blue}} \boldsymbol{w}_{\text{red}}}} \right). \tag{71}$$

This reduces to (66) when $\boldsymbol{\Sigma}_{\text{blue}} = \sigma^2 \boldsymbol{I}_D$. The prediction in §E.1.1 substitutes empirical class means and the pooled blue within-class covariance $\widehat{\boldsymbol{\Sigma}}_{\text{blue}}$. For those learned nonlinear representations, it is a Gaussian approximation using estimated moments, rather than an exact finite-sample identity.

A.6 Estimation of steering vectors from finite samples

Assumptions and sampling scheme. We derive the squared-error formula in paragraph (ii) of Section 3.1 and the angular error of the same estimator. Fix the representation and its Fourier components, suppressing training time. For each shape $s \in \{-1, +1\}$, draw $n \geq 1$ independent samples

$$\begin{aligned} \boldsymbol{z}_{s,i} &= \hat{\boldsymbol{z}}_\emptyset + s\hat{\boldsymbol{z}}_S + c_{s,i}\hat{\boldsymbol{z}}_C + s c_{s,i}\hat{\boldsymbol{z}}_{SC} + \boldsymbol{\varepsilon}_{s,i}, \\ \mathbb{P}(c_{s,i} = +1) = \mathbb{P}(c_{s,i} = -1) &= \tfrac{1}{2}, \qquad \boldsymbol{\varepsilon}_{s,i} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_D). \end{aligned} \tag{72}$$

All colors and noises are mutually independent, including across the two shape groups. Thus there are $2n$ samples in total, and colors are balanced in expectation within each shape; their realized counts fluctuate. We use $\sigma \geq 0$, allowing the noiseless case as a boundary. As in the minimal model under Assum. 2, take the shape, color, and interaction components to be mutually orthogonal and write $q_m = \|\hat{\boldsymbol{z}}_m\|^2$. The squared-error calculation below does not require this orthogonality; its expression through α and the angular calculation use it. Assume $q_S > 0$ whenever taking an angle with the population steering vector or dividing by q_S.

Population vector and sampling error. Since each color has mean zero, the shape-conditional population means and their empirical estimates are

$$\begin{aligned} \mathbb{E}[\boldsymbol{z} \mid s] &= \hat{\boldsymbol{z}}_\emptyset + s\hat{\boldsymbol{z}}_S, & \bar{\boldsymbol{z}}_s^{(n)} &= \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{z}_{s,i}, \\ \Delta &= 2\hat{\boldsymbol{z}}_S, & \widehat{\Delta} &= \bar{\boldsymbol{z}}_+^{(n)} - \bar{\boldsymbol{z}}_-^{(n)}. \end{aligned} \tag{73}$$

Define the empirical color averages $\bar{c}_s = n^{-1}\sum_i c_{s,i}$. They are independent across shapes, with $\mathbb{E}\bar{c}_s = 0$ and $\mathbb{E}\bar{c}_s^2 = 1/n$. Subtracting the two sample means gives

$$\begin{aligned} \boldsymbol{e} \coloneqq \widehat{\Delta} - \Delta &= (\bar{c}_+ - \bar{c}_-)\hat{\boldsymbol{z}}_C + (\bar{c}_+ + \bar{c}_-)\hat{\boldsymbol{z}}_{SC} + \boldsymbol{\eta}, \\ \boldsymbol{\eta} &= \frac{1}{n}\sum_i (\boldsymbol{\varepsilon}_{+,i} - \boldsymbol{\varepsilon}_{-,i}) \sim \mathcal{N}\left(\mathbf{0}, \frac{2\sigma^2}{n}\boldsymbol{I}_D\right). \end{aligned} \tag{74}$$

The noise $\boldsymbol{\eta}$ is independent of both color averages. Moreover,

$$\mathbb{E}[(\bar{c}_+ - \bar{c}_-)^2] = \mathbb{E}[(\bar{c}_+ + \bar{c}_-)^2] = \frac{2}{n}, \qquad \mathbb{E}[(\bar{c}_+ - \bar{c}_-)(\bar{c}_+ + \bar{c}_-)] = 0. \tag{75}$$

The two color coefficients are uncorrelated, although generally not independent. Consequently the estimator is unbiased and

$$\mathbb{E}[\widehat{\Delta}] = \Delta, \qquad \mathrm{Cov}(\widehat{\Delta}) = \frac{2}{n}\left[\hat{\boldsymbol{z}}_C\hat{\boldsymbol{z}}_C^\top + \hat{\boldsymbol{z}}_{SC}\hat{\boldsymbol{z}}_{SC}^\top + \sigma^2\boldsymbol{I}_D\right]. \tag{76}$$

Taking the trace yields the exact mean-squared error

$$\begin{aligned} \mathbb{E}\left[\|\widehat{\Delta} - \Delta\|^2\right] &= \frac{2}{n}\left[q_C + q_{SC} + D\sigma^2\right] \\ &= \frac{2}{n}\left[q_S\frac{1-\alpha}{1+\alpha} + q_C + D\sigma^2\right], \end{aligned} \tag{77}$$

where the second equality uses $\alpha = (q_S - q_{SC})/(q_S + q_{SC})$ with $q_S > 0$. At fixed q_S, q_C, D, σ, n, its derivative with respect to α is $-4q_S/[n(1+\alpha)^2] < 0$, giving precisely the main-text claim. The color and interaction terms reflect fluctuations in the sampled contexts. If instead exactly $n/2$ red and $n/2$ blue examples were imposed within each shape (for even n), both color averages would vanish and the error would be $2D\sigma^2/n$.

Angular error. Let $\theta \in [0, \pi]$ be the angle between Δ and $\widehat{\Delta}$, and set

$$\boldsymbol{e}_S = \frac{\hat{\boldsymbol{z}}_S}{\sqrt{q_S}}, \qquad r = \|\Delta\| = 2\sqrt{q_S}, \qquad \tau^2 = \frac{2\sigma^2}{n}. \tag{78}$$

By orthogonality, both color-sampling terms in (74) are perpendicular to $\boldsymbol{e}_S$. We can therefore write

$$\begin{aligned} &\widehat{\Delta} = (r+G)\boldsymbol{e}_S + \boldsymbol{R}, && G = \boldsymbol{e}_S^\top\boldsymbol{\eta} \sim \mathcal{N}(0, \tau^2), \\ &\boldsymbol{R} = (\boldsymbol{I}_D - \boldsymbol{e}_S\boldsymbol{e}_S^\top)\boldsymbol{e}, && \boldsymbol{R} \perp \boldsymbol{e}_S. \end{aligned} \tag{79}$$

The random variable G is independent of $\boldsymbol{R}$: isotropic Gaussian noise has independent orthogonal projections, and the colors are independent of all noise. In particular,

$$V_\perp \coloneqq \mathbb{E}\|\boldsymbol{R}\|^2 = \frac{2}{n}\left[q_C + q_{SC} + (D-1)\sigma^2\right], \qquad \tan^2\theta = \frac{\|\boldsymbol{R}\|^2}{(r+G)^2}. \tag{80}$$

For the noiseless case $\sigma = 0$, we have:

$$\mathbb{E}[\tan^2\theta] = \frac{q_C + q_{SC}}{2nq_S} = \frac{1}{2n}\left[\frac{1-\alpha}{1+\alpha} + \frac{q_C}{q_S}\right]. \tag{81}$$

B Population derivations for linear networks

This appendix derives the results in §3 and §4 directly from the population loss. We use $m \in \mathcal{M} = \{\emptyset, S, C, SC\}$, with characters $\chi_m(s, c) = 1, s, c, sc$, respectively. Throughout,

$$d_m^{(x)} = \|\hat{\boldsymbol{x}}_m\|^2, \qquad d_m^{(y)} = \|\hat{\boldsymbol{y}}_m\|^2, \qquad q_m(t) = \|\hat{\boldsymbol{z}}_m(t)\|^2$$

are *squared* Fourier-component lengths. The scalar latent s is unrelated to the indexed target energy $d_m^{(y)}$. We retain input noise in the population dynamics and exact solution, then specialize to $\sigma_x = 0$ for overshoot, initialization, and depth. Time denotes ordinary Euclidean gradient flow in the representation weights, with unit learning rate and the readout at its instantaneous ridge optimum.

B.1 Population gradient flow and decoupled Fourier modes

Population moments. The Fourier expansion and uniform latent distribution imply $\mathbb{E}[\chi_m \chi_j] = \delta_{mj}$. Using $\boldsymbol{x} = \sum_m \hat{\boldsymbol{x}}_m \chi_m + \sigma_x \boldsymbol{\epsilon}$, $\boldsymbol{y} = \sum_m \hat{\boldsymbol{y}}_m \chi_m$, and independence of the zero-mean noise with $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^\top] = \boldsymbol{I}$ gives

$$\boldsymbol{\Sigma}_{XX} = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top] = \sum_m \hat{\boldsymbol{x}}_m \hat{\boldsymbol{x}}_m^\top + \sigma_x^2 \boldsymbol{I}_{D_x}, \qquad \boldsymbol{\Sigma}_{XY} = \mathbb{E}[\boldsymbol{x}\boldsymbol{y}^\top] = \sum_m \hat{\boldsymbol{x}}_m \hat{\boldsymbol{y}}_m^\top, \tag{82}$$

$$\mathbb{E}\|\boldsymbol{y}\|^2 = \sum_m d_m^{(y)}. \tag{83}$$

These are uncentered second moments; the constant Fourier component is included. No Gaussian assumption on the input noise is needed for these identities or the following linear training calculation: its first two moments suffice.

To see the noise contribution to the loss explicitly, write $\hat{\boldsymbol{z}}_m = \boldsymbol{W}\hat{\boldsymbol{x}}_m$. Expanding the squared residual, the signal–noise cross term averages to zero, while $\mathbb{E}\|\boldsymbol{A}\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{A}\|_F^2$ for any fixed matrix $\boldsymbol{A}$. Thus (2) becomes

$$\mathcal{L}(\boldsymbol{W}_r, \boldsymbol{W}) = \frac{1}{2}\sum_m \|\boldsymbol{W}_r \hat{\boldsymbol{z}}_m - \hat{\boldsymbol{y}}_m\|^2 + \frac{\sigma_x^2}{2}\|\boldsymbol{W}_r \boldsymbol{W}\|_F^2 + \frac{1}{2\gamma}\|\boldsymbol{W}_r\|_F^2. \tag{84}$$

The middle term is the prediction variance induced by input noise. It remains present even though the class-mean noise is zero.

Eliminate the readout. Use the main-text orientation $\boldsymbol{W} \in \mathbb{R}^{D \times D_x}$ and $\boldsymbol{W}_r \in \mathbb{R}^{D_y \times D}$, so the prediction is $\boldsymbol{W}_r \boldsymbol{W}\boldsymbol{x}$. Differentiating the population loss with respect to $\boldsymbol{W}_r$ yields the normal equation

$$\boldsymbol{W}_r^* \left(\boldsymbol{W}\boldsymbol{\Sigma}_{XX}\boldsymbol{W}^\top + \gamma^{-1}\boldsymbol{I}_D\right) = \boldsymbol{\Sigma}_{XY}^\top \boldsymbol{W}^\top. \tag{85}$$

For every nonzero $\boldsymbol{u}$, $\boldsymbol{u}^\top(\boldsymbol{W}\boldsymbol{\Sigma}_{XX}\boldsymbol{W}^\top + \gamma^{-1}\boldsymbol{I}_D)\boldsymbol{u} = \mathbb{E}[(\boldsymbol{u}^\top \boldsymbol{W}\boldsymbol{x})^2] + \gamma^{-1}\|\boldsymbol{u}\|^2 > 0$. The readout optimum is therefore unique, including for singular input moments, and is

$$\boldsymbol{W}_r^* = \gamma \boldsymbol{\Sigma}_{XY}^\top \boldsymbol{W}^\top \left(\boldsymbol{I}_D + \gamma \boldsymbol{W}\boldsymbol{\Sigma}_{XX}\boldsymbol{W}^\top\right)^{-1}. \tag{86}$$

Completing the square in the readout gives the reduced objective

$$\begin{aligned}\mathcal{F}(\boldsymbol{W}) &\coloneqq \mathcal{L}(\boldsymbol{W}_r^*(\boldsymbol{W}), \boldsymbol{W}) \\ &= \frac{1}{2}\sum_m d_m^{(y)} - \frac{\gamma}{2}\operatorname{Tr}\left[\boldsymbol{\Sigma}_{XY}^\top \boldsymbol{W}^\top (\boldsymbol{I}_D + \gamma \boldsymbol{W}\boldsymbol{\Sigma}_{XX}\boldsymbol{W}^\top)^{-1}\boldsymbol{W}\boldsymbol{\Sigma}_{XY}\right].\end{aligned} \tag{87}$$

The factor $1/4$ for uniform classes has already been included in the expectation defining the Fourier coefficients. There is no sample-size factor in this population loss.

Differentiate at the optimum. The dependence of $\boldsymbol{W}_r^*$ on $\boldsymbol{W}$ creates no additional term in the gradient of $\mathcal{F}$. Indeed, the chain-rule term containing $\nabla_{\boldsymbol{W}_r}\mathcal{L}$ vanishes at its optimum. Define the residual cross moment

$$\boldsymbol{R} \coloneqq \mathbb{E}[\boldsymbol{x}(\boldsymbol{y} - \boldsymbol{W}_r^*\boldsymbol{W}\boldsymbol{x})^\top] = \boldsymbol{\Sigma}_{XY} - \boldsymbol{\Sigma}_{XX}\boldsymbol{W}^\top(\boldsymbol{W}_r^*)^\top. \tag{88}$$

Differentiating the original loss while holding the optimal readout fixed now gives

$$\dot{\boldsymbol{W}} = -\nabla_{\boldsymbol{W}}\mathcal{F} = (\boldsymbol{W}_r^*)^\top \boldsymbol{R}^\top = \gamma \boldsymbol{W}\boldsymbol{R}\boldsymbol{R}^\top. \tag{89}$$

For the second equality on the right, transpose the readout normal equation and rearrange it as $(\boldsymbol{W}_r^*)^\top = \gamma \boldsymbol{W}\boldsymbol{R}$. As a sign check, $d\mathcal{F}/dt = -\|\nabla_{\boldsymbol{W}}\mathcal{F}\|_F^2 \le 0$.

Sufficient initialization for mode decoupling with noisy inputs. This paper largely focuses on the case $\sigma_x \to 0$, and these results only require Assum. 2 as stated. When generalizing to the noisy case $\sigma_x > 0$, exact decoupling of the representation's Fourier modes can be achieved by a *task-aligned initialization* commonly used in previous linear network work (see, e.g. [31, 25, 27]):

$$\boldsymbol{W}(0)^\top \boldsymbol{W}(0)\hat{\boldsymbol{x}}_m = g_{m,0}^2 \hat{\boldsymbol{x}}_m, \quad m \in \mathcal{M}. \tag{90}$$

This can be motivated by the standard practice of using Gaussian initialized weights: independent $W_{ij}(0) \sim \mathcal{N}(0, g_0^2/D)$ give $\mathbb{E}[\boldsymbol{W}(0)^\top \boldsymbol{W}(0)] = g_0^2 \boldsymbol{I}_{D_x}$, so orthogonal input modes have zero expected hidden cross inner products at any width (§A.3).

For each present input mode $d_m^{(x)} > 0$, write $\hat{\boldsymbol{x}}_m = \sqrt{d_m^{(x)}}\boldsymbol{e}_m$, where the $\boldsymbol{e}_m$'s are orthonormal. Choose orthonormal hidden directions $\boldsymbol{u}_m$, and set

$$\boldsymbol{W}(0) = \sum_{m:d_m^{(x)}>0} g_{m,0}\boldsymbol{u}_m\boldsymbol{e}_m^\top + \boldsymbol{W}_\perp, \qquad \boldsymbol{W}_\perp \boldsymbol{e}_m = 0, \quad \boldsymbol{W}_\perp^\top \boldsymbol{u}_m = 0. \tag{91}$$

The second map sends input directions with no class-mean signal into a hidden subspace orthogonal to the signal directions. It may be zero. A scaled identity $\boldsymbol{W}(0) = g_0\boldsymbol{I}$ in a square network realizes the expected Gaussian Gram matrix exactly; allowing separate $g_{m,0}$'s also permits arbitrary positive initial mode energies. The hidden width must accommodate the nonzero input modes. Modes with $d_m^{(x)} = 0$ have $\hat{\boldsymbol{z}}_m = 0$ for every linear representation and require no direction $\boldsymbol{u}_m$.

Consider the candidate trajectory

$$\boldsymbol{W}(t) = \sum_{m:d_m^{(x)}>0} g_m(t)\boldsymbol{u}_m\boldsymbol{e}_m^\top + \boldsymbol{W}_\perp, \qquad \Delta_m = 1 + \gamma(d_m^{(x)} + \sigma_x^2)g_m^2. \tag{92}$$

Its hidden second moment is

$$\boldsymbol{W}\boldsymbol{\Sigma}_{XX}\boldsymbol{W}^\top = \sum_{m:d_m^{(x)}>0} (d_m^{(x)} + \sigma_x^2)g_m^2\boldsymbol{u}_m\boldsymbol{u}_m^\top + \sigma_x^2\boldsymbol{W}_\perp\boldsymbol{W}_\perp^\top. \tag{93}$$

The last term acts orthogonally to all $\boldsymbol{u}_m$. The readout inverse therefore divides the m-th signal direction by Δ_m, giving

$$\boldsymbol{W}_r^* = \sum_{m:d_m^{(x)}>0} \frac{\gamma\sqrt{d_m^{(x)}}g_m}{\Delta_m}\hat{\boldsymbol{y}}_m\boldsymbol{u}_m^\top, \qquad \boldsymbol{W}^\top(\boldsymbol{W}_r^*)^\top = \sum_{m:d_m^{(x)}>0} \frac{\gamma\sqrt{d_m^{(x)}}g_m^2}{\Delta_m}\boldsymbol{e}_m\hat{\boldsymbol{y}}_m^\top. \tag{94}$$

Substituting this into (88) produces a cancellation that is essential in the noisy case:

$$\boldsymbol{R} = \sum_{m:d_m^{(x)}>0} \sqrt{d_m^{(x)}}\left[1 - \frac{\gamma(d_m^{(x)} + \sigma_x^2)g_m^2}{\Delta_m}\right]\boldsymbol{e}_m\hat{\boldsymbol{y}}_m^\top = \sum_{m:d_m^{(x)}>0} \frac{\sqrt{d_m^{(x)}}}{\Delta_m}\boldsymbol{e}_m\hat{\boldsymbol{y}}_m^\top. \tag{95}$$

Target orthogonality then gives

$$\boldsymbol{R}\boldsymbol{R}^\top = \sum_{m:d_m^{(x)}>0} \frac{d_m^{(x)}d_m^{(y)}}{\Delta_m^2}\boldsymbol{e}_m\boldsymbol{e}_m^\top, \qquad \dot{\boldsymbol{W}} = \sum_{m:d_m^{(x)}>0} \frac{\gamma d_m^{(x)}d_m^{(y)}g_m}{\Delta_m^2}\boldsymbol{u}_m\boldsymbol{e}_m^\top. \tag{96}$$

Thus the full gradient changes only the gains:

$$\dot{g}_m = \frac{\gamma d_m^{(x)}d_m^{(y)}g_m}{[1 + \gamma(d_m^{(x)} + \sigma_x^2)g_m^2]^2}, \qquad \dot{\boldsymbol{W}}_\perp = 0. \tag{97}$$

The candidate form is therefore invariant. The readout formula is smooth for $\gamma > 0$, so uniqueness of the weight ODE identifies this candidate with the actual gradient-flow trajectory. In particular, $\hat{\boldsymbol{z}}_m = \sqrt{d_m^{(x)}}g_m\boldsymbol{u}_m$ remain mutually orthogonal.

Mode energies and the noise factor. For a present input mode, define

$$q_m = d_m^{(x)} g_m^2, \qquad b_m = 1 + \frac{\sigma_x^2}{d_m^{(x)}}, \qquad \kappa_m(q_m) = \frac{\gamma d_m^{(x)} d_m^{(y)}}{(1+\gamma b_m q_m)^2}. \tag{98}$$

Differentiating the squared gain gives the mode equation

$$\dot{q}_m = 2 d_m^{(x)} g_m \dot{g}_m = 2 q_m \kappa_m(q_m) = \frac{2\gamma d_m^{(x)} d_m^{(y)} q_m}{(1+\gamma b_m q_m)^2}. \tag{99}$$

The within-class noise variance along $\boldsymbol{u}_m$ is $\sigma_x^2 g_m^2$, so the total hidden second moment along that direction is $(d_m^{(x)} + \sigma_x^2) g_m^2 = b_m q_m$. The representation amplifies input noise along with the class-mean signal; this is why the denominator contains $b_m q_m$, rather than $q_m + \sigma_x^2$.

An independent check follows from the minimized scalar loss on this invariant family:

$$\mathcal{F}_m(g_m) = \frac{d_m^{(y)}}{2} \frac{1 + \gamma \sigma_x^2 g_m^2}{1 + \gamma (d_m^{(x)} + \sigma_x^2) g_m^2}, \qquad \frac{d\mathcal{F}_m}{dg_m} = -\frac{\gamma d_m^{(x)} d_m^{(y)} g_m}{\Delta_m^2}. \tag{100}$$

The full minimized loss is $\mathcal{F} = \sum_{m: d_m^{(x)} > 0} \mathcal{F}_m + \frac{1}{2} \sum_{m: d_m^{(x)} = 0} d_m^{(y)}$; targets in absent input modes contribute a constant error. The squared coordinate q_m is an observable, not a Euclidean weight parameter. Accordingly its flow is $\dot{q}_m = -4 d_m^{(x)} q_m \, \partial \mathcal{F}_m / \partial q_m$, which agrees with (99).

Abstraction from the two relevant energies. The constant and color components cancel when shape is changed at fixed color. Hence

$$\boldsymbol{u}_{\text{red}} = 2(\hat{\boldsymbol{z}}_S + \hat{\boldsymbol{z}}_{SC}), \qquad \boldsymbol{u}_{\text{blue}} = 2(\hat{\boldsymbol{z}}_S - \hat{\boldsymbol{z}}_{SC}). \tag{101}$$

Their inner product is $4(q_S - q_{SC})$, and both squared lengths are $4(q_S + q_{SC})$. Whenever this sum is positive,

$$\alpha = \frac{q_S - q_{SC}}{q_S + q_{SC}} = \frac{1 - \nu_z}{1 + \nu_z}, \qquad \nu_z = \frac{q_{SC}}{q_S}. \tag{102}$$

The ratio expression assumes $q_S > 0$; the first expression also covers the boundary $q_S = 0 < q_{SC}$. For positive energies,

$$\frac{\dot{\nu}_z}{\nu_z} = \frac{\dot{q}_{SC}}{q_{SC}} - \frac{\dot{q}_S}{q_S} = 2(\kappa_{SC} - \kappa_S), \tag{103}$$

$$\dot{\alpha} = -\frac{2\dot{\nu}_z}{(1+\nu_z)^2} = (1 - \alpha^2)(\kappa_S - \kappa_{SC}). \tag{104}$$

These are the ODEs in (10) and (11). Noise does not appear explicitly in the angle of the class means, but it affects that angle through the two growth rates.

B.2 Exact solutions for the Fourier energies and abstraction

B.2.1 Integrating a single mode

For $d_m^{(x)} > 0$ and $q_m(0) > 0$, define the dimensionless energy $\tilde{q}_m = \gamma b_m q_m$. Its ODE is

$$\dot{\tilde{q}}_m = \frac{2\gamma d_m^{(x)} d_m^{(y)} \tilde{q}_m}{(1+\tilde{q}_m)^2}. \tag{105}$$

Separate variables and integrate from the initial value:

$$\int_{\tilde{q}_{m,0}}^{\tilde{q}_m(t)} \frac{(1+a)^2}{a} \, da = 2\gamma d_m^{(x)} d_m^{(y)} t, \qquad \frac{(1+a)^2}{a} = a + 2 + \frac{1}{a}. \tag{106}$$

Thus, for $a > 0$, defining

$$F(a) = \frac{1}{2} a^2 + 2a + \log a, \qquad F'(a) = \frac{(1+a)^2}{a} > 0, \tag{107}$$

we obtain

$$F(\tilde{q}_m(t)) = F(\tilde{q}_{m,0}) + 2\gamma d_m^{(x)} d_m^{(y)} t. \tag{108}$$

Because $F(a) \to -\infty$ as $a \downarrow 0$ and $F(a) \to \infty$ as $a \to \infty$, strict monotonicity makes F a bijection from $(0, \infty)$ onto $\mathbb{R}$. Its inverse therefore defines a unique trajectory at every finite time.

Proposition B.1 (Exact population mode trajectory)**.** *For $d_m^{(x)} > 0$, $q_m(0) > 0$, and $d_m^{(y)} \geq 0$,*

$$q_m(t) = \frac{1}{\gamma b_m} F^{-1}\Big(F(\gamma b_m q_m(0)) + 2\gamma d_m^{(x)} d_m^{(y)} t\Big). \tag{109}$$

If $d_m^{(y)} = 0$, this expression is the constant $q_m(0)$. A zero initialized energy stays zero and is treated directly by the ODE.

The solution can also be written without a rescaled variable:

$$\log \frac{q_m(t)}{q_m(0)} + 2\gamma b_m[q_m(t) - q_m(0)] + \frac{\gamma^2 b_m^2}{2}[q_m(t)^2 - q_m(0)^2] = 2\gamma d_m^{(x)} d_m^{(y)} t. \tag{110}$$

Differentiating either implicit expression recovers the mode ODE and setting $t = 0$ recovers the initialization. These checks also fix the placement of the time shift: it is added to $F(\tilde{q}_{m,0})$ *after* evaluating F.

B.2.2 Combining the shape and interaction modes

Assume $d_S^{(x)}, d_{SC}^{(x)} > 0$, $d_S^{(y)} > 0$, and positive initial energies. Introduce the shared clock and two fixed ratios

$$\tau = 2\gamma d_S^{(x)} d_S^{(y)} t, \qquad r = \frac{d_{SC}^{(x)} d_{SC}^{(y)}}{d_S^{(x)} d_S^{(y)}} = \nu_x \nu_y, \qquad B = \frac{b_S}{b_{SC}} = \frac{1 + \sigma_x^2/d_S^{(x)}}{1 + \sigma_x^2/d_{SC}^{(x)}}. \tag{111}$$

The two exact mode solutions are

$$\tilde{q}_S(\tau) = F^{-1}(F(\tilde{q}_{S,0}) + \tau), \qquad \tilde{q}_{SC}(\tau) = F^{-1}(F(\tilde{q}_{SC,0}) + r\tau). \tag{112}$$

Converting back to the physical Fourier energies is mode dependent:

$$\nu_z = \frac{q_{SC}}{q_S} = B\frac{\tilde{q}_{SC}}{\tilde{q}_S}, \qquad \alpha = \frac{\tilde{q}_S - B\tilde{q}_{SC}}{\tilde{q}_S + B\tilde{q}_{SC}} = \tanh\left(\frac{1}{2}\log\frac{\tilde{q}_S}{B\tilde{q}_{SC}}\right). \tag{113}$$

Substituting (112) into this equation proves Theorem 1. Omitting B would incorrectly identify a ratio of noise-rescaled energies with the ratio of class-mean energies. In the noiseless case $B = 1$.

Eliminating time also gives a useful conserved relation:

$$F(\tilde{q}_{SC}) - rF(\tilde{q}_S) = F(\tilde{q}_{SC,0}) - rF(\tilde{q}_{S,0}) =: C. \tag{114}$$

Equivalently, the abstraction trajectory can be parameterized by the strictly increasing shape energy:

$$F\left(\frac{\tilde{q}_S}{B}\frac{1-\alpha}{1+\alpha}\right) - rF(\tilde{q}_S) = C, \tag{115}$$

$$t = \frac{F(\tilde{q}_S) - F(\tilde{q}_{S,0})}{2\gamma d_S^{(x)} d_S^{(y)}}. \tag{116}$$

These formulas do not require an elementary expression for F^{-1}: a monotone scalar root solve suffices to evaluate it.

B.3 Early-time growth, terminal abstraction, and boundary cases

The small-energy regime. When both the initial and current dimensionless energies are small, $F(a) = \log a + O(a)$. The exact solution then gives

$$\log \frac{\tilde{q}_m(t)}{\tilde{q}_{m,0}} = 2\gamma d_m^{(x)} d_m^{(y)} t + O(\tilde{q}_m(t) + \tilde{q}_{m,0}), \qquad q_m(t) \approx q_m(0) e^{2\gamma d_m^{(x)} d_m^{(y)} t}. \tag{117}$$

The approximation is valid while the energies remain small, rather than for all times after a small initialization. If both relevant modes are in this regime, their ratio and abstraction obey

$$\nu_z(t) \approx \nu_z(0) e^{2\gamma(d_{SC}^{(x)} d_{SC}^{(y)} - d_S^{(x)} d_S^{(y)})t}, \tag{118}$$

$$\alpha(t) \approx \tanh\Big[\operatorname{arctanh}\alpha(0) + \gamma(d_S^{(x)} d_S^{(y)} - d_{SC}^{(x)} d_{SC}^{(y)})t\Big]. \tag{119}$$

The initial improvement of abstraction in the signal-dominant regime comes from faster fractional growth of the shape mode.

The large-energy regime. For $u \to \infty$, inversion of $F(a) = a^2/2 + 2a + \log a$ gives

$$F^{-1}(u) = \sqrt{2u} - 2 + O\left(\frac{\log u}{\sqrt{u}}\right). \tag{120}$$

One way to verify the remainder is to insert $a = \sqrt{2u} - 2$ into F. The discrepancy is $O(\log u)$, while $F'(a)$ is of order $\sqrt{u}$; the resulting correction to a is therefore $O(\log u/\sqrt{u})$. For an active mode, meaning $d_m^{(x)} d_m^{(y)} > 0$ and $q_m(0) > 0$, this yields

$$\tilde{q}_m(t) = 2\sqrt{\gamma d_m^{(x)} d_m^{(y)} t} - 2 + O\left(\frac{\log t}{\sqrt{t}}\right), \qquad q_m(t) \sim \frac{2}{b_m}\sqrt{\frac{d_m^{(x)} d_m^{(y)}}{\gamma}}\, t. \tag{121}$$

The fixed initialization shift in F does not affect the leading coefficient. If both shape and interaction modes are active, taking their ratio proves

$$\nu_{z,\infty} = B\sqrt{\nu_x \nu_y}, \qquad \alpha_\infty = \frac{1 - B\sqrt{\nu_x \nu_y}}{1 + B\sqrt{\nu_x \nu_y}}. \tag{122}$$

When $\sigma_x = 0$, this becomes the terminal law in Proposition 2:

$$\alpha_\infty = \frac{1 - \sqrt{\nu_x \nu_y}}{1 + \sqrt{\nu_x \nu_y}}. \tag{123}$$

For noisy inputs with unequal mode energies, the additional factor B is necessary. The limiting angle is independent of positive initialization scales and finite $\gamma > 0$, although the finite-time trajectory depends on both.

The word "terminal" refers here to the relative geometry. Active q_m's diverge as $t^{1/2}$, and their gains grow as $t^{1/4}$; there is no finite positive fixed point for the unregularized hidden weights. The readout shrinks while the prediction converges. In an active mode, the residual loss approaches

$$\lim_{t\to\infty} \mathcal{F}_m = \frac{d_m^{(y)}}{2} \frac{\sigma_x^2}{d_m^{(x)} + \sigma_x^2}, \tag{124}$$

which is zero in the noiseless case.

Approach to the limiting angle. For positive shape and interaction target energies, let $r = \nu_x \nu_y > 0$ and use the clock $\tau = 2\gamma d_S^{(x)} d_S^{(y)} t$. Applying (120) to both modes gives

$$\begin{aligned} \nu_z(\tau) &= B\sqrt{r}\left[1 + \frac{2(1 - r^{-1/2})}{\sqrt{2\tau}} + O\left(\frac{\log \tau}{\tau}\right)\right], \\ \alpha(\tau) &= \alpha_\infty + \frac{(1 - \alpha_\infty^2)(r^{-1/2} - 1)}{\sqrt{2\tau}} + O\left(\frac{\log \tau}{\tau}\right). \end{aligned} \tag{125}$$

Thus for $0 < r < 1$ abstraction approaches its limit from above at sufficiently large finite times. This motivates the exact overshoot calculation below. The formula is not used at $r = 0$, where one mode does not grow.

In the rest of the linear appendix, we set $\sigma_x = 0$, so $b_m = B = 1$ and $\tilde{q}_m = \gamma q_m$.

B.4 Non-monotonicity and overshoot

We now specialize to noiseless inputs, $\sigma_x = 0$, as in the main-text analysis of the trajectory. The shape and interaction Fourier energies are sufficient to determine abstraction. To keep the calculation separate from the binary latent s and target energies $d_m^{(y)}$, introduce

$$a = \gamma q_S, \qquad b = \gamma q_{SC}, \qquad r = \frac{b}{a} = \nu_z, \qquad \rho = \sqrt{\nu_x \nu_y}, \qquad \tau = 2\gamma d_S^{(x)} d_S^{(y)} t. \tag{126}$$

Thus a, b are dimensionless mode energies, r is their entanglement ratio, and τ is a common learning clock. Throughout this subsection assume $d_S^{(x)} d_S^{(y)} > 0$, $a_0, b_0 > 0$, and

$$0 < \rho < 1, \qquad r_0 = \frac{b_0}{a_0} > \rho. \tag{127}$$

The first condition is the nondegenerate signal-dominant regime. The second is equivalent to $\alpha_0 < \alpha_\infty$, since

$$\alpha = \frac{1-r}{1+r}, \qquad \frac{d\alpha}{dr} = -\frac{2}{(1+r)^2} < 0, \qquad \alpha_\infty = \frac{1-\rho}{1+\rho}. \tag{128}$$

The case $\rho = 0$ has terminal abstraction one, so strict overshoot above the terminal value is impossible. It must be excluded from the strict statements below.

In the rescaled time, the mode ODEs and their exact integrals are

$$\frac{da}{d\tau} = \frac{a}{(1+a)^2}, \qquad \frac{db}{d\tau} = \frac{\rho^2 b}{(1+b)^2}, \tag{129}$$

$$F(a) = F(a_0) + \tau, \qquad F(b) = F(b_0) + \rho^2\tau, \qquad F(u) = \tfrac{1}{2}u^2 + 2u + \log u. \tag{130}$$

Both energies increase, even when their ratio decreases. Subtracting ρ^2 times the first integral from the second eliminates time:

$$F(b) - \rho^2 F(a) = C, \qquad C = F(b_0) - \rho^2 F(a_0). \tag{131}$$

This identity describes the trajectory in the (a, b)-plane. We will intersect it with two lines: $b = \rho a$, where abstraction equals its terminal value, and the line on which the ratio stops changing.

Crossing the terminal ratio. At $b = \rho a$, the trajectory equation becomes $H_\infty(a) = C$, where

$$\begin{aligned} H_\infty(a) &= F(\rho a) - \rho^2 F(a) = 2\rho(1-\rho)a + \log\rho + (1-\rho^2)\log a, \\ H'_\infty(a) &= 2\rho(1-\rho) + \frac{1-\rho^2}{a} > 0. \end{aligned} \tag{132}$$

The limits of H_∞ are $-\infty$ at zero and $+\infty$ at infinity, so the crossing energy $a_{\text{cross}} > 0$ is unique. Moreover, $b_0 > \rho a_0$ and the strict increase of F imply $C > H_\infty(a_0)$, hence $a_{\text{cross}} > a_0$. The crossing therefore occurs at a finite positive time,

$$T_{\text{cross}} = \frac{F(a_{\text{cross}}) - F(a_0)}{2\gamma d_S^{(x)} d_S^{(y)}}. \tag{133}$$

To determine which side of the line the trajectory occupies, observe that

$$r > \rho \quad \Longleftrightarrow \quad b > \rho a \quad \Longleftrightarrow \quad C > H_\infty(a). \tag{134}$$

Because a increases with time, $r > \rho$ before the crossing and $r < \rho$ at every finite time after it. Thus abstraction crosses its terminal value once and subsequently approaches that value from above.

The unique peak. Differentiate the logarithm of the ratio:

$$\frac{1}{r}\frac{dr}{d\tau} = \frac{1}{b}\frac{db}{d\tau} - \frac{1}{a}\frac{da}{d\tau} = \frac{\rho^2}{(1+b)^2} - \frac{1}{(1+a)^2}. \tag{135}$$

Since the denominators are positive, the ratio is stationary precisely on

$$b = \rho(1+a) - 1, \qquad a > a_c := \frac{1-\rho}{\rho}. \tag{136}$$

This line is the *nullcline* of the ratio: above it r decreases, and below it r increases. The restriction $a > a_c$ ensures that its interaction coordinate is positive.

Substitute the line into the conserved trajectory and define

$$G(a) = F\big(\rho(1+a) - 1\big) - \rho^2 F(a), \qquad a > a_c. \tag{137}$$

Using $F'(u) = (1+u)^2/u$, we find

$$\begin{aligned} G'(a) &= \rho F'\big(\rho(1+a)-1\big) - \rho^2 F'(a) \\ &= \rho^2(1+a)^2\left[\frac{\rho}{\rho(1+a)-1} - \frac{1}{a}\right] = \frac{\rho^2(1+a)^2(1-\rho)}{a[\rho(1+a)-1]} > 0. \end{aligned} \tag{138}$$

At the lower endpoint $G(a) \to -\infty$, because the argument of the first F tends to zero. At infinity,

$$G(a) = \rho(1-\rho)a + (1-\rho^2)\log a + O(1) \longrightarrow +\infty. \tag{139}$$

Hence there is a unique $a_* > a_c$ with

$$G(a_*) = C. \tag{140}$$

It lies after initialization. If $a_0 \le a_c$, this is immediate. If $a_0 > a_c$, then $b_0 - [\rho(1+a_0) - 1] = a_0(r_0 - \rho) + 1 - \rho > 0$, giving $C > G(a_0)$ and again $a_* > a_0$.

For $a_c < a < a_*$, the inequality $G(a) < C$, together with the strict increase of F, gives $b > \rho(1+a) - 1$, so $dr/d\tau < 0$. For $a \le a_c$, this same sign follows because $b > 0$ while the nullcline is nonpositive. For $a > a_*$, the inequalities reverse and $dr/d\tau > 0$. The unique peak time and minimum ratio are therefore

$$t_* = \frac{F(a_*) - F(a_0)}{2\gamma d_S^{(x)} d_S^{(y)}}, \tag{141}$$

$$r_* = \frac{\rho(1+a_*) - 1}{a_*} = \rho - \frac{1-\rho}{a_*}, \qquad 0 < r_* < \rho. \tag{142}$$

The peak occurs after the terminal-ratio crossing because $r_* < \rho$ and the ratio decreases until the peak.

Proposition B.2 (Overshoot and a unique maximum)**.** *Under (127), abstraction increases strictly to a unique maximum at t_*, then decreases strictly to α_∞. It crosses α_∞ exactly once, at $0 < T_{\text{cross}} < t_*$, and $\alpha(t_*) = (1-r_*)/(1+r_*) > \alpha_\infty$.*

The calculation identifies the mechanism. At small mode energies the relative rates are approximately one and $\rho^2 < 1$, so the shape mode grows faster. At larger energies, its ridge denominator reduces its relative growth rate. The ratio eventually turns around and returns to the terminal value ρ.

B.5 Initialization scale

Keep $0 < \rho < 1$ and $r_0 > \rho$ fixed, and vary the common scale

$$a_0 = \zeta, \qquad b_0 = \zeta r_0, \qquad \zeta > 0. \tag{143}$$

The preceding calculation applies to every member of this family. Its trajectory constant is

$$\begin{aligned} C_\zeta &= F(\zeta r_0) - \rho^2 F(\zeta) \\ &= \tfrac{1}{2}\zeta^2(r_0^2 - \rho^2) + 2\zeta(r_0 - \rho^2) + (1-\rho^2)\log\zeta + \log r_0, \end{aligned} \tag{144}$$

and its peak is determined by a single scalar equation,

$$G(a_*(\zeta)) = C_\zeta, \qquad r_*(\zeta) = \rho - \frac{1-\rho}{a_*(\zeta)}, \qquad \alpha_{\max}(\zeta) = \frac{1 - r_*(\zeta)}{1 + r_*(\zeta)}. \tag{145}$$

No integration of the learning ODE is required to find the maximum.

Why larger initialization lowers the maximum. Differentiating the trajectory constant gives

$$C'_\zeta = \zeta(r_0^2 - \rho^2) + 2(r_0 - \rho^2) + \frac{1-\rho^2}{\zeta} > 0. \tag{146}$$

Every term is positive: $r_0 > \rho$, and $\rho > \rho^2$ for $0 < \rho < 1$. Since $G'(a) > 0$, implicit differentiation gives

$$\frac{da_*}{d\zeta} = \frac{C'_\zeta}{G'(a_*)} > 0, \qquad \frac{dr_*}{d\zeta} = \frac{1-\rho}{a_*^2}\frac{C'_\zeta}{G'(a_*)} > 0. \tag{147}$$

Finally,

$$\frac{d\alpha_{\max}}{d\zeta} = -\frac{2(1-\rho)}{(1+r_*)^2 a_*^2}\frac{C'_\zeta}{G'(a_*)} < 0. \tag{148}$$

This proves the strict monotonicity in Proposition 3.

Small- and large-initialization limits. As $\zeta \downarrow 0$,

$$C_\zeta = (1-\rho^2)\log\zeta + \log r_0 + o(1) \longrightarrow -\infty. \tag{149}$$

The only way $G(a_*)$ can tend to $-\infty$ is for $a_* \downarrow a_c = (1-\rho)/\rho$. Equation (145) therefore gives

$$r_*(\zeta) \longrightarrow 0, \qquad \alpha_{\max}(\zeta) \longrightarrow 1. \tag{150}$$

As $\zeta \to \infty$, instead,

$$C_\zeta = \tfrac{1}{2}\zeta^2(r_0^2 - \rho^2) + O(\zeta) \longrightarrow +\infty, \tag{151}$$

so $a_* \to \infty$, $r_* \to \rho$, and

$$\alpha_{\max}(\zeta) \longrightarrow \frac{1-\rho}{1+\rho} = \alpha_\infty. \tag{152}$$

For every finite initialization scale the maximum remains strictly above α_∞; its excess vanishes in the large-scale limit. The terminal score is set by the data energies through ρ, whereas the maximum also depends on the trajectory constant C_ζ.

B.6 Rich-limit residence-time metastability

We retain the initialization family above and quantify how long abstraction stays close to one. Fix a threshold strictly above the terminal value,

$$0 < \varepsilon < 1 - \alpha_\infty, \qquad r_\varepsilon = \frac{\varepsilon}{2-\varepsilon}. \tag{153}$$

Then

$$\alpha \geq 1-\varepsilon \quad \Longleftrightarrow \quad r \leq r_\varepsilon, \qquad 0 < r_\varepsilon < \rho < r_0. \tag{154}$$

For sufficiently small ζ, the minimum ratio lies below r_ε. The unique-peak result then ensures one entrance and one exit, so the total residence time is the exit time minus the entrance time.

Locate the two threshold crossings. At a crossing $b = r_\varepsilon a$, the trajectory equation is

$$\begin{aligned} H_\varepsilon(a) &= C_\zeta, \\ H_\varepsilon(a) &\coloneqq F(r_\varepsilon a) - \rho^2 F(a) \\ &= \tfrac{1}{2}(r_\varepsilon^2 - \rho^2)a^2 + 2(r_\varepsilon - \rho^2)a \\ &\quad + \log r_\varepsilon + (1-\rho^2)\log a. \end{aligned} \tag{155}$$

This function tends to $-\infty$ at both zero and infinity. Furthermore,

$$aH_\varepsilon'(a) = -(\rho^2 - r_\varepsilon^2)a^2 + 2(r_\varepsilon - \rho^2)a + (1-\rho^2). \tag{156}$$

The quadratic has negative leading coefficient and positive constant term, hence exactly one positive root. Thus H_ε first increases and then decreases. Since $C_\zeta \to -\infty$ as $\zeta \downarrow 0$, there are exactly two positive crossing energies for all sufficiently small ζ: $a_\varepsilon^{\mathrm{in}} < a_\varepsilon^{\mathrm{out}}$. The entrance is after initialization. Indeed, ζ lies below the fixed maximizer of H_ε when it is sufficiently small, and $H_\varepsilon(\zeta) < C_\zeta$ because $r_\varepsilon < r_0$. The corresponding physical times are

$$T_\varepsilon^{\mathrm{in/out}}(\zeta) = \frac{F(a_\varepsilon^{\mathrm{in/out}}) - F(\zeta)}{2\gamma d_S^{(x)} d_S^{(y)}}, \qquad R_\varepsilon = T_\varepsilon^{\mathrm{out}} - T_\varepsilon^{\mathrm{in}}. \tag{157}$$

The entrance time stays finite. The smaller root tends to zero. Using $F(u) = \log u + O(u)$ in (155) gives

$$(1-\rho^2)\log a_\varepsilon^{\mathrm{in}} + \log r_\varepsilon + o(1) = (1-\rho^2)\log\zeta + \log r_0 + o(1). \tag{158}$$

Therefore

$$\log\frac{a_\varepsilon^{\mathrm{in}}}{\zeta} = \frac{1}{1-\rho^2}\log\frac{r_0}{r_\varepsilon} + o(1), \tag{159}$$

and the entrance-time formula becomes

$$T_\varepsilon^{\mathrm{in}}(\zeta) = \frac{1}{2\gamma d_S^{(x)} d_S^{(y)}(1-\rho^2)}\log\frac{r_0}{r_\varepsilon} + o(1) = O(1). \tag{160}$$

The time needed to enter a fixed near-perfect band consequently has a finite limit as the initialization scale shrinks.

The exit time diverges logarithmically. The larger root tends to infinity. Set $\Lambda_\zeta = \log(1/\zeta)$. Since $C_\zeta = -(1-\rho^2)\Lambda_\zeta + O(1)$, the leading quadratic term in H_ε implies $a_\varepsilon^{\mathrm{out}} = \Theta(\sqrt{\Lambda_\zeta})$. Keeping the lower-order linear and logarithmic terms then gives

$$\left(a_\varepsilon^{\mathrm{out}}\right)^2 = \frac{2(1-\rho^2)}{\rho^2 - r_\varepsilon^2}\Lambda_\zeta + O(\sqrt{\Lambda_\zeta}). \tag{161}$$

Substitute this into (157). The large root contributes $F(a_\varepsilon^{\mathrm{out}}) = (1-\rho^2)\Lambda_\zeta/(\rho^2 - r_\varepsilon^2) + O(\sqrt{\Lambda_\zeta})$, while $-F(\zeta) = \Lambda_\zeta + O(1)$. Hence

$$T_\varepsilon^{\mathrm{out}}(\zeta) = \frac{1 - r_\varepsilon^2}{2\gamma d_S^{(x)} d_S^{(y)}(\rho^2 - r_\varepsilon^2)} \log\frac{1}{\zeta} + O\left(\sqrt{\log\frac{1}{\zeta}}\right). \tag{162}$$

Subtracting the bounded entrance time gives the more precise residence law

$$R_\varepsilon(\zeta) = \frac{1 - r_\varepsilon^2}{2\gamma d_S^{(x)} d_S^{(y)}(\rho^2 - r_\varepsilon^2)} \log\frac{1}{\zeta} + O\left(\sqrt{\log\frac{1}{\zeta}}\right) = \Theta\left(\log\frac{1}{\zeta}\right). \tag{163}$$

This proves Proposition 4. All data energies, γ, r_0, and the threshold are fixed in this limit. The constants need not remain uniform if the threshold approaches α_∞ or the target interaction energy tends to zero.

Fixed-time behavior and the order of limits. The residence law concerns a long interval of *near*-perfect abstraction; it does not say that every positive time has score exactly one. The limiting trajectory on any fixed training horizon can be found directly. Since the dimensionless ODEs imply

$$a(\tau;\zeta) \le \zeta e^{\tau}, \qquad b(\tau;\zeta) \le \zeta r_0 e^{\rho^2 \tau}, \tag{164}$$

both energies tend uniformly to zero on each bounded τ-interval. Integrating their logarithmic derivatives then yields

$$\begin{aligned} \log\frac{a(\tau;\zeta)}{\zeta} &= \int_0^\tau \frac{du}{[1+a(u;\zeta)]^2} \longrightarrow \tau, \\ \log\frac{b(\tau;\zeta)}{\zeta r_0} &= \int_0^\tau \frac{\rho^2\, du}{[1+b(u;\zeta)]^2} \longrightarrow \rho^2\tau. \end{aligned} \tag{165}$$

Consequently,

$$\begin{aligned} \lim_{\zeta\downarrow 0} r(\tau;\zeta) &= r_0 e^{-(1-\rho^2)\tau}, \\ \lim_{\zeta\downarrow 0} \alpha(\tau;\zeta) &= \frac{1 - r_0 e^{-(1-\rho^2)\tau}}{1 + r_0 e^{-(1-\rho^2)\tau}}. \end{aligned} \tag{166}$$

This expression reaches any fixed threshold $1-\varepsilon$ after the finite limiting entrance time above, and then remains above it over any prescribed finite remaining horizon as $\zeta \downarrow 0$. In contrast, taking infinite training time first returns the terminal value for every positive initialization scale. Therefore

$$\lim_{\zeta\downarrow 0}\lim_{t\to\infty} \alpha(t;\zeta) = \frac{1-\rho}{1+\rho}, \qquad \lim_{t\to\infty}\lim_{\zeta\downarrow 0} \alpha(t;\zeta) = 1. \tag{167}$$

These are limits of scores for positive ζ. At exactly zero initialization the hidden concept vectors vanish, so their cosine is undefined.

B.7 Depth dynamics

We now derive the results in §4.2 for a network with L hidden linear layers. Throughout this subsection $\sigma_x = 0$, so inputs and targets are deterministic functions of the two latents. The readout is optimized at each time, and every hidden layer follows Euclidean gradient flow with the same learning rate, absorbed into time. The quantities $d_m^{(x)} = \|\hat{\boldsymbol{x}}_m\|^2$ and $d_m^{(y)} = \|\hat{\boldsymbol{y}}_m\|^2$ are energies. In particular, $q_m^{(0)} = d_m^{(x)}$ is the input Fourier energy, and $q_m^{(\ell)}$ denotes the corresponding energy after hidden layer ℓ.

An aligned family of multilayer representations. Write

$$\boldsymbol{z}^{(\ell)} = \boldsymbol{W}_\ell \cdots \boldsymbol{W}_1 \boldsymbol{x}, \qquad \boldsymbol{f} = \boldsymbol{W}_r \boldsymbol{z}^{(L)}. \tag{168}$$

For every present input mode, let $\boldsymbol{e}_m^{(0)} = \hat{\boldsymbol{x}}_m / \sqrt{d_m^{(x)}}$. Choose orthonormal vectors $\boldsymbol{e}_m^{(\ell)}$ in each hidden layer. Thus each layer must be wide enough to contain the present input modes, but its width need not equal the input dimension or the widths of other layers. A concrete aligned initialization is

$$\boldsymbol{W}_j(0) = \sum_{m: d_m^{(x)} > 0} g_{j,m}(0) \boldsymbol{e}_m^{(j)} (\boldsymbol{e}_m^{(j-1)})^\top, \qquad g_{j,m}(0) > 0. \tag{169}$$

The factors are zero on the unused orthogonal complements. One may also include decoupled blocks there, provided they do not mix the displayed mode directions. The positive-gain convention fixes signs for convenience; the energy calculations depend only on squared gains.

This family stays aligned under gradient flow. To see both the geometry and the update, let $p_m = \prod_{j=1}^{L} g_{j,m}$. On the aligned family,

$$\hat{\boldsymbol{z}}_m^{(\ell)} = \sqrt{d_m^{(x)}} \left(\prod_{j=1}^{\ell} g_{j,m} \right) \boldsymbol{e}_m^{(\ell)}, \qquad q_m^{(\ell)} = d_m^{(x)} \prod_{j=1}^{\ell} g_{j,m}^2, \tag{170}$$

$$q_m \equiv q_m^{(L)} = d_m^{(x)} p_m^2. \tag{171}$$

Hidden Fourier vectors are therefore orthogonal at every layer. The optimal readout and its prediction in each mode are

$$\boldsymbol{W}_r^* = \sum_{m: d_m^{(x)} > 0} \frac{\gamma \sqrt{d_m^{(x)}} p_m}{1 + \gamma q_m} \hat{\boldsymbol{y}}_m (\boldsymbol{e}_m^{(L)})^\top, \qquad \boldsymbol{W}_r^* \hat{\boldsymbol{z}}_m^{(L)} = \frac{\gamma q_m}{1 + \gamma q_m} \hat{\boldsymbol{y}}_m. \tag{172}$$

Consequently the minimized population loss is

$$\mathcal{L}^* = \frac{1}{2} \sum_m \frac{d_m^{(y)}}{1 + \gamma q_m}, \qquad \mu_m(q_m) \coloneqq -2 \frac{\partial \mathcal{L}^*}{\partial q_m} = \frac{\gamma d_m^{(y)}}{(1 + \gamma q_m)^2}. \tag{173}$$

An absent input mode contributes the constant $d_m^{(y)}/2$ to this loss.

For completeness, the full layer gradient, evaluated on the aligned family, is

$$\nabla_{\boldsymbol{W}_j} \mathcal{L}^* = - \sum_{m: d_m^{(x)} > 0} \mu_m d_m^{(x)} g_{j,m} \left(\prod_{k \neq j} g_{k,m}^2 \right) \boldsymbol{e}_m^{(j)} (\boldsymbol{e}_m^{(j-1)})^\top. \tag{174}$$

Indeed, the prediction residual in mode m is $-\hat{\boldsymbol{y}}_m / (1 + \gamma q_m)$. Backpropagating through $\boldsymbol{W}_r^*$ and layers $L, \ldots, j+1$, then multiplying by the mode's input to layer j, produces the displayed coefficient. Character orthogonality removes products of distinct latent modes, and target-vector orthogonality makes the backpropagated residual point along $\boldsymbol{e}_m^{(L)}$. Thus the gradient has no component outside the aligned family. This verifies that the scalar calculation describes the full weight dynamics for these initial conditions.

For nonzero gains, the chain rule now gives

$$\frac{\partial q_m}{\partial g_{j,m}} = \frac{2 q_m}{g_{j,m}}, \qquad \dot{g}_{j,m} = -\frac{\partial \mathcal{L}^*}{\partial g_{j,m}} = \mu_m \frac{q_m}{g_{j,m}}, \tag{175}$$

$$\dot{q}_m = \sum_{j=1}^{L} \frac{2 q_m}{g_{j,m}} \dot{g}_{j,m} = 2 \mu_m q_m^2 \sum_{j=1}^{L} g_{j,m}^{-2}. \tag{176}$$

The polynomial expression in (174), rather than division by $g_{j,m}$, is used at zero gains.

Why balanced gains remain balanced. Let $h_{j,m} = g_{j,m}^2$. Multiplying the gain equation by $2g_{j,m}$ gives

$$\dot{h}_{j,m} = 2\mu_m q_m = -4q_m \frac{\partial \mathcal{L}^*}{\partial q_m}. \tag{177}$$

The right-hand side is independent of the layer index. We have therefore proved the following statement.

Lemma B.3 (Persistence of mode-wise balancedness)**.** *Within the aligned family, gradient flow with a common hidden-layer learning rate preserves every difference $g_{j,m}^2 - g_{k,m}^2$. In particular, equal squared gains at initialization remain equal throughout training.*

Thus the equal initial gains in Setting 2 remain equal throughout training. We therefore drop the layer index and write $g_m(t)$ for the common positive gain of mode m:

$$g_{1,m} = \cdots = g_{L,m} = g_m, \qquad q_m^{(\ell)} = d_m^{(x)} g_m^{2\ell}. \tag{178}$$

For example, a prescribed positive final-layer energy $q_{m,0}$ is obtained by choosing

$$g_m(0)^2 = \left(\frac{q_{m,0}}{d_m^{(x)}}\right)^{1/L}. \tag{179}$$

In matrix form, this is the familiar balanced initialization condition $\boldsymbol{W}_{j+1}^\top \boldsymbol{W}_{j+1} = \boldsymbol{W}_j \boldsymbol{W}_j^\top$ on the active subspace, expressed in the input Fourier directions [41, 26]. The readout is not included among these balanced factors: it is the separately optimized ridge solution.

Since $q_m = d_m^{(x)} g_m^{2L}$, we have $\sum_{j=1}^L g_{j,m}^{-2} = L g_m^{-2} = L(d_m^{(x)}/q_m)^{1/L}$. Substitution into (176) yields the scalar ODE stated in §4.2:

$$\boxed{\dot{q}_m = 2L\gamma d_m^{(y)} \left(d_m^{(x)}\right)^{1/L} \frac{q_m^{2-1/L}}{(1+\gamma q_m)^2}.} \tag{180}$$

For $L = 1$, this is exactly the shallow noiseless population equation. For $d_m^{(y)} > 0$ and positive initial energy the mode grows monotonically. For $d_m^{(y)} = 0$, all of its gains remain fixed.

Exact trajectories and their growth scales. Set $a_m = \gamma q_m$ and $c_{L,m} = 2L\gamma^{1/L} \left(d_m^{(x)}\right)^{1/L} d_m^{(y)}$. Then

$$\dot{a}_m = c_{L,m} \frac{a_m^{2-1/L}}{(1+a_m)^2}. \tag{181}$$

Separating variables requires the antiderivative

$$G_L(a) = \begin{cases} \frac{1}{2}a^2 + 2a + \log a, & L = 1, \\ \dfrac{L}{L+1} a^{1+1/L} + 2La^{1/L} - \dfrac{L}{L-1} a^{-(L-1)/L}, & L > 1. \end{cases} \tag{182}$$

The logarithm at $L = 1$ comes from integrating a^{-1}; the power formula in the second line must not be used at that endpoint. Direct differentiation verifies $G_L'(a) = (1+a)^2/a^{2-1/L} > 0$. Moreover, $G_L(a) \to -\infty$ as $a \downarrow 0$, and $G_L(a) \to \infty$ as $a \to \infty$. It therefore has a unique inverse on the real line. Integrating the scalar ODE gives

$$G_L(a_m(t)) = G_L(a_m(0)) + c_{L,m}t, \qquad q_m(t) = \gamma^{-1} G_L^{-1}(G_L(\gamma q_m(0)) + c_{L,m}t)\,. \tag{183}$$

This expression assumes $q_m(0) > 0$. A zero mode remains zero under the balanced ODE, and should be handled directly rather than by substitution into G_L.

When $a_m \ll 1$, $L = 1$ gives the familiar early exponential growth. For $L > 1$, the leading small-mode equation instead gives

$$a_m(t)^{-(L-1)/L} \simeq a_m(0)^{-(L-1)/L} - \frac{L-1}{L} c_{L,m} t. \tag{184}$$

This approximation applies only while the mode is small; it does not predict a finite-time divergence of the full ODE. For a fixed small threshold a_*, the time to grow from $a_m(0) \downarrow 0$ to a_* scales as $\log(1/a_m(0))/c_{1,m}$ when $L = 1$, and as $La_m(0)^{-(L-1)/L}/[(L-1)c_{L,m}]$ when $L > 1$. Thus depth changes the duration of the small-initialization transient.

At large time, every active mode with $d_m^{(x)}, d_m^{(y)}, q_m(0) > 0$ diverges, and the leading term of G_L gives

$$a_m(t)^{1+1/L} \sim 2(L+1)\gamma^{1/L} \left(d_m^{(x)}\right)^{1/L} d_m^{(y)} t,$$
$$q_m(t) \sim \left[\frac{2(L+1)}{\gamma} \left(d_m^{(x)}\right)^{1/L} d_m^{(y)} t\right]^{L/(L+1)}. \tag{185}$$

There is no finite-time blowup: the exact inverse exists for every finite t, and the late-time growth is a power smaller than one. The terminal abstraction below is a limit of relative geometry while the individual active energies continue growing.

Layerwise interpolation at every learning time. For positive input and representation energies, balancing implies

$$q_m^{(\ell)} = \left(d_m^{(x)}\right)^{1-\ell/L} \left(q_m^{(L)}\right)^{\ell/L}, \qquad \nu_z^{(\ell)} = \nu_x^{1-\ell/L} \left(\nu_z^{(L)}\right)^{\ell/L}, \tag{186}$$

where $\nu_z^{(\ell)} = q_{SC}^{(\ell)}/q_S^{(\ell)}$ and $\nu_x = d_{SC}^{(x)}/d_S^{(x)}$. The first identity follows by eliminating g_m between $q_m^{(\ell)} = d_m^{(x)} g_m^{2\ell}$ and $q_m^{(L)} = d_m^{(x)} g_m^{2L}$; the second follows by taking the ratio of the two relevant modes.

For any positive ratio ν, its abstraction score satisfies

$$\operatorname{arctanh}\left(\frac{1-\nu}{1+\nu}\right) = \frac{1}{2}\log\frac{1+(1-\nu)/(1+\nu)}{1-(1-\nu)/(1+\nu)} = -\frac{1}{2}\log\nu. \tag{187}$$

Applying this identity to (186) therefore gives

$$\operatorname{arctanh}\alpha^{(\ell)}(t) = \left(1-\frac{\ell}{L}\right)\operatorname{arctanh}\alpha_x + \frac{\ell}{L}\operatorname{arctanh}\alpha^{(L)}(t). \tag{188}$$

This proves (16). In particular, at a fixed time abstraction increases from one hidden layer to the next exactly when $\alpha^{(L)}(t) > \alpha_x$; it decreases when the reverse inequality holds. Balancing alone does not assert that deeper layers are more abstract at every time, irrespective of the final-layer geometry.

Terminal abstraction at each depth. First assume $d_S^{(x)}, d_{SC}^{(x)}, d_S^{(y)}, d_{SC}^{(y)} > 0$, with positive initial gains for both modes. Taking the ratio of the two late-time asymptotics gives

$$\nu_{z,\infty}^{(L)} = \left(\frac{\left(d_{SC}^{(x)}\right)^{1/L} d_{SC}^{(y)}}{\left(d_S^{(x)}\right)^{1/L} d_S^{(y)}}\right)^{L/(L+1)} = \nu_x^{1/(L+1)}\nu_y^{L/(L+1)}, \qquad \nu_y = \frac{d_{SC}^{(y)}}{d_S^{(y)}}. \tag{189}$$

Neither the initial gains nor γ appear in this ratio. The corresponding score is $(1-\nu_{z,\infty}^{(L)})/(1+\nu_{z,\infty}^{(L)})$. Taking $-\frac{1}{2}\log$ gives the second statement of Proposition 5:

$$\operatorname{arctanh}\alpha_\infty^{(L)} = \frac{1}{L+1}\operatorname{arctanh}\alpha_x + \frac{L}{L+1}\operatorname{arctanh}\alpha_y. \tag{190}$$

Combining this with the finite-time layerwise identity yields

$$\nu_{z,\infty}^{(\ell;L)} = \nu_x^{1-\ell/(L+1)}\nu_y^{\ell/(L+1)},$$
$$\operatorname{arctanh}\alpha_\infty^{(\ell;L)} = \left(1-\frac{\ell}{L+1}\right)\operatorname{arctanh}\alpha_x + \frac{\ell}{L+1}\operatorname{arctanh}\alpha_y. \tag{191}$$

Here the notation $(\ell; L)$ distinguishes the position of a layer from the total number of hidden layers. The limiting hidden-layer scores are equally spaced in arctanh-coordinates between input and target geometry. The final hidden layer is at fraction $L/(L+1)$ of this interval, and approaches the target score as the total depth increases.

Corollary B.4 (Depth monotonicity of terminal abstraction)**.** *For $0 < \nu_x, \nu_y < \infty$, with the positive initial gains above, the final hidden layer's terminal abstraction strictly increases with network depth when $\nu_y < \nu_x$, strictly decreases when $\nu_y > \nu_x$, and is independent of depth when the ratios are equal.*

To see this without treating network depth as continuous, put $A_x = \operatorname{arctanh} \alpha_x$, $A_y = \operatorname{arctanh} \alpha_y$, and subtract $(A_x + LA_y)/(L+1)$ from the same expression with depth $L+1$:

$$\operatorname{arctanh} \alpha_\infty^{(L+1)} - \operatorname{arctanh} \alpha_\infty^{(L)} = \frac{A_y - A_x}{(L+1)(L+2)} = \frac{\log(\nu_x/\nu_y)}{2(L+1)(L+2)}. \tag{192}$$

The function $\tanh$ is increasing, so the same sign governs the change in α. If the closed expression is extended to real $L > 0$, its derivative is

$$\frac{\partial}{\partial L} \operatorname{arctanh} \alpha_\infty^{(L)} = \frac{A_y - A_x}{(L+1)^2} = \frac{1}{2(L+1)^2} \log \frac{\nu_x}{\nu_y}. \tag{193}$$

Similarly, for a network of fixed depth L, successive hidden layers have terminal $\operatorname{arctanh}$-scores differing by $(A_y - A_x)/(L+1)$. Thus both notions of improvement with depth hold when the targets are more abstract than the inputs.

C Nonlinear networks: Fourier geometry and learning dynamics

This part derives the results of §5. Throughout, $\sigma_x = 0$: an input and its deterministic target are functions of the four equally likely classes $u = (s, c) \in \{-1, +1\}^2$. We use $\mathcal{M} = \{\emptyset, S, C, SC\}$, with $\chi_m(u) = 1, s, c, sc$, and retain the squared-norm convention $d_m^{(x)} = \|\hat{\boldsymbol{x}}_m\|^2$, $d_m^{(y)} = \|\hat{\boldsymbol{y}}_m\|^2$. We first derive the population equations and identify the additional Gaussian assumption needed to close them. We then analyze the erf and leaky-ReLU dynamics, and prove the instantaneous attenuation law.

C.1 Fourier coordinates and nonlinear feature geometry

Network and normalization. For hidden width D, write

$$\boldsymbol{z}(u) = \boldsymbol{W}\boldsymbol{x}(u), \qquad \boldsymbol{h}(u) = \phi(\boldsymbol{z}(u)), \qquad \boldsymbol{f}(u) = D^{-1/2}\boldsymbol{W}_r\boldsymbol{h}(u), \tag{194}$$

where $\boldsymbol{W}_r \in \mathbb{R}^{D_y \times D}$. The population objective is

$$\mathcal{L}(\boldsymbol{W}_r, \boldsymbol{W}) = \frac{1}{2}\mathbb{E}_u\|D^{-1/2}\boldsymbol{W}_r\boldsymbol{h}(u) - \boldsymbol{y}(u)\|^2 + \frac{1}{2\gamma}\|\boldsymbol{W}_r\|_F^2, \qquad \mathbb{E}_u = \frac{1}{4}\sum_u. \tag{195}$$

The readout features $D^{-1/2}\boldsymbol{h}$ have order-one inner products. We measure learning time using $\dot{\boldsymbol{W}} = -D\nabla_{\boldsymbol{W}}\mathcal{L}^*$, where $\mathcal{L}^* = \min_{\boldsymbol{W}_r}\mathcal{L}$. This absorbs the width factor in the gradient. With unit learning rate on $\boldsymbol{W}$, the same trajectory takes D times as long. Neither this time convention nor a common normalization of the representation changes its abstraction score.

A typical neuron's preactivation has the scalar Fourier expansion

$$z(u) = \sum_m a_m\chi_m(u), \qquad a_m = \boldsymbol{w}^\top\hat{\boldsymbol{x}}_m. \tag{196}$$

We write $\mathbb{E}_a$ for an average over neurons, distinct from the class average $\mathbb{E}_u$. In the infinite-width description,

$$q_m = \mathbb{E}_a[a_m^2] = \lim_{D\to\infty} D^{-1}\|\hat{\boldsymbol{z}}_m\|^2. \tag{197}$$

A centered isotropic Gaussian weight initialization with covariance $g_0^2\boldsymbol{I}$ gives independent $a_m(0) \sim \mathcal{N}(0, g_0^2 d_m^{(x)})$. More general independent Gaussian initial mode variances are possible, subject to $q_m(0) = 0$ whenever $d_m^{(x)} = 0$.

Relabeling symmetry and Fourier orthogonality. For $g = (g_s, g_c) \in \{-1, +1\}^2$, let $ug = (sg_s, cg_c)$. Flipping latent labels transforms the coefficients as $a_m \mapsto \chi_m(g)a_m$. Suppose the neuron law is invariant under these transformations. For distinct modes m, n, there is a flip with $\chi_m(g)\chi_n(g) = -1$, so $\mathbb{E}_a[a_m a_n] = 0$. Consequently

$$Q(u, u') := \mathbb{E}_a[z(u)z(u')] = r_{uu'}, \qquad r_g = \sum_m q_m\chi_m(g), \qquad v = r_{(+1,+1)} = \sum_m q_m. \tag{198}$$

Here uu' denotes componentwise multiplication. The quantity v is the common preactivation second moment across classes; under the centered Gaussian assumption below, it is also their common variance.

For a single neuron's activation $h(u) = \phi(z(u))$, define its Fourier coefficient by

$$b_j(a) = \mathbb{E}_u[h(u)\chi_j(u)]. \tag{199}$$

Under the same relabeling, $b_j \mapsto \chi_j(g)b_j$. Hence $\mathbb{E}_a[b_j b_m] = 0$ for $j \neq m$: distinct feature Fourier modes have vanishing width-normalized inner products in the infinite-width limit. Their energies are

$$k_j = \mathbb{E}_a[b_j^2] = \lim_{D\to\infty} D^{-1}\left\|\hat{\boldsymbol{h}}_j\right\|^2, \qquad \hat{\boldsymbol{h}}_j = \mathbb{E}_u[\boldsymbol{h}(u)\chi_j(u)]. \tag{200}$$

In particular, a nonlinearity changes the Fourier energies but preserves this population orthogonality. The preactivation and feature abstraction scores are

$$\alpha_{\text{pre}} = \frac{q_S - q_{SC}}{q_S + q_{SC}}, \qquad \alpha_{\text{post}} = \frac{k_S - k_{SC}}{k_S + k_{SC}}, \tag{201}$$

provided the respective concept directions are nonzero. Unlike the linear analysis, this argument does not require fixed hidden directions for the feature Fourier vectors.

Gaussian closure assumption. We evaluate nonlinear moments using

$$a_m \text{ independent}, \qquad a_m \sim \mathcal{N}(0, q_m). \tag{202}$$

Equivalently, the four preactivations of a neuron are jointly centered Gaussian with covariance Q. This is exact for a fixed Gaussian neuron law. During training it is an additional moment-closure assumption: infinite width removes sampling fluctuations, but does not guarantee that a nonlinear flow preserves a Gaussian initialization. The covariance maps below are exact Gaussian integrals; the training ODEs are exact within the stated Gaussian closure.

For a centered Gaussian pair (U, V) with common variance v and covariance r, define

$$\psi_\phi(r, v) = \mathbb{E}[\phi(U)\phi(V)]. \tag{203}$$

Changing variables from (u, u') to $(u, g = uu')$, and using character orthogonality, gives

$$k_j = \frac{1}{16} \sum_{u,u'} \chi_j(u)\chi_j(u')\mathbb{E}_a[h(u)h(u')] = \frac{1}{4} \sum_g \chi_j(g)\psi_\phi(r_g, v). \tag{204}$$

This is the four-term transform used in (18). Nonnegative q_m's ensure $|r_g| \le v$, and $k_j \ge 0$ follows directly from (200).

The erf covariance map. Let $\phi_\beta(z) = \mathrm{erf}(\beta z)$, $\beta > 0$. If $\xi \sim \mathcal{N}(0, 1)$, then $\mathrm{erf}(\beta z) = \mathbb{E}_\xi[\mathrm{sign}(z + \xi/(\sqrt{2}\beta))]$. Using two independent auxiliary noises, the covariance of the two erf outputs is therefore the correlation of the signs of a Gaussian pair whose correlation is $\rho_\beta = 2\beta^2 r/(1+2\beta^2 v)$. For a standard Gaussian pair of correlation ρ, the angle between the two defining half-planes gives $\Pr(U > 0, V > 0) = 1/4 + \arcsin(\rho)/(2\pi)$. By symmetry, the sign correlation is four times this probability minus one. Thus [45]

$$\psi_\beta(r, v) = \frac{2}{\pi} \arcsin\left(\frac{2\beta^2 r}{1 + 2\beta^2 v}\right). \tag{205}$$

The leaky-ReLU covariance map. Write $z_+ = \max\{z, 0\}$ and $\phi_\omega(z) = \omega z + (1-\omega)z_+$, with $0 \le \omega \le 1$. If $\vartheta = \arccos(r/v)$, represent the Gaussian pair as $U = \sqrt{v}R\cos\varphi$, $V = \sqrt{v}R\cos(\varphi - \vartheta)$, where φ is uniform on $[0, 2\pi)$, independent of R, and $\mathbb{E}[R^2] = 2$. Both cosines are positive on an interval of length $\pi - \vartheta$. Integrating over that interval yields

$$\mathbb{E}[U_+V_+] = \frac{v}{\pi} \int_{\vartheta - \pi/2}^{\pi/2} \cos\varphi \cos(\varphi - \vartheta)\, d\varphi = \frac{v}{2\pi}[\sin\vartheta + (\pi - \vartheta)\cos\vartheta]. \tag{206}$$

Also, $\mathbb{E}[UV_+] = \mathbb{E}[U_+V] = r/2$, by the symmetry $(U, V) \mapsto (-U, -V)$. Expanding the two leaky-ReLU factors gives

$$\begin{aligned} \psi_\omega(r, v) &= \omega r + (1-\omega)^2 \mathbb{E}[U_+V_+] \\ &= \frac{1+\omega^2}{2} r + \frac{(1-\omega)^2}{2\pi} \left[\sqrt{v^2 - r^2} - r\arccos(r/v)\right]. \end{aligned} \tag{207}$$

This recovers the arc-cosine covariance formula [44] and its linear endpoint $\omega = 1$. If $v = 0$, all preactivation energies vanish and abstraction is undefined; below we assume $v > 0$.

C.2 Population learning dynamics under Gaussian closure

Eliminate the readout in feature Fourier coordinates. Character orthogonality decomposes the loss into

$$\mathcal{L} = \frac{1}{2} \sum_j \|D^{-1/2}\boldsymbol{W}_r\hat{\boldsymbol{h}}_j - \hat{\boldsymbol{y}}_j\|^2 + \frac{1}{2\gamma}\|\boldsymbol{W}_r\|_F^2. \tag{208}$$

Because the feature vectors $\hat{\boldsymbol{h}}_j$ are mutually orthogonal, the optimal readout is

$$\boldsymbol{W}_r^* = D^{-1/2} \sum_j \frac{\gamma}{1 + \gamma k_j} \hat{\boldsymbol{y}}_j \hat{\boldsymbol{h}}_j^\top, \qquad D^{-1/2}\boldsymbol{W}_r^*\hat{\boldsymbol{h}}_j = \frac{\gamma k_j}{1 + \gamma k_j}\hat{\boldsymbol{y}}_j. \tag{209}$$

A zero feature mode causes no difficulty: its target component remains unpredicted. Substituting the readout into the full objective, including its penalty, gives

$$\mathcal{L}^*_\phi(\boldsymbol{q}) = \frac{1}{2}\sum_j \frac{d_j^{(y)}}{1+\gamma k_j(\boldsymbol{q})}, \qquad \mu_j \coloneqq \frac{\gamma d_j^{(y)}}{(1+\gamma k_j)^2}, \qquad \frac{\partial \mathcal{L}^*}{\partial k_j} = -\frac{1}{2}\mu_j. \tag{210}$$

The quantities μ_j are the strengths with which the different target components enter the hidden update.

First derive the exact neuron equation. At the optimal readout, the residual on class u is

$$\boldsymbol{e}(u) = \boldsymbol{y}(u) - D^{-1/2}\boldsymbol{W}_r^*\boldsymbol{h}(u) = \sum_j \frac{\hat{\boldsymbol{y}}_j}{1+\gamma k_j}\chi_j(u). \tag{211}$$

The envelope theorem permits differentiating the loss while holding the optimal readout fixed. For a neuron with coefficients a_m, target orthogonality then gives

$$\begin{aligned} \dot{\boldsymbol{w}} &= \mathbb{E}_u\left[\phi'(z(u))\boldsymbol{x}(u)\sum_j \mu_j b_j(a)\chi_j(u)\right], \\ \dot{a}_m &= d_m^{(x)}\sum_j \mu_j b_j(a)\,\partial_{a_m} b_j(a), \qquad \partial_{a_m} b_j(a) = \mathbb{E}_u[\chi_j(u)\chi_m(u)\phi'(z(u))]. \end{aligned} \tag{212}$$

The factor $d_m^{(x)}$ follows from $\hat{\boldsymbol{x}}_m^\top \boldsymbol{x}(u) = d_m^{(x)}\chi_m(u)$. This equation is exact for a relabeling-invariant limiting neuron law, with $k_j = \mathbb{E}_a[b_j^2]$ evaluated under that law. Its velocity transforms as $\dot{a}_m \mapsto \chi_m(g)\dot{a}_m$ under a latent relabeling. The initial relabeling symmetry is consequently preserved, although Gaussianity need not be.

The exact evolution of a mode energy is

$$\dot{q}_m = 2d_m^{(x)}\sum_j \mu_j \mathbb{E}_a[a_m b_j(a)\partial_{a_m} b_j(a)]. \tag{213}$$

The right-hand side generally depends on the full neuron distribution. This is the point at which we apply the Gaussian closure.

Gaussian integration closes the second moments. For a centered Gaussian coordinate of variance q_m, integration by parts gives $\mathbb{E}_a[a_m f(a)] = q_m \mathbb{E}_a[\partial_{a_m} f(a)]$. A second useful identity is

$$\frac{\partial}{\partial q_m}\mathbb{E}_a[f(a)] = \frac{1}{2}\mathbb{E}_a[\partial_{a_m}^2 f(a)]. \tag{214}$$

It follows by differentiating the Gaussian density, using $\partial_{q_m} p = \frac{1}{2}\partial_{a_m}^2 p$, and integrating by parts twice. Apply these identities to $f = b_j\partial_{a_m} b_j$ and to b_j^2, respectively:

$$\begin{aligned} \mathbb{E}_a[a_m b_j \partial_{a_m} b_j] &= q_m \mathbb{E}_a[(\partial_{a_m} b_j)^2 + b_j\partial_{a_m}^2 b_j] \\ &= q_m \frac{\partial k_j}{\partial q_m} = q_m J_{jm}, \qquad J_{jm} \coloneqq \frac{\partial k_j}{\partial q_m}. \end{aligned} \tag{215}$$

For ReLU, the identities can be obtained by smooth approximation; the limiting second derivative includes the delta function at its kink. Discarding that term would give a different, incorrect Jacobian. Combining (213) and (215) proves

$$\begin{aligned} \kappa_m^\phi &= d_m^{(x)}\sum_j \mu_j J_{jm} = \gamma d_m^{(x)}\sum_j \frac{d_j^{(y)} J_{jm}}{(1+\gamma k_j)^2}, \\ \dot{q}_m &= 2q_m\kappa_m^\phi, \qquad \dot{\alpha}_{\mathrm{pre}} = (1-\alpha_{\mathrm{pre}}^2)(\kappa_S^\phi - \kappa_{SC}^\phi). \end{aligned} \tag{216}$$

The last identity follows by differentiating $(q_S - q_{SC})/(q_S + q_{SC})$, exactly as in the linear case. This proves (20) and (21) with the squared-norm convention.

An explicit four-term formula for the Jacobian. Each q_m changes both r_g and v, since $\partial r_g/\partial q_m = \chi_m(g)$ and $\partial v/\partial q_m = 1$. Differentiating (204) therefore gives

$$J_{jm} = \frac{1}{4}\sum_g \chi_j(g)\left[\psi_r(r_g, v)\chi_m(g) + \psi_v(r_g, v)\right]. \tag{217}$$

The derivative ψ_v holds the covariance r fixed while changing the common variance. It cannot be omitted, even for a nonconstant mode. For erf, put

$$G_\beta(r, v) = \frac{4\beta^2/\pi}{\sqrt{(1+2\beta^2 v)^2 - 4\beta^4 r^2}}, \qquad \zeta_\beta = \frac{2\beta^2}{1+2\beta^2 v}. \tag{218}$$

Direct differentiation of (205) yields

$$\psi_r(r, v) = G_\beta(r, v), \qquad \psi_v(r, v) = -\zeta_\beta r G_\beta(r, v). \tag{219}$$

For leaky ReLU, let $\rho = r/v$. The square-root terms cancel when (207) is differentiated, leaving

$$\psi_r(r, v) = \omega + \frac{(1-\omega)^2}{2\pi}(\pi - \arccos\rho), \qquad \psi_v(r, v) = \frac{(1-\omega)^2}{2\pi}\sqrt{1-\rho^2}. \tag{220}$$

These formulas extend continuously to $\rho = \pm 1$. They also have a direct Gaussian interpretation: $\psi_r = \mathbb{E}[\phi'(U)\phi'(V)]$ and $\psi_v = \mathbb{E}[\phi''(U)\phi(V)]$. For leaky ReLU, $\phi'' = (1-\omega)\delta_0$; evaluating this expectation gives the second formula in (220).

At the diagonal group element $g = (+1, +1)$, the covariance and variance both equal v. Its derivative in (217) is $\psi_r(v, v) + \psi_v(v, v)$, not $\psi_r(v, v)$ alone. In particular,

$$\frac{d}{dv}\psi_\beta(v, v) = \frac{4\beta^2}{\pi(1+2\beta^2 v)\sqrt{1+4\beta^2 v}}, \qquad \frac{d}{dv}\psi_\omega(v, v) = \frac{1+\omega^2}{2}. \tag{221}$$

Equations (204), (217), and (216) form a closed four-dimensional system.

Linear limit, boundary modes, and decreasing loss. For $\phi(z) = z$, we have $k_j = q_j$, $J_{jm} = \delta_{jm}$, and $\kappa_m = \gamma d_m^{(x)} d_m^{(y)}/(1+\gamma q_m)^2$. Thus the noiseless linear dynamics are recovered exactly, including the leaky-ReLU endpoint $\omega = 1$. Within the closure, $q_m(0) = 0$ implies $q_m(t) = 0$; an absent input mode also cannot learn. A zero target energy $d_m^{(y)} = 0$, however, does not force q_m to stay fixed, because other target components can act through J_{jm}. The gains need not all be positive, so the difference of gains in (216) is preferable to a ratio when describing general trajectories.

The chain rule gives another check on the dynamics:

$$\dot{q}_m = -4 d_m^{(x)} q_m \frac{\partial \mathcal{L}_\phi^*}{\partial q_m}, \qquad \frac{d\mathcal{L}_\phi^*}{dt} = -4\sum_m d_m^{(x)} q_m \left(\frac{\partial \mathcal{L}_\phi^*}{\partial q_m}\right)^2 \le 0. \tag{222}$$

The factor $4 d_m^{(x)} q_m$ is inherited from the weight dynamics. In particular, setting $\dot{q}_m = -\partial \mathcal{L}^*/\partial q_m$ would not describe the original parameterization.

C.3 Erf in the weakly nonlinear regime

We show that, while preactivations remain in the nearly linear part of $\phi_\beta(z) = \mathrm{erf}(\beta z)$ and the relevant modes remain nondegenerate, the Gaussian-closure dynamics reduce at leading order to those of a linear network with effective ridge parameter $\gamma_{\mathrm{eff}} = 4\beta^2\gamma/\pi$. At a ratio-balanced state with large relevant feature energies relative to the ridge scale, this recovers the linear prediction for abstraction up to controlled errors. These results give a mechanism for the numerical agreement in Fig. 5A, B, F, but do not establish an unrestricted infinite-time erf limit. The expansion parameter is $\beta^2 v$, the variance of the argument of erf, rather than β alone.

Kernel and Jacobian expansions. Define

$$c_\beta = \frac{4\beta^2}{\pi}, \qquad \varepsilon = 2\beta^2 v, \qquad b_\beta = \frac{c_\beta}{1+\varepsilon}, \qquad \zeta_\beta = \frac{2\beta^2}{1+\varepsilon}. \tag{223}$$

Fix $\varepsilon_0 < 1$ and suppose $0 \le \varepsilon \le \varepsilon_0$. Since $|\zeta_\beta r| \le \varepsilon/(1+\varepsilon)$, Taylor expansion of $\arcsin$ gives, uniformly for $|r| \le v$,

$$\psi_\beta(r,v) = b_\beta r + O(b_\beta |r| \varepsilon^2), \qquad G_\beta(r,v) = b_\beta[1 + O(\varepsilon^2)]. \tag{224}$$

Substitution into the Fourier map and its Jacobian yields

$$\begin{aligned} k_j &= b_\beta q_j + O(b_\beta v \varepsilon^2), \\ J_{jm} &= b_\beta[\delta_{jm} - \zeta_\beta q_j] + O(b_\beta \varepsilon^2). \end{aligned} \tag{225}$$

For example, the term $-b_\beta \zeta_\beta q_j$ follows from $\frac{1}{4}\sum_g \chi_j(g) r_g = q_j$. It is the derivative of the common variance-dependent prefactor, and it contributes at order ε, earlier than the cubic term of $\arcsin$. Since $b_\beta = c_\beta[1 + O(\varepsilon)]$, these formulas imply

$$\psi_\beta(r,v) = c_\beta r + O(\beta^4 v^2), \qquad J_{jm} = c_\beta \delta_{jm} + O(\beta^4 v), \tag{226}$$

which exhibit the leading linear covariance map and diagonal Jacobian.

C.4 Leaky-ReLU dynamics and terminal abstraction

We now impose the centered, signal-balanced conditions of Setting 3:

$$d_S^{(x)} = d_C^{(x)} > 0, \quad d_S^{(y)} = d_C^{(y)} > 0, \quad q_S(0) = q_C(0) > 0, \qquad d_\emptyset^{(y)} = q_\emptyset(0) = 0. \tag{227}$$

These are preserved by the closed equations. Centering the preactivations does not center their nonlinear features: generally $k_\emptyset > 0$. Its target strength is zero, however, so $\mu_\emptyset = 0$. For interior trajectories assume also $d_{SC}^{(x)} > 0$ and $q_{SC}(0) > 0$, and set

$$q = q_S = q_C, \qquad \nu = \frac{q_{SC}}{q}, \qquad \nu_x = \frac{d_{SC}^{(x)}}{d_S^{(x)}}, \qquad \nu_y = \frac{d_{SC}^{(y)}}{d_S^{(y)}}. \tag{228}$$

The scale q and ratio ν determine all preactivation energies.

Correlations and feature energies. The covariance of two class preactivations is $q(ss' + cc' + \nu ss' cc')$, with common variance $v = q(2+\nu)$. The correlations for classes differing in one or both latents are

$$\rho_1 = -\frac{\nu}{2+\nu}, \qquad \rho_0 = \frac{\nu - 2}{\nu + 2}. \tag{229}$$

Introduce the angle

$$\theta = \arccos \rho_0, \qquad \nu = \frac{2(1+\cos\theta)}{1-\cos\theta}, \qquad \rho_1 = -\frac{1+\cos\theta}{2}, \qquad \alpha_{\mathrm{pre}}(\theta) = -\frac{1+3\cos\theta}{3+\cos\theta}. \tag{230}$$

Thus $\theta = \pi$ corresponds to $\nu = 0$ and $\alpha_{\mathrm{pre}} = 1$; the limit $\theta \downarrow 0$ gives $\nu \to \infty$ and $\alpha_{\mathrm{pre}} \to -1$.

Let $a_\omega = (1+\omega^2)/2$ and normalize the covariance map as

$$g_\omega(\rho) = \rho + \frac{(1-\omega)^2}{\pi(1+\omega^2)}[\sqrt{1-\rho^2} - \rho \arccos \rho], \qquad \psi_\omega(v\rho, v) = a_\omega v g_\omega(\rho). \tag{231}$$

The four-term transform gives

$$\begin{aligned} k_S = k_C &= \frac{a_\omega v}{4}[1 - g_\omega(\rho_0)], \\ k_{SC} &= \frac{a_\omega v}{4}[1 - 2g_\omega(\rho_1) + g_\omega(\rho_0)], \\ k_\emptyset &= \frac{a_\omega v}{4}[1 + 2g_\omega(\rho_1) + g_\omega(\rho_0)]. \end{aligned} \tag{232}$$

Their sum is $a_\omega v$, as required. Positive homogeneity implies $k_j = q f_j(\nu)$ for functions f_j independent of scale. In particular,

$$\nu_K(\theta, \omega) = \frac{k_{SC}}{k_S} = \frac{1 - 2g_\omega(\rho_1) + g_\omega(\rho_0)}{1 - g_\omega(\rho_0)}. \tag{233}$$

Target strengths at finite ridge. Define

$$\mu_S = \mu_C = \frac{\gamma d_S^{(y)}}{(1+\gamma k_S)^2}, \qquad \mu := \frac{\mu_{SC}}{\mu_S} = \nu_y \left(\frac{1+\gamma k_S}{1+\gamma k_{SC}}\right)^2. \tag{234}$$

This is the exact ratio at the current scale. If both relevant feature energies grow without bound, it tends to $\mu_\infty = \nu_y/\nu_K^2$. When $\nu_y = 0$, it is exactly zero at every scale and for every finite $\gamma > 0$.

Evaluate the two modal gains explicitly. For pure ReLU, define

$$D(\rho) = \frac{\sqrt{1-\rho^2}}{2\pi}, \qquad E(\rho) = \frac{\pi - \arccos\rho}{2\pi}, \qquad D_i = D(\rho_i), \quad E_i = E(\rho_i), \quad i = 0, 1. \tag{235}$$

These are the variance and covariance derivatives of its Gaussian map. The gain $\mathcal{G}_m = \sum_j \mu_j J_{jm}$ can be evaluated using (217) without any matrix diagonalization. At the four relative class labels $g = (++), (+-), (-+), (--)$, the sums $\sum_j \mu_j \chi_j(g)$ are respectively $\mu_S(2+\mu), -\mu_S\mu, -\mu_S\mu, \mu_S(\mu-2)$. In the shape transform the last entry has sign minus, while in the interaction transform it has sign plus; the two one-flip entries cancel in the first and add in the second. The variance-derivative term has no factor χ_m. Collecting these terms gives

$$\begin{aligned} A_S(\theta,\mu) &= \frac{2+\mu}{8} - \frac{\mu}{2}D_1 + \frac{\mu-2}{4}D_0 - \frac{\mu-2}{4}E_0, \\ A_{SC}(\theta,\mu) &= \frac{2+\mu}{8} - \frac{\mu}{2}D_1 + \frac{\mu-2}{4}D_0 + \frac{\mu}{2}E_1 + \frac{\mu-2}{4}E_0. \end{aligned} \tag{236}$$

Here $\mathcal{G}_S = \mu_S A_S$ and $\mathcal{G}_{SC} = \mu_S A_{SC}$ for pure ReLU. Because $\psi_\omega = \omega\psi_{\text{linear}} + (1-\omega)^2\psi_{\text{ReLU}}$, the Jacobians have the same linear decomposition. Thus for general leaky ReLU set

$$B_S = \omega + (1-\omega)^2 A_S(\theta,\mu), \qquad B_{SC} = \omega\mu + (1-\omega)^2 A_{SC}(\theta,\mu). \tag{237}$$

The exact finite-ridge equations within the closure reduce to

$$\begin{aligned} \dot{q} &= 2d_S^{(x)}\mu_S q B_S, \\ \dot{\nu} &= 2d_S^{(x)}\mu_S \nu(\nu_x B_{SC} - B_S), \\ \dot{\alpha}_{\text{pre}} &= d_S^{(x)}\mu_S(1-\alpha_{\text{pre}}^2)(B_S - \nu_x B_{SC}). \end{aligned} \tag{238}$$

All quantities on the right are explicit functions of (q, ν).

A positive scale-growth rate. For $\mu \geq 0$ and $0 < \theta \leq \pi$, the shape gain is bounded away from zero. To see this, write

$$A_S(\theta,\mu) = \frac{2\pi - \theta - \sin\theta}{4\pi} + \mu\frac{\theta + \sin\theta - 2\sqrt{1-\cos^4(\theta/2)}}{8\pi}. \tag{239}$$

The first term is at least $1/4$. For the numerator of the second, put $t = \theta/2 \in (0, \pi/2]$. It equals $2\sin t[t/\sin t + \cos t - \sqrt{1+\cos^2 t}]$, which is nonnegative because $t/\sin t \geq 1$ and $1 + \cos t \geq \sqrt{1+\cos^2 t}$. Hence

$$B_S \geq \omega + \frac{(1-\omega)^2}{4} = \frac{(1+\omega)^2}{4} > 0. \tag{240}$$

The scale q therefore increases along every interior trajectory.

Scale growth and convergence of the ratio. For $0 \leq \omega < 1$, every trajectory with positive initial shape and interaction energies has a finite limiting ratio. We establish this before identifying the equation for an interior limit. First, put $a = (1+\omega)/2$. The decomposition $\phi_\omega(z) = az + (1-\omega)|z|/2$, together with symmetry under $\boldsymbol{a} \mapsto -\boldsymbol{a}$, makes its odd and even parts orthogonal. Hence

$$a^2 \leq f_S(\nu) \leq a_\omega, \qquad f_{SC}(\nu) \geq a^2\nu. \tag{241}$$

The upper bound follows also from $g_\omega(\rho) \geq \rho$ in (232). Since $q \geq q(0) > 0$, these bounds give $\mu = O(\nu^{-2})$ as $\nu \to \infty$, uniformly along the trajectory. In this limit, $\theta = O(\nu^{-1/2})$, $A_{SC}(\theta, 0) =$

$(\theta - \sin\theta)/(4\pi) = O(\nu^{-3/2})$, and the coefficient of μ in A_{SC} is bounded. Thus $B_{SC} \to 0$, while $B_S \geq a^2$; the ratio drift is negative for all sufficiently large ν. This bounds $\nu(t)$ above by some finite M.

On $[0, M]$, both f_S and f_{SC} are continuous and positive: the only additional endpoint to check is

$$f_{SC}(0) = \frac{(1-\omega)^2}{4}\left(1 - \frac{2}{\pi}\right) > 0. \tag{242}$$

It follows that μ and B_S are bounded. The scale equation is therefore bounded above and below by positive multiples of $q/(1+q)^2$. Consequently $q \to \infty$, $q(t) = \Theta(t^{1/2})$, and both relevant feature energies diverge. Define the increasing growth clock

$$\tau(t) = \int_0^t 2d_S^{(x)} \mu_S B_S \, dt' = \log \frac{q(t)}{q(0)} \longrightarrow \infty. \tag{243}$$

In this clock the ratio equation is $d\nu/d\tau = \nu(\nu_x B_{SC}/B_S - 1)$. As $q \to \infty$, its right-hand side converges uniformly on $[0, M]$ to the expression obtained by replacing μ with $\mu_\infty = \nu_y/\nu_K^2$. The limiting field is analytic for $\nu > 0$ and is not identically zero, since its drift is negative at sufficiently large ν. A bounded scalar trajectory that had two distinct accumulation points would cross every interval between them in both directions arbitrarily late. At a point where the limiting field is nonzero, its eventual fixed sign makes those crossings impossible. All intermediate points would therefore have to be zeros of the analytic field, contradicting nonidentity. Thus $\nu(t)$ converges, either to zero or to an interior root. This argument does not assert uniqueness of that root.

The scalar equation for an interior terminal ratio. With $\mu_\infty(\theta) = \nu_y/\nu_K(\theta, \omega)^2$, define

$$\nu_R(\theta; \omega, \nu_y) = \frac{\omega\mu_\infty + (1-\omega)^2 A_{SC}(\theta, \mu_\infty)}{\omega + (1-\omega)^2 A_S(\theta, \mu_\infty)}. \tag{244}$$

If the limiting ratio did not satisfy

$$\nu_x \nu_R(\theta; \omega, \nu_y) = 1, \tag{245}$$

its velocity in the diverging clock would stay bounded away from zero near the limit, a contradiction. This proves the scalar condition in (25). For an isolated stable root, the limiting ratio flow $\nu[\nu_x\nu_R(\nu) - 1]$ points towards that root; a negative derivative there is sufficient for local stability. Boundary limits must be considered separately. Merely setting the ordinary-time velocity to zero would not justify the condition, since $\mu_S \to 0$.

When $\omega = 1$, $k_m = q_m$, $B_S = 1$, and $B_{SC} \to \nu_y/\nu^2$. The root condition becomes $\nu^2 = \nu_x\nu_y$, recovering the linear terminal law. For the parameters of Fig. 5E, $(\omega, \nu_x, \nu_y) = (0.5, 0.6, 0.1)$, the stable root is $\theta \simeq 2.514815$, giving $\alpha_{\mathrm{pre},\infty} \simeq 0.652837$.

Proposition C.1 (Perfect-abstraction boundary for ReLU)**.** *Under (227), let $\omega = 0$ and $r = 1 - 2/\pi$. In the limiting ratio dynamics, the perfect-abstraction state $\nu = 0$ is locally attracting for $\nu_x < \nu_x^{\mathrm{crit}}(\nu_y)$ and repelling for $\nu_x > \nu_x^{\mathrm{crit}}(\nu_y)$, where*

$$\nu_x^{\mathrm{crit}}(\nu_y) = \frac{r(2r + \nu_y)}{2r^2 + (1+r)\nu_y}. \tag{246}$$

Equivalently, the boundary is

$$\nu_y(\nu_x) = \frac{2r^2(1-\nu_x)}{(1+r)\nu_x - r}, \qquad \frac{r}{1+r} < \nu_x \leq 1. \tag{247}$$

In normalized coordinates, its endpoints are $\eta_x(1) = (\pi - 2)/(3\pi - 4)$ and $\eta_x(0) = 1/2$, and $d\eta_y/d\eta_x < -1$ everywhere along the curve. This is a local stability boundary; it does not assert convergence to perfect abstraction from every initialization.

Proof. Perfect abstraction corresponds to $\theta = \pi$, where $\rho_0 = -1$ and $\rho_1 = 0$. For pure ReLU, $g_0(-1) = 0$ and $g_0(0) = 1/\pi$, so (232) gives

$$k_S = \frac{q}{4}, \qquad k_{SC} = \frac{rq}{4}, \qquad \nu_K(\pi, 0) = r, \qquad \mu_\infty(\pi) = \frac{\nu_y}{r^2}. \tag{248}$$

In particular, the interaction feature energy remains positive even when the preactivation interaction ratio vanishes. Substituting $D_0 = E_0 = 0$, $D_1 = 1/(2\pi)$, and $E_1 = 1/4$ into (236), and writing $\mu_0 = \nu_y/r^2$, yields

$$B_S(\pi, \mu_0) = \frac{1}{4} + \frac{r\mu_0}{8}, \qquad B_{SC}(\pi, \mu_0) = \frac{1}{4} + \frac{(1+r)\mu_0}{8}. \tag{249}$$

In the diverging growth clock (243), the limiting ratio flow near zero is therefore

$$\frac{d\nu}{d\tau} = \nu\left[\nu_x \frac{2 + (1+r)\mu_0}{2 + r\mu_0} - 1\right] + o(\nu). \tag{250}$$

The coefficient changes sign at $\nu_x = (2 + r\mu_0)/(2 + (1+r)\mu_0)$, proving (246); rearranging gives (247).

Using $\eta_i = \nu_i/(1+\nu_i)$, the same curve becomes

$$\eta_x(\eta_y) = \frac{2r^2(1-\eta_y) + r\eta_y}{4r^2(1-\eta_y) + (1+2r)\eta_y}. \tag{251}$$

Evaluating at $\eta_y = 0, 1$ gives the stated endpoints, with the top endpoint understood as $\nu_y \to \infty$. Differentiation gives

$$\frac{d\eta_y}{d\eta_x} = -\frac{[4r^2 + (1+2r-4r^2)\eta_y]^2}{2r^2} \le -8r^2 < -1, \tag{252}$$

since $1 + 2r - 4r^2 > 0$ and $8r^2 > 1$. □

C.5 Magnitude attenuation for erf and leaky ReLU

We now fix a representation, independently of its training rule, and compare its abstraction before and after applying a nonlinearity. Assume its preactivations have the centered Gaussian law (202) with $q_m \ge 0$ and $q_S + q_{SC} > 0$. We first prove magnitude attenuation for erf and leaky ReLU, then give sign preservation as a separate corollary. §C.6 extends the magnitude bound to arbitrary square-integrable activations.

Write the angle as a ratio of covariance differences. Introduce

$$X = q_\emptyset + q_C, \quad U = q_\emptyset - q_C, \qquad Y = q_S + q_{SC}, \quad V = q_S - q_{SC}. \tag{253}$$

Nonnegative Fourier energies imply $X \ge |U|$, $Y \ge |V|$, and $v = X + Y$. The four covariance entries are

$$r_{++} = X + Y, \quad r_{-+} = X - Y, \qquad r_{+-} = U + V, \quad r_{--} = U - V. \tag{254}$$

The entry r_{+-} compares the same shape across color contexts; r_{-+} compares opposite shapes in the same context. Taking the shape and interaction Fourier coefficients in (204) gives

$$\alpha_{\mathrm{pre}} = \frac{V}{Y}, \qquad \alpha_{\mathrm{post}} = \frac{\psi_\phi(U+V, v) - \psi_\phi(U-V, v)}{\psi_\phi(X+Y, v) - \psi_\phi(X-Y, v)}. \tag{255}$$

The denominator is $2(k_S + k_{SC})$, so it is positive whenever the feature concept directions are nonzero.

Theorem C.2 (Magnitude attenuation for erf and leaky ReLU)**.** *For the centered Gaussian preactivation law (202) with nonnegative Fourier energies, $q_S + q_{SC} > 0$, and either $\phi_\beta(z) = \mathrm{erf}(\beta z)$, $\beta > 0$, or $\phi_\omega(z) = \max\{z, \omega z\}$, $0 \le \omega \le 1$, the feature concept directions are nonzero and*

$$|\alpha_{\mathrm{post}}| \le |\alpha_{\mathrm{pre}}|. \tag{256}$$

Proof. **Step 1: a bound for a single power.** Normalize the covariance map by $g(\rho) = \psi_\phi(v\rho, v)/\psi_\phi(v, v)$. The positive normalizing factor cancels from abstraction. In the following calculation divide X, U, Y, V by v, retaining their names, so $X + Y = 1$. For an integer $p \ge 1$, define

$$N_p = (U+V)^p - (U-V)^p, \qquad \Delta_p = (X+Y)^p - (X-Y)^p. \tag{257}$$

The even terms cancel in each binomial difference. Thus

$$|N_p| \le 2|V| \sum_{\substack{1\le j\le p\\ j \text{ odd}}} \binom{p}{j} |U|^{p-j}|V|^{j-1}$$
$$\le 2|V| \sum_{\substack{1\le j\le p\\ j \text{ odd}}} \binom{p}{j} X^{p-j}Y^{j-1} = \frac{|V|}{Y}\Delta_p. \tag{258}$$

Here $X, Y \ge 0$ also imply $\Delta_p \ge 0$. The bound includes $V = 0$, when $N_p = 0$; it does not require any individual covariance entry to be nonnegative.

Step 2: a nonnegative mixture of powers. Suppose the normalized covariance map has an absolutely convergent series

$$g(\rho) = \sum_{p\ge 0} a_p \rho^p, \qquad a_p \ge 0, \qquad \sum_p a_p < \infty. \tag{259}$$

The constant cancels from the two differences, and (258) gives

$$|g(U+V) - g(U-V)| \le \sum_{p\ge 1} a_p |N_p|$$
$$\le \frac{|V|}{Y} \sum_{p\ge 1} a_p \Delta_p = |\alpha_{\text{pre}}|[g(X+Y) - g(X-Y)]. \tag{260}$$

Whenever $g(X+Y) - g(X-Y) > 0$, division proves $|\alpha_{\text{post}}| \le |\alpha_{\text{pre}}|$. In particular, $a_1 > 0$ guarantees a strictly positive denominator: $\sum_{p\ge 1} a_p\Delta_p \ge a_1\Delta_1 = 2a_1 Y > 0$. When $V = 0$, the numerator vanishes directly and both scores are zero.

Step 3: verify the conditions for erf. Let $\xi_\beta = 2\beta^2 v/(1 + 2\beta^2 v) \in (0, 1)$. Equation (205) gives

$$g_\beta(\rho) = \frac{\arcsin(\xi_\beta \rho)}{\arcsin \xi_\beta}, \qquad \arcsin x = \sum_{j=0}^{\infty} \frac{(2j)!}{4^j (j!)^2 (2j+1)} x^{2j+1}. \tag{261}$$

All nonzero coefficients are positive, including the linear coefficient, and the series converges absolutely on $[-1, 1]$. Step 2 therefore proves magnitude attenuation and nonzero feature concept directions.

Step 4: verify the conditions for leaky ReLU. The normalized pure-ReLU map is

$$g_R(\rho) = \frac{\sqrt{1-\rho^2} + (\pi - \arccos \rho)\rho}{\pi}. \tag{262}$$

Since $g_R'(\rho) = 1/2 + \arcsin(\rho)/\pi$ and $g_R(0) = 1/\pi$, integrating the arcsine series gives

$$g_R(\rho) = \frac{1}{\pi} + \frac{\rho}{2} + \frac{1}{\pi} \sum_{j=0}^{\infty} \frac{(2j)!\rho^{2j+2}}{4^j (j!)^2 (2j+1)(2j+2)}. \tag{263}$$

All coefficients are nonnegative and the linear coefficient is positive. For leaky ReLU,

$$g_\omega(\rho) = \frac{2\omega}{1+\omega^2}\rho + \frac{(1-\omega)^2}{1+\omega^2} g_R(\rho). \tag{264}$$

Both mixture weights are nonnegative and sum to one. Thus g_ω has an absolutely convergent series with nonnegative coefficients and a positive linear coefficient. Step 2 proves the result, including the identity case $\omega = 1$, where $\alpha_{\text{post}} = \alpha_{\text{pre}}$. □

Corollary C.3 (No sign reversal for erf and leaky ReLU)**.** *Under the assumptions of Theorem C.2,* $\alpha_{\text{post}}\alpha_{\text{pre}} \ge 0$. *Together with magnitude attenuation, this gives* $0 \le \alpha_{\text{post}} \le \alpha_{\text{pre}}$ *when* $\alpha_{\text{pre}} \ge 0$, *and* $\alpha_{\text{pre}} \le \alpha_{\text{post}} \le 0$ *when* $\alpha_{\text{pre}} \le 0$.

Proof. The normalized covariance maps are nondecreasing on $[-1, 1]$:

$$g'_\beta(\rho) = \frac{\xi_\beta}{\arcsin(\xi_\beta)\sqrt{1-\xi_\beta^2\rho^2}} > 0,$$
$$g'_R(\rho) = \frac{1}{2} + \frac{\arcsin\rho}{\pi} \geq 0. \tag{265}$$

Equation (264) expresses g_ω as a nonnegative mixture of the identity and g_R, so it too is nondecreasing. With the original coordinates in (253), the difference $g((U+V)/v) - g((U-V)/v)$ has the same weak sign as V. The denominator in (255) is positive, and $\alpha_{\mathrm{pre}} = V/Y$ with $Y > 0$, proving the claim. □

C.6 Magnitude attenuation for square-integrable activations

Magnitude attenuation extends beyond erf and leaky ReLU to every scalar activation with a finite second moment under the preactivation law. We prove this by fixing the context offsets, applying Gebelein's inequality to the resulting concept differences, and then averaging over the offsets.

Gebelein's inequality and maximal correlation. The Hirschfeld–Gebelein–Rényi (HGR) maximal correlation of two random variables is the supremum of the Pearson correlations between separate measurable transformations of them, with positive finite variances. For standard jointly Gaussian G_1, G_2 with Pearson correlation ρ, Gebelein's inequality states that

$$|\mathbb{E}[f(G_1)g(G_2)]| \leq |\rho|\sqrt{\mathbb{E}[f(G_1)^2]\,\mathbb{E}[g(G_2)^2]} \tag{266}$$

for square-integrable f, g with $\mathbb{E}[f(G_1)] = \mathbb{E}[g(G_2)] = 0$ [42, 43]. Linear transformations attain this bound, so the HGR maximal correlation of a jointly Gaussian pair equals the absolute value of its Pearson correlation. We use the inequality itself below; at $|\rho| = 1$, it follows directly from Cauchy–Schwarz. The following theorem is the formal version of Theorem 6.

Theorem C.4 (General Gaussian magnitude attenuation)**.** *Let the four class preactivations have the centered Gaussian law (202), with $q_m \geq 0$, $q_S + q_{SC} > 0$, and common variance $v = \sum_m q_m$. Let $\phi : \mathbb{R} \to \mathbb{R}$ be measurable and satisfy $\mathbb{E}[\phi(\sqrt{v}Z)^2] < \infty$ for $Z \sim \mathcal{N}(0,1)$. Define the feature energies k_m by (204). Whenever $k_S + k_{SC} > 0$,*

$$|\alpha_{\mathrm{post}}| \leq |\alpha_{\mathrm{pre}}|. \tag{267}$$

Proof. Using the independent Gaussian Fourier coefficients $a_m \sim \mathcal{N}(0, q_m)$ from (202), write the preactivation in context $c \in \{-1, +1\}$ as

$$z_{s,c} = B_c + sT_c, \qquad B_c = a_\emptyset + ca_C, \qquad T_c = a_S + ca_{SC}. \tag{268}$$

Here B_c is the offset shared by the two shapes in context c. The pair (B_+, B_-) is independent of (T_+, T_-), although the two offsets may be correlated with each other. Moreover,

$$\mathrm{Var}(T_\pm) = q_S + q_{SC} = Y > 0,$$
$$\rho := \mathrm{Corr}(T_+, T_-) = \frac{q_S - q_{SC}}{q_S + q_{SC}} = \alpha_{\mathrm{pre}}, \tag{269}$$

since the preactivation concept differences are $2T_+$ and $2T_-$.

For a fixed offset b, define

$$F_b(t) = \phi(b+t) - \phi(b-t), \qquad D_c = F_{B_c}(T_c). \tag{270}$$

The D_c's are the per-neuron feature differences. Averaging across neurons gives the concept-vector norms and inner product, so

$$\mathbb{E}[D_\pm^2] = 4(k_S + k_{SC}) > 0, \qquad \alpha_{\mathrm{post}} = \frac{\mathbb{E}[D_+ D_-]}{\sqrt{\mathbb{E}[D_+^2]\,\mathbb{E}[D_-^2]}}. \tag{271}$$

Since $B_c \pm T_c \sim \mathcal{N}(0, v)$, $\mathbb{E}[D_c^2] \leq 4\mathbb{E}[\phi(\sqrt{v}Z)^2] < \infty$.

The function F_b is odd in t, regardless of whether ϕ is odd. Hence, for $T \sim \mathcal{N}(0, Y)$, $\mathbb{E}_T[F_b(T)] = 0$ for almost every b under the distribution of B_c. Write $\sigma(b)^2 = \mathbb{E}_T[F_b(T)^2]$; this is finite for almost every offset by the preceding second-moment bound. Conditional on (B_+, B_-), the pair $(T_+/\sqrt{Y}, T_-/\sqrt{Y})$ retains its standard jointly Gaussian law with correlation ρ. Applying (266) to the two mean-zero functions $F_{B_+}(\sqrt{Y}\,\cdot)$ and $F_{B_-}(\sqrt{Y}\,\cdot)$ gives

$$|\mathbb{E}[D_+ D_- \mid B_+, B_-]| \le |\rho|\, \sigma(B_+)\sigma(B_-). \tag{272}$$

Averaging over the offsets and applying Cauchy–Schwarz yields

$$\begin{aligned} |\mathbb{E}[D_+ D_-]| &\le |\rho|\, \mathbb{E}[\sigma(B_+)\sigma(B_-)] \\ &\le |\rho| \sqrt{\mathbb{E}[\sigma(B_+)^2]\, \mathbb{E}[\sigma(B_-)^2]} \\ &= |\rho| \sqrt{\mathbb{E}[D_+^2]\, \mathbb{E}[D_-^2]}. \end{aligned} \tag{273}$$

Division by the positive denominator in (271) proves $|\alpha_{\mathrm{post}}| \le |\rho| = |\alpha_{\mathrm{pre}}|$. □

C.7 Normalization and prescribed offsets in the simulations

The preceding derivation uses a uniform population over four noiseless classes. For comparison with simulations on N fixed examples, let $\boldsymbol{X}_{\mathrm{data}}$ and $\boldsymbol{Y}$ have the input and target columns, and write $\boldsymbol{X} = \boldsymbol{X}_{\mathrm{data}}^\top \boldsymbol{X}_{\mathrm{data}}$, $\boldsymbol{T} = \boldsymbol{Y}^\top \boldsymbol{Y}$. Here these finite Gram matrices are used only to specify the normalization of the simulator. With the same width normalization and an empirical average $1/N$, readout elimination gives

$$\mathcal{L}^*(\boldsymbol{K}) = \frac{1}{2N} \operatorname{Tr}[\boldsymbol{T}(\boldsymbol{I} + \gamma \boldsymbol{K}/N)^{-1}], \qquad \boldsymbol{M} = \frac{\gamma}{N} (\boldsymbol{I} + \gamma \boldsymbol{K}/N)^{-1} \boldsymbol{T} (\boldsymbol{I} + \gamma \boldsymbol{K}/N)^{-1}. \tag{274}$$

For a neuron row, write the preactivations $\boldsymbol{z}$ and activations $\boldsymbol{h} = \phi(\boldsymbol{z})$ as column vectors across examples. The exact row equation is

$$\dot{\boldsymbol{z}} = \frac{\boldsymbol{X}}{N} [\phi'(\boldsymbol{z}) \odot (\boldsymbol{M}\boldsymbol{h})]. \tag{275}$$

If its Gaussian covariance has common diagonal v, let $K_{ij} = \psi_\phi(Q_{ij}, v)$, $G_{ij} = \psi_r(Q_{ij}, v)$, and $H_{ij} = \psi_v(Q_{ij}, v)$. The same Gaussian integration as above gives

$$\begin{aligned} \boldsymbol{R} &= \boldsymbol{M} \odot \boldsymbol{G} + \operatorname{Diag}[(\boldsymbol{M} \odot \boldsymbol{H})\mathbf{1}], \\ \dot{\boldsymbol{Q}} &= \frac{1}{N} (\boldsymbol{X}\boldsymbol{R}\boldsymbol{Q} + \boldsymbol{Q}\boldsymbol{R}\boldsymbol{X}). \end{aligned} \tag{276}$$

For noiseless repeated classes, the ordinary matrix eigenvalues along the class Fourier sign vectors are $Nq_m, Nk_m, Nd_m^{(x)}, Nd_m^{(y)}$. Consequently an implementation using $(\boldsymbol{I} + \gamma_{\mathrm{sim}} \boldsymbol{K})^{-1}$ and $\boldsymbol{X}$ without the $1/N$ factor corresponds to

$$\gamma = N \gamma_{\mathrm{sim}}, \qquad t = N t_{\mathrm{sim}}. \tag{277}$$

The entanglement ratios and abstraction scores do not change under this normalization conversion.

The row simulator in Fig. 5F also permits prescribed initial preactivation components in $\ker \boldsymbol{X}$. Equation (275) leaves these components fixed. They can be represented as a fixed offset $\boldsymbol{Z}_0$ in $\boldsymbol{Z}(t) = \boldsymbol{W}(t)\boldsymbol{X}_{\mathrm{data}} + \boldsymbol{Z}_0$, with $\boldsymbol{Z}_0$ supported on that nullspace. This is a noiseless prescribed-offset model; it does not add observational input noise. Such initial conditions need not be realizable by $\boldsymbol{W}\boldsymbol{X}_{\mathrm{data}}$ alone, which necessarily satisfies $\ker \boldsymbol{X} \subseteq \ker \boldsymbol{Q}$. The four-class calculation is recovered when these offsets vanish.

For n fixed examples per class, denote the individual diagonal and same-class off-diagonal preactivation covariances by q_d, q_2. The class-centroid diagonal is $\bar{q} = [q_d + (n-1)q_2]/n$. After the nonlinearity, the feature-centroid diagonal is instead

$$\bar{k} = \frac{\psi_\phi(q_d, q_d) + (n-1)\psi_\phi(q_2, q_d)}{n}. \tag{278}$$

D Generalization to multiple latent variables

We extend the two-layer linear theory to $K > 2$ binary latents. Throughout this section, the latents are uniform, $\sigma_x = 0$, and the readout remains at its ridge optimum as in Assum. 1.

D.1 Fourier modes and context-dependent concept directions

Let $\boldsymbol{s} = (s_1, \dots, s_K) \sim \mathrm{Unif}(\{-1, +1\}^K)$, $\boldsymbol{x} = \boldsymbol{\mu}(\boldsymbol{s})$, $\boldsymbol{y} = \boldsymbol{y}(\boldsymbol{s})$, and $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{x}$. Index modes by subsets $m \in \mathcal{M} = 2^{[K]}$, where $[K] = \{1, \dots, K\}$, and write $\chi_m(\boldsymbol{s}) = \prod_{a \in m} s_a$, with $\chi_\emptyset = 1$. By Theorem A.2, each vector-valued function has the unique expansion

$$\boldsymbol{f}(\boldsymbol{s}) = \sum_{m \subseteq [K]} \hat{\boldsymbol{f}}_m \chi_m(\boldsymbol{s}), \qquad \hat{\boldsymbol{f}}_m = 2^{-K} \sum_{\boldsymbol{s} \in \{-1,+1\}^K} \boldsymbol{f}(\boldsymbol{s}) \chi_m(\boldsymbol{s}). \tag{279}$$

Apply this to $\boldsymbol{\mu}, \boldsymbol{y}, \boldsymbol{z}$, so $\hat{\boldsymbol{z}}_m = \boldsymbol{W}\hat{\boldsymbol{x}}_m$. We extend Assum. 2 to all 2^K modes:

$$\langle \hat{\boldsymbol{x}}_m, \hat{\boldsymbol{x}}_{m'} \rangle = d_m^{(x)} \delta_{mm'}, \quad \langle \hat{\boldsymbol{y}}_m, \hat{\boldsymbol{y}}_{m'} \rangle = d_m^{(y)} \delta_{mm'}, \quad \langle \hat{\boldsymbol{z}}_m(0), \hat{\boldsymbol{z}}_{m'}(0) \rangle = q_{m,0} \delta_{mm'}. \tag{280}$$

As before, $d_m^{(x)}, d_m^{(y)}, q_m$ are squared component norms. Orthogonality of the characters under the uniform distribution is automatic; orthogonality of these vector coefficients is an additional assumption. The input, target, and hidden dimensions must accommodate their respective nonzero orthogonal components.

Fix a nonempty proper subset $i \subsetneq [K]$. The concept of interest is the binary parity χ_i; a singleton $i = \{a\}$ recovers the original latent s_a. Define its strict-superset modes and their number by

$$\mathcal{A}_i = \{j \subseteq [K] : i \subsetneq j\}, \qquad M_i = |\mathcal{A}_i| = 2^{K-|i|} - 1. \tag{281}$$

A context $\boldsymbol{c} \in \{-1, +1\}^{[K]\setminus i}$ fixes all latents outside i, giving $M_i + 1$ contexts. The concept direction is

$$\boldsymbol{u}_{i,\boldsymbol{c}} = \mathbb{E}[\boldsymbol{z} \mid \chi_i = +1, \boldsymbol{s}_{[K]\setminus i} = \boldsymbol{c}] - \mathbb{E}[\boldsymbol{z} \mid \chi_i = -1, \boldsymbol{s}_{[K]\setminus i} = \boldsymbol{c}]. \tag{282}$$

For $|i| > 1$, each conditional mean averages uniformly over the $2^{|i|-1}$ assignments within i with the specified parity. Characters involving a nonempty proper subset of i average to zero within each parity class; those involving none of i cancel in the difference. Only modes containing all of i survive, giving

$$\boldsymbol{u}_{i,\boldsymbol{c}} = 2 \left[\hat{\boldsymbol{z}}_i + \sum_{j \in \mathcal{A}_i} \chi_{j\setminus i}(\boldsymbol{c}) \hat{\boldsymbol{z}}_j \right]. \tag{283}$$

Thus a mode can be an interaction relative to one concept and the signal for another. The empty mode has no binary contrast, and the full mode $i = [K]$ has only one context, so neither has a pairwise abstraction score of the form below.

D.2 Abstraction as a ratio of Fourier energies

Generalize the two-context cosine by averaging over all distinct context pairs:

$$\alpha_i = \binom{M_i + 1}{2}^{-1} \sum_{\boldsymbol{c} < \boldsymbol{c}'} \frac{\boldsymbol{u}_{i,\boldsymbol{c}}^\top \boldsymbol{u}_{i,\boldsymbol{c}'}}{\|\boldsymbol{u}_{i,\boldsymbol{c}}\| \, \|\boldsymbol{u}_{i,\boldsymbol{c}'}\|}, \tag{284}$$

where the ordering only avoids double-counting. Write $R_i = \sum_{j \in \mathcal{A}_i} q_j$ for total interaction energy and, when the denominators are positive, define

$$\nu_{z,j}^i = \frac{q_j}{q_i}, \qquad \nu_{x,j}^i = \frac{d_j^{(x)}}{d_i^{(x)}}, \qquad \nu_{y,j}^i = \frac{d_j^{(y)}}{d_i^{(y)}}, \qquad j \in \mathcal{A}_i. \tag{285}$$

Proposition D.1 (Abstraction of an arbitrary Fourier mode)**.** *If the representation components are orthogonal and $q_i + R_i > 0$, then*

$$\alpha_i = \frac{q_i - R_i/M_i}{q_i + R_i} = \frac{1 - M_i^{-1} \sum_{j \in \mathcal{A}_i} \nu_{z,j}^i}{1 + \sum_{j \in \mathcal{A}_i} \nu_{z,j}^i}, \tag{286}$$

where the second equality requires $q_i > 0$.

Proof. Orthogonality and (283) give $\|\boldsymbol{u}_{i,\boldsymbol{c}}\|^2 = 4(q_i + R_i)$ for every context, and

$$\tfrac{1}{4}\boldsymbol{u}_{i,\boldsymbol{c}}^\top \boldsymbol{u}_{i,\boldsymbol{c}'} = q_i + \sum_{j \in \mathcal{A}_i} q_j \chi_{j\setminus i}(\boldsymbol{c})\chi_{j\setminus i}(\boldsymbol{c}'). \tag{287}$$

Each context character in this sum is nonconstant and balanced. Hence $\sum_{\boldsymbol{c}\neq\boldsymbol{c}'} \chi_{j\setminus i}(\boldsymbol{c})\chi_{j\setminus i}(\boldsymbol{c}') = -(M_i + 1)$, so its average over distinct pairs is $-1/M_i$. Averaging the inner products and dividing by the common squared norm proves (286). □

Consequently, $-1/M_i \le \alpha_i \le 1$, with perfect abstraction exactly when $q_i > 0$ and $R_i = 0$. If $q_i = 0 < R_i$, the score is $-1/M_i$; if both vanish, it is undefined. The normalization M_i counts all possible strict supersets, even when some modes have zero energy. In particular, the lower endpoint depends on the number of contexts. Equivalently, the fraction of contrast energy in the concept mode is $q_i/(q_i + R_i) = (1 + M_i\alpha_i)/(M_i + 1)$.

Example: three latents. For shape s, color c, and odor o, the shape directions are $\boldsymbol{u}_{S,(c,o)} = 2(\hat{\boldsymbol{z}}_S + c\hat{\boldsymbol{z}}_{SC} + o\hat{\boldsymbol{z}}_{SO} + co\hat{\boldsymbol{z}}_{SCO})$. Shape–color parity instead has two directions $\boldsymbol{u}_{SC,o} = 2(\hat{\boldsymbol{z}}_{SC} + o\hat{\boldsymbol{z}}_{SCO})$. Thus

$$\alpha_S = \frac{q_S - (q_{SC} + q_{SO} + q_{SCO})/3}{q_S + q_{SC} + q_{SO} + q_{SCO}}, \qquad \alpha_{SC} = \frac{q_{SC} - q_{SCO}}{q_{SC} + q_{SCO}}. \tag{288}$$

An interaction concept can therefore be perfectly abstract across its remaining contexts even when its constituent latents are not.

E Experimental details

E.1 Experimental details for motivating examples

E.1.1 Fig. 2A: synthetic ReLU network and probe generalization

For this figure we create a synthetic task reflecting the balanced 2×2 factorial design, where the two factors are shape and color. This gives us four classes: □, □, ○, ○. Each example is generated from latent coordinates for shape, color, and interaction. The figure uses input dimension 16, one nuisance coordinate with zero strength, and latent amplitudes 3.0 for shape, 2.0 for color, and 2.0 for the interaction. These amplitudes specify the generator before mixing and are not the Fourier energies of the resulting inputs. To simulate realistic noisy data, the latent coordinates are mixed by a random matrix with normalized columns, corrupted with additive Gaussian noise of standard deviation 2.0, passed through a pointwise $\tanh$ function, and then corrupted again with additive Gaussian noise of standard deviation 1.0. For each seed we generate 1000 training examples per class using the above procedure.

The model is a feedforward ReLU network with a bias-free linear readout. The network has 2 hidden layers with width 32, and ReLU after each hidden linear map. We train it to predict the two signed factor labels with mean-squared error using Adam, learning rate 10^{-3}, batch size 512, for 100 epochs. The plotted curves average over seeds $0, 1, 2$, with shaded bands showing the SEM across seeds.

At each checkpoint, we measure abstraction and probe transfer on a separate evaluation sample of 500 examples per class. The probe direction is the normalized red-context mean difference $\boldsymbol{w}_{\mathrm{red}} = \boldsymbol{u}_{\mathrm{red}}/\|\boldsymbol{u}_{\mathrm{red}}\|$. Blue examples are centered at the midpoint of their two class means, as in §A.5. The Gaussian prediction uses the empirical blue within-class covariance $\widehat{\boldsymbol{\Sigma}}_{\mathrm{blue}}$, giving $\Phi[-\|\boldsymbol{u}_{\mathrm{blue}}\|\alpha/(2\sqrt{\boldsymbol{w}_{\mathrm{red}}^{\top}\widehat{\boldsymbol{\Sigma}}_{\mathrm{blue}}\boldsymbol{w}_{\mathrm{red}}})]$. This is the empirical version of (71); the evaluation sample supplies both these moments and the empirical blue-context error.

E.1.2 Fig. 2B: steering-vector estimation from finite samples

We generate representations in $D = 16$ dimensions as

$$\boldsymbol{z} = s\sqrt{q_S}\boldsymbol{e}_S + c\sqrt{q_C}\boldsymbol{e}_C + sc\sqrt{q_{SC}}\boldsymbol{e}_{SC} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_D),$$

where $s, c \in \{-1, +1\}$ encode shape and color, and $\boldsymbol{e}_S, \boldsymbol{e}_C, \boldsymbol{e}_{SC}$ are the first three canonical unit vectors. We fix $q_S = 16$, $q_C = 4$, and $\sigma = 0.5$, set the constant Fourier component to zero, and sweep $\alpha \in \{0, 0.1, \ldots, 1\}$ by setting $q_{SC} = q_S(1-\alpha)/(1+\alpha)$. Thus α is the prescribed population abstraction, and the population steering vector $\Delta = 2\sqrt{q_S}\boldsymbol{e}_S$ remains fixed. These are population Fourier energies without sample-count rescaling; no network is trained.

For each (α, n), with $n \in \{100, 200, 300\}$, we draw n independent examples per shape ($2n$ total), sampling each color uniformly and each noise vector independently. Colors are balanced only in expectation, so their realized counts fluctuate within each shape. We compute $\widehat{\Delta} = \bar{\boldsymbol{z}}_+ - \bar{\boldsymbol{z}}_-$ and record $\|\widehat{\Delta} - \Delta\|^2$. Markers average this error over $R = 1000$ independent datasets per condition. Error bars are two-sided 95% Student-t confidence intervals for the mean, with half-width $t_{0.975,R-1}\,\mathrm{SD}/\sqrt{R}$, where SD is the sample standard deviation of the squared errors. Solid curves evaluate the exact prediction $\frac{2}{n}[q_S(1-\alpha)/(1+\alpha) + q_C + D\sigma^2]$ from (77) using the generating parameters, with no fitting.

E.2 Experimental details for theory-simulation agreement figures

We use the Fourier modes $m \in \mathcal{M} = \{\emptyset, S, C, SC\}$, with input and target energies $d_m^{(x)}, d_m^{(y)}$, representation energies q_m, and entanglement ratios $\nu_x = d_{SC}^{(x)}/d_S^{(x)}$, $\nu_y = d_{SC}^{(y)}/d_S^{(y)}$, and $\nu_z = q_{SC}/q_S$. All four-component tuples below follow the order $(\emptyset, S, C, SC)$. To retain the numerical scale of the simulation configurations, we report absolute energies as $Nd_m^{(x)}$, $Nd_m^{(y)}$, and Nq_m, where $N = 4n$ is the total number of examples in the balanced dataset specified for each experiment. Dividing the reported values by N recovers the Fourier energies used in the main text; entanglement ratios and abstraction scores are unchanged. For nonlinear networks, q_m and the feature energies k_m include the width normalization of §5, so no additional factor of D is needed in this conversion.

E.2.1 Fig. 3: deep linear networks

The simulations optimize squared error with the ridge readout at its instantaneous optimum (Assum. 1), using $n = 4$ examples per class and $N = 16$ examples in total. We report N times the class-centroid Fourier energies. To specify the additional finite-sample variation, let $\boldsymbol{\chi}_m \in \mathbb{R}^N$ contain the repeated class characters, and define

$$\boldsymbol{P}_{\text{within}} = \boldsymbol{I}_N - \frac{1}{N}\sum_{m\in\mathcal{M}} \boldsymbol{\chi}_m \boldsymbol{\chi}_m^\top, \qquad \boldsymbol{X}_{\text{data}}^\top \boldsymbol{X}_{\text{data}} = \sum_{m\in\mathcal{M}} d_m^{(x)} \boldsymbol{\chi}_m \boldsymbol{\chi}_m^\top + \rho_x \boldsymbol{P}_{\text{within}}.$$

We analogously define ρ_y and $\rho_z^{(\ell)}$ for targets and hidden representations. These residual directions have zero mean within each class and are separate from the four Fourier components; in particular, ρ_x is not the population noise variance σ_x^2. The residual coefficients $\rho_x, \rho_y, \rho_z^{(\ell)}$ are reported directly, without multiplication by N. This convention retains the fixed-sample geometry of the simulations. The simulator uses a summed squared loss with ridge coefficient 1; its population counterpart has $\gamma = N = 16$. All plotted times use the simulator clock t_{sim}, related to population time by $t = N t_{\text{sim}}$. Weight-space gradient flow is integrated with `DOP853`.

Panel A uses $L = 1$, with

$$(N d_m^{(x)})_m = (0.03, 0.16, 0.12, 0.0624), \qquad (N d_m^{(y)})_m = (0, 0.16, 0.12, 0.0128),$$

so $\nu_x = 0.39$ and $\nu_y = 0.08$. We initialize with $(N q_{m,0})_m = a(0.03, 0.16, 0.12, 0.08)$, with $a \in \{10, 1, 10^{-1}, 10^{-10}\}$, hence $\nu_{z,0} = 0.5$. This Fourier-aligned initialization scales representation energies by a. The residual coefficients are $\rho_x = 0.1$, $\rho_y = 0$, and $\rho_z(0) = 0.1a$. The horizon is $t_{\text{sim}} = 10^6$. Solid curves show the weight-space simulations, dotted curves the scalar energy predictions, and the horizontal reference the analytic terminal abstraction.

Panel B uses the saved $L = 6$ trajectory, with

$$(N d_m^{(x)})_m = (3.2, 16.0, 13.6, 7.2), \qquad (N d_m^{(y)})_m = (0, 16.8, 12.0, 2.4).$$

Thus $\alpha_x = (16.0 - 7.2)/(16.0 + 7.2) = 0.3793$ and $\alpha_y = 0.75$. The Fourier-aligned balanced initialization sets $(N q_m^{(L)}(0))_m = (0.08, 0.32, 0.22, 0.58)$. Intermediate energies obey $q_m^{(\ell)}(0) = (d_m^{(x)})^{1-\ell/L} (q_m^{(L)}(0))^{\ell/L}$. The residual coefficients are $\rho_x = 0.1$, $\rho_y = 0$, and $\rho_z^{(L)}(0) = 0.05$, with the same geometric interpolation across hidden layers. The horizon is $t_{\text{sim}} = 100$. Solid curves show $\alpha^{(\ell)}(t)$; dotted curves apply the layerwise interpolation law to α_x and the contemporaneous final-layer abstraction.

Panel C sweeps $L = 1, \ldots, 40$, with

$$(N d_m^{(x)})_m = (3.2, 16.0, 14.4, 11.2), \qquad (N d_m^{(y)})_m = (0, 16.0, 12.0, 1.6),$$

so $d_C^{(x)}/d_S^{(x)} = 0.90$, $d_C^{(y)}/d_S^{(y)} = 0.75$, $\nu_x = 0.70$, and $\nu_y = 0.10$. It uses the same final-layer initial energies, residual coefficients, and balanced factorization as panel B. The plotted simulation endpoints at $t_{\text{sim}} = 100$ are compared with the terminal depth law; the red reference is $\alpha_y = (1-0.1)/(1+0.1)$.

E.2.2 Fig. 5E–F: leaky-ReLU and erf nonlinear dynamics

The nonlinear simulators use the normalization described in §C.7: $\gamma = N\gamma_{\text{sim}}$ and $t = N t_{\text{sim}}$. We report N times the Fourier energies below and retain the simulator clock for the plotted horizons. The erf implementation uses $\phi(z) = \text{erf}(\beta_{\text{sim}} z/\sqrt{2})$, so its parameter corresponds to $\beta = \beta_{\text{sim}}/\sqrt{2}$ in §5. The numerical β and γ settings attached to the erf figure panels refer to these simulator parameters.

For Fig. 5E, the four-class calculation has $N = 4$, leaky ReLU with $\omega = 0.5$, and $\gamma_{\text{sim}} = 100$, or $\gamma = 400$. Its scaled energies are

$$(N d_m^{(x)})_m = (0, 1, 1, 0.6), \qquad (N d_m^{(y)})_m = (0, 1, 1, 0.1),$$
$$(N q_{m,0})_m = (0, 0.01, 0.01, 0.01).$$

Thus $\nu_x = 0.6$, $\nu_y = 0.1$, and $\nu_{z,0} = 1$. The width-512 simulation uses Gaussian projection: at each covariance derivative evaluation it resamples Gaussian neuron rows and matches their empirical covariance to the current covariance. This explicitly enforces the Gaussian closure used by the

theory. It uses seed 1, RK4 with step 0.01, and horizon $t_{\mathrm{sim}} = 10^4$. The theory curve integrates the corresponding closed covariance ODE; the red reference is the terminal fixed point $\alpha_{\mathrm{pre},\infty} = 0.6528$.

For Fig. 5F, the erf row simulation has $n = 4$, $N = 16$, $\beta_{\mathrm{sim}} = 0.1$, and $\gamma_{\mathrm{sim}} = 1000$, or $\gamma = 16000$. Its scaled energies are

$$(Nd_m^{(x)})_m = (0, 1, 1, 1), \qquad (Nd_m^{(y)})_m = (0, 1, 1, 0.1),$$

with $\nu_x = 1$, $\nu_y = 0.1$, and nominal initial scaled energies $(Nq_{m,0})_m = (3.0, 0.5, 0.5, 0.3)$. The inputs and targets have no within-class residual component; the initial preactivation covariance has $\rho_z(0) = 0.2$. Its constant component and within-class residuals are prescribed offsets in the input nullspace and remain fixed during training, as explained in §C.7. The width-2048 rows are sampled from the nominal Gaussian initialization with seed 0; their empirical covariance fluctuates about that initialization. Euler integration uses step 0.1 to $t_{\mathrm{sim}} = 10^4$, while the Gaussian-closure covariance ODE starts from the nominal covariance and is integrated with `DOP853`. The red reference is the linear terminal abstraction from Proposition 2.

The phase diagrams use entanglement fractions $\eta_x = \nu_x/(1 + \nu_x)$ and $\eta_y = \nu_y/(1 + \nu_y)$. Panels C–D evaluate a 1000×1000 grid over $[0, 0.999999]^2$ for $\omega = 0.2$ and $\omega = 0$, respectively, using Brent's method for (25). Panel B evaluates the linear terminal law. For the erf phase diagram in panel A, $N = 16$, $\beta_{\mathrm{sim}} = 1$, and $\gamma_{\mathrm{sim}} = 1000$. The scaled input and target energies satisfy $Nd_S^{(x)} = Nd_C^{(x)} = Nd_S^{(y)} = Nd_C^{(y)} = 1$, $Nd_{SC}^{(x)} = \nu_x$, and $Nd_{SC}^{(y)} = \nu_y$; their constant and within-class components vanish. The initialization multiplies the default preactivation covariance by 0.1, giving $(Nq_{m,0})_m = (0.3, 0.05, 0.05, 0.02)$, with $\rho_z(0) = 0.02$. The closed ODE uses RK4 with step 0.01 to $t_{\mathrm{sim}} = 10$ on a 10×10 grid over $\eta_x, \eta_y \in [0, 1]$. This panel reports abstraction at the simulated horizon.

E.3 Experimental details for 3dshapes experiments

E.3.1 Description of task

All 3dshapes experiments in this subsection use the red/blue cube/sphere subset of the 3dshapes dataset. It contains four classes, Red Sphere, Blue Sphere, Red Cube, and Blue Cube, with balanced class counts. The object hue is fixed to the nearest grid value matching red 0.0 or blue 0.7, and the object shape is matched exactly. Floor hue, wall hue, object scale, and object orientation are left free to vary. Unless otherwise stated, the processed dataset contains 1000 examples per class, for 4000 images total, with an 80/20 train/test split. Images use 64×64 resolution and are normalized channelwise by $(x - 0.5)/0.5$ for the trained convolutional models.

For the trained ConvNet and ResNet experiments, supervised targets are deterministic functions of the two binary latents with orthogonal Fourier components. The implementation normalizes the factor vectors over the full balanced manifest of $N = 4n$ images, giving

$$Nd_\emptyset^{(y)} = 0, \qquad Nd_S^{(y)} = Nd_C^{(y)} = 1, \qquad Nd_{SC}^{(y)} = \nu_y.$$

Targets are constant within each class, and the target entanglement ratio ν_y sets the interaction energy relative to the shape energy. Models are trained with vector mean-squared error, and the reported abstraction is measured on the training split.

E.3.2 Fig. 1B: feedforward convolutional network

The panel uses a feedforward convolutional network with 2 convolutional layers and 2 fully connected layers. It applies a 3×3 convolution from 3 channels to 32 channels, ReLU, a second 3×3 convolution from 32 to 32 channels, ReLU, adaptive average pooling, a width-256 fully connected ReLU layer, a width-64 fully connected ReLU layer, and a linear head to the label dimension. The final analyzed hidden representation is the width-64 `fc2` activation.

The targets use $\nu_y = 0.125$. The model is initialized with $\sigma_w^2 = 0.15$ and trained with AdamW, learning rate 10^{-3}, weight decay 10^{-5}, batch size 128, and a cosine learning-rate schedule with minimum learning rate 10^{-5}. Training runs for 100 epochs, and abstraction is measured every epoch. The plotted curve reports the mean and SEM across 5 seeds.

E.3.3 Fig. 4: Small ResNets

Fig. 4 uses small ResNets with a 3×3 convolutional stem with ReLU, followed by L residual blocks. Each block contains two 3×3 convolutions with a ReLU between them, adds the residual connection, and applies a final ReLU. The representation at layer ℓ is the adaptive-average-pooled post-block activation. A linear readout maps the final pooled representation to the output.

The initialization-scale sweep in panel B uses $n = 1000$, $\nu_y = 0.01$, depth $L = 2$, width 32, and initialization scales $\sigma_w^2 \in \{0.03, 0.3, 3.0, 30.0\}$. For each scale we train seeds $0, \ldots, 19$ for 50 epochs with AdamW, learning rate 10^{-3}, weight decay 10^{-5}, batch size 128, and cosine decay to 10^{-5}. Panel B plots maximum shape abstraction attained during training for each scale averaged over seeds and with SEM error bars.

The depth sweep in panel C uses $n = 1000$, $\nu_y = 0.25$, width 32, initialization scale $\sigma_w^2 = 0.1$, and depths $L \in \{2, 4, 6, 8\}$. For each depth we train seeds $0, \ldots, 19$ for 50 epochs with AdamW, learning rate 10^{-3}, weight decay 10^{-8}, batch size 128, and cosine decay to 10^{-5}. Panel C plots the final layer's shape abstraction after the last epoch averaged across seeds with SEM error bars.

Panel D uses the same training run data as panel C, but uses just the depth $L = 8$ runs to plot the abstraction layerwise across $\ell \in \{1, \ldots, 8\}$. The plotted curves average over seeds $0, \ldots, 19$ with SEM error bars.

E.3.4 Fig. 6A: DINOv3 ViT-L/16 local GELU ablation

We use pretrained DINOv3 ViT-L/16 from the official `facebookresearch/dinov3` repository, loaded as `dinov3_vitl16`. The run uses all 4000 examples from `rsbc_3dshapes_64`, batch size 32, and model input size 256. Representations are extracted from the post-block residual stream by forward hooks on each of the 24 transformer blocks. To obtain one embedding vector per sample we pool by averaging the captured token representations over the token dimension.

For the baseline, we run the unmodified DINOv3 model and compute shape abstraction from the post-block representation at each residual layer. For the ablation condition at layer ℓ, we temporarily replace the first activation site in `blocks.<layer>.mlp` with the identity map, run a forward pass, and compute abstraction from the post-block output of the same layer ℓ. The patch is local to that forward pass and is restored before the next layer is evaluated. In this model the patched object paths are `blocks.<layer>.mlp.act` for layers $0, \ldots, 23$.

E.4 Experimental details for Gemma 4 ablation experiments

This section gives details for the Gemma 4 ablation experiments in Fig. 6B-C.

Gemma model and concept set. Fig. 6B–C uses `google/gemma-4-E2B` from Hugging Face, loaded as a causal language model with 35 decoder blocks. User-facing layer index $\ell = 0$ denotes the output of decoder block 0, and $\ell = 34$ denotes the final residual stream. Following [3], we study bilingual concepts in the form of translation directions from one language to another. We use the same four language pairs French-Spanish, French-German, English-French and German-Spanish considered in [3] as well as an additional 14 English–X translation directions (where X is Arabic, Chinese, Dutch, German, Indonesian, Italian, Japanese, Korean, Persian, Polish, Portuguese, Russian, Spanish, and Turkish). Each concept is instantiated by 80 ordered word pairs comprising common words. See Table 1 for the English–Spanish word pairs.

English	Spanish	English	Spanish	English	Spanish	English	Spanish
water	agua	rain	lluvia	sister	hermana	one	uno
food	comida	snow	nieve	son	hijo	two	dos
bread	pan	dog	perro	daughter	hija	three	tres
fish	pez	cat	gato	red	rojo	four	cuatro
fruit	fruta	horse	caballo	green	verde	five	cinco
apple	manzana	bird	pájaro	blue	azul	six	seis
house	casa	chicken	pollo	white	blanco	seven	siete
car	coche	day	día	black	negro	eight	ocho
money	dinero	night	noche	month	mes	nine	nueve
book	libro	big	grande	year	año	ten	diez
pen	bolígrafo	small	pequeño	gold	oro	music	música
road	carretera	good	bueno	silver	plata	mathematics	matemáticas
sun	sol	bad	malo	wood	madera	physics	física
moon	luna	new	nuevo	metal	metal	chemistry	química
fire	fuego	old	viejo	stone	piedra	biology	biología
hand	mano	hot	caliente	grass	hierba	engineering	ingeniería
eye	ojo	cold	frío	leaf	hoja	history	historia
tree	árbol	man	hombre	sand	arena	geography	geografía
mountain	montaña	woman	mujer	river	río	law	derecho
cloud	nube	brother	hermano	beach	playa	english	inglés

Table 1: The 80 ordered English–Spanish word pairs used for the English–Spanish concept in the Gemma ablation experiments.

Tokenization and representation extraction. Each lexical item is presented in the standalone template `{word}` with `add_special_tokens = false`. Each word is represented by the hidden state of its final non-padding token. This lets the experiment use both single-token and multi-token lexical items while still attaching each representation to a single isolated word. The final run uses a fixed split seed of 1234 and exactly 80 valid pairs for every concept.

Abstraction metric. For a concept with ordered pairs $\{(u_i, v_i)\}_{i=1}^{m}$, let $\mathbf{h}_i^{(\ell)}$ denote the residual-stream representation of word u_i at layer ℓ, and let $\mathbf{g}_i^{(\ell)}$ denote the representation of word v_i. We form ordered difference vectors

$$\mathbf{d}_i^{(\ell)} = \mathbf{h}_i^{(\ell)} - \mathbf{g}_i^{(\ell)} \tag{289}$$

The abstraction score is the mean pairwise cosine across all concept vectors:

$$\alpha^{(\ell)} = \frac{2}{m(m-1)} \sum_{1 \le i < j \le m} \frac{\mathbf{d}_i^{(\ell)} \cdot \mathbf{d}_j^{(\ell)}}{\max(\|\mathbf{d}_i^{(\ell)}\|, \varepsilon) \max(\|\mathbf{d}_j^{(\ell)}\|, \varepsilon)}, \qquad \varepsilon = 10^{-8}. \tag{290}$$

In the final Gemma run, $m = 80$, so each concept contributes $\binom{80}{2} = 3160$ pairwise terms. The numerical convention assigns zero contribution to zero difference vectors, including identical source and target words.

Local GELU ablation. For Fig. 6B–C, we replace the activation function in the final decoder block's MLP with the identity map for that forward pass and read out the final residual stream. The patched module is `model.language_model.layers.34.mlp.act_fn`. Attention, the remaining MLP operations, and all other blocks are unchanged.

Probe generalization error. To test the practical consequence predicted by (8), we evaluate whether a probe trained on one subset of lexical contexts transfers to unseen pairs from the same concept. Each concept's 80 ordered pairs are split into 40 train pairs and 40 test pairs. Because every concept has the same number of pairs, the same index split is reused across concepts. Each ordered pair contributes two labeled examples, $u_i \mapsto 0$ and $v_i \mapsto 1$. Train and test features are centered by the training mean, and we fit a logistic-regression probe with LBFGS, learning rate 1.0, 200 iterations, and tolerance 10^{-7}. Reported accuracies are test accuracies on the held-out 40 pairs.

Numerical summary. In the final-layer diagonal comparison, abstraction increases for all 18 concepts, with concept-wise means moving from 0.207 in baseline to 0.242 after ablation. Held-out probe accuracy increases on average from 0.944 to 0.953, with 17 non-decreases out of 18 concepts. The only decrease occurs for French–Spanish, where accuracy changes slightly from 0.850 to 0.838. These results are the basis for Fig. 6B–C.

E.5 Experimental details for macaque ventral-stream analysis

This section describes the analysis pipeline used to produce the results in Fig. 7 and the associated robustness checks.

Dataset. We use the public Brain-Score release `MajajHong2015.public`. The public release is close to, but not identical with, the original internal dataset used in the Majaj/Hong studies. In particular, it exposes only variation labels 0 and 3 rather than the explicit low/medium/high variation bins described in the original paper. All results reported here are computed directly from the public release.

Preprocessing. Raw repeated presentations are first averaged to a single response vector per stimulus. Each neuroid is then z-scored across all 3200 averaged public stimuli before subsetting to the four focal categories. This yields a normalized response matrix that does not rely on blank-screen baselines, which are not exposed in the public release in the same form as in the original analyses.

Conceptual design. We restrict attention to four categories: Boats, Tables, Fruits, and Animals. These define a 2×2 factorial structure in which limbedness and naturalness act as the two abstract variables. Writing $\boldsymbol{\mu}_{\text{Boat}}, \boldsymbol{\mu}_{\text{Table}}, \boldsymbol{\mu}_{\text{Fruit}}, \boldsymbol{\mu}_{\text{Animal}}$ for category centroids in a given neural population, we define

$$\mathbf{u}_{\text{limbed},1} = \boldsymbol{\mu}_{\text{Table}} - \boldsymbol{\mu}_{\text{Boat}}, \qquad \mathbf{u}_{\text{limbed},2} = \boldsymbol{\mu}_{\text{Animal}} - \boldsymbol{\mu}_{\text{Fruit}}, \tag{291}$$

$$\mathbf{u}_{\text{natural},1} = \boldsymbol{\mu}_{\text{Fruit}} - \boldsymbol{\mu}_{\text{Boat}}, \qquad \mathbf{u}_{\text{natural},2} = \boldsymbol{\mu}_{\text{Animal}} - \boldsymbol{\mu}_{\text{Table}} \tag{292}$$

The abstraction score for each concept is the cosine similarity between its two context-specific vectors. The code variable `is_organic` corresponds to what we call *naturalness* in the paper.

Balanced centroid construction. The public release contains 8 objects per focal category. In each bootstrap resample, we draw 8 objects per category with replacement. Within each sampled object and variation bin, images are resampled with replacement, averaged within the cell, averaged equally across the available variation bins for that object, and then averaged equally across sampled objects to form the category centroid. This hierarchical averaging ensures that the abstraction score is not dominated by categories, objects, or variation bins with more images.

Site matching and bootstrap design. The public release contains 168 IT sites (58 from Chabo and 110 from Tito) and 88 V4 sites (70 from Chabo and 18 from Tito). Our primary abstraction analysis is the pooled site-matched comparison used in Fig. 7B. On each bootstrap iteration, we retain all 88 V4 sites and sample 88 IT sites without replacement to match the representation dimensions. This panel uses 500 bootstrap resamples with random seed 0. In each resample, we jointly redraw objects, images, and the matched IT site subset.

Linear-probe generalization. Fig. 7C uses bias-free logistic-regression probes for limbedness (Animals and Tables versus Boats and Fruits) and naturalness (Animals and Fruits versus Boats and Tables). We run 10 repeats with seed 7, each using a stratified 50/50 split within each category, giving 800 training and 800 test images. Splits are shared across concepts and areas; each repeat uses all 88 V4 sites and samples 88 IT sites without replacement. Both splits are centered by the training mean, and unregularized probes are fit with LBFGS, 200 iterations, and tolerance 10^{-7}. The panel reports mean accuracy and the 2.5th and 97.5th percentiles across repeats. Mean IT/V4 accuracies are $0.870/0.761$ for limbedness and $0.864/0.775$ for naturalness. All four categories occur in both splits, so this measures generalization to held-out images.

Additional metrics and controls. Besides the raw cosine score, the pipeline also records vector magnitude, a split-half cross-validated cosine, a permutation null obtained by randomly permuting the category labels, and a separate context-transfer decoder control. This control follows the cross-context design of (8): for each concept, we train a linear direction in one context and test it on the opposite context, then average the two transfer directions.

Robustness summary. The pooled site-matched IT–V4 gap reported in the main text is 0.676 for limbedness and 0.682 for naturalness, with both 95% percentile bootstrap intervals excluding zero. The descriptive all-sites comparison gives essentially the same means and intervals. The Chabo-only matched analysis is even stronger in the same direction, while the Tito-only analysis is directionally similar but noisier because the public release contains only 18 Tito V4 sites. The permutation null is centered near zero in both areas, and the transfer-decoder control shows the same ordering as the cosine metric: in the pooled site-matched analysis, transfer accuracy is 0.599 versus 0.371 for limbedness and 0.655 versus 0.389 for naturalness in IT and V4 respectively.

Abstraction vs factorization. Our findings are consistent with prior work showing greater factorization in IT than V4 [54]. However, factorization and abstraction measure different things: factorization measures orthogonality between different concept's subspaces (sometimes called disentanglement), while abstraction measures the alignment of counterfactual concept vectors. As discussed in [54], the two measures are related in the sense that a fully factorized representation should be highly abstract.

F Other

F.1 Compute resources

Neural network training experiments were run on a cluster with 8 NVIDIA H100 80GB HBM3 GPUs, 192 logical CPU cores from two Intel Xeon Platinum 8468 sockets, and approximately 1 TB of system RAM. Individual training runs used a single H100 GPU. All other analysis was run locally on a MacBook Air M2 with 8 GB of RAM. All code was implemented in Python 3.10-3.12. All neural network training experiments used stable 2.x versions of PyTorch.

F.2 Assets and licenses

We use the 3dshapes dataset under Apache-2.0, DINOv3 under the DINOv3 License, Gemma 4 E2B under Google's Gemma usage terms, and the public Brain-Score MajajHong2015.public assembly under the public Brain-Score/DiCarlo Lab access terms.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made, including the contributions made in the paper. In both, we explicitly restrict the scope of our claims to a minimal linear network setting. The claims made match theoretical and experimental results, and in the main text we discuss how much the results can be expected to generalize to other settings.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: The paper discusses several limitations of the work, including the simplifying assumptions made in the minimal linear network setting and the difficulty of extending the theory to deep nonlinear networks. Limitations are discussed both while introducing simplfying assumption (i.e. near each assumption and setting environment) and in the separate "Limitations" (bolded) section in the final Discussion section.

   Guidelines:

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

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [Yes]

   Justification: All assumptions are explicitly stated in the main text and also more extensively detailed in the appendix. In the main text we sometimes provide an informal / abridged version or sketch of a mathematical argument, but all mathematical results are proved and derived in the appendix.

   Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: Yes, Appendix E provides full details for the main experimental results (i.e. §4.3 and §6). In addition, we plan to release the code used for these experiments at a late date to further facilitate reproducibility.

   Guidelines:

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

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No]

   We plan to release the code at a later date and in the meantime provide explicit instructions for how to reproduce the main experimental results in Appendix E.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: All the training and test details are specified in Appendix E, including data splits, hyperparameters, how they were chosen, and type of optimizer.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

Justification: In the training experiments §4.3 we report error bars for the standard error of the mean (SEM) across random seeds, and for the neural data experiment in Fig. 7 we report error bars using 95% bootstrap intervals. In Appendix E we provide full statistical details.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: The paper provides information on the compute resources used for the experiments in Appendix F.1. In addition, we plan to release the code used for these experiments at a later date.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The authors have reviewed the NeurIPS Code of Ethics and ensure that the research conducted in the paper conforms to its principles.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: As a theory-based paper, we do not see a direct path to any negative societal impacts of the work performed. We also do not see any direct positive societal impacts, but we hope that the work will contribute to the scientific community's understanding of representation learning in neural networks.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [N/A]

    Justification: As a theory-based work, the paper poses no such risks.

    Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: All datasets and models used in the paper are properly cited, and the license and terms of use are properly respected. For example, for the neural data used in Fig. 7, we cite the original paper that produced the dataset, and state which version of the dataset is used in the appendix.

Guidelines:

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

13. **New assets**

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

Answer: [N/A]

Justification: The paper does not introduce new assets.

Guidelines:

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

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

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

Answer: [N/A]

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

Guidelines:

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

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

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

    Answer: [N/A]

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

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: The paper's core method development does not involve LLMs as any important, original, or non-standard components

    Guidelines:

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