Learning Gaussian Multi-Index Models with Gradient Flow: Time Complexity and Directional Convergence

Berfin Şimşek
Flatiron Institute

Amire Bendjeddou
EPFL

Daniel Hsu
Columbia University

Abstract

This work focuses on the gradient flow dynamics of a neural network model that uses correlation loss to approximate a multi-index function on high-dimensional standard Gaussian data. Specifically, the multi-index function we consider is a sum of neurons $f^*(\mathbf{x}) = \sum_{j=1}^{k} \sigma^*(\mathbf{v}_j^T \mathbf{x})$ where $\mathbf{v}_1, \dots, \mathbf{v}_k$ are unit vectors, and σ^* lacks the first and second Hermite polynomials in its Hermite expansion. It is known that, for the single-index case ($k=1$), overcoming the search phase requires polynomial time complexity. We first generalize this result to multi-index functions characterized by vectors in arbitrary directions. After the search phase, it is not clear whether the network neurons converge to the index vectors, or get stuck at a sub-optimal solution. When the index vectors are orthogonal, we give a complete characterization of the fixed points and prove that neurons converge to the nearest index vectors. Therefore, using $n \asymp k \log k$ neurons ensures finding the full set of index vectors with gradient flow with high probability over random initialization. When $\mathbf{v}_i^T \mathbf{v}_j = \beta \geq 0$ for all $i \neq j$, we prove the existence of a sharp threshold $\beta_c = c/(c+k)$ at which the fixed point that computes the average of the index vectors transitions from a saddle point to a minimum. Numerical simulations show that using a correlation loss and a mild overparameterization suffices to learn all of the index vectors when they are nearly orthogonal, however, the correlation loss fails when the dot product between the index vectors exceeds a certain threshold.

Preliminary work. Under review by AISTATS 2025. Do not distribute.

1 Introduction

Suppose a neural network model is trained to approximate the input-output pairs generated by a multi-index function by following the gradients of the loss function. Such loss functions are notoriously non-convex and complex to analyze in general due to the commonly observed degeneracy at initialization and a large number of fixed points, including local minima. As a result, a randomly initialized algorithm may take a long time to find the subspace of the index vectors, and even then, may fail to match the index vectors with network neurons. The failure of the algorithm may be due to poor initialization–which can be overcome using overparameterization–or due to the emergence of local minima related to the geometry of the index vectors.

More concretely, we study a stylized setting where the multi-index function is $f^*(\mathbf{x}) = \sum_{j=1}^{k} \sigma^*(\mathbf{v}_j^T \mathbf{x})$ for unit-norm index vectors $\mathbf{v}_1, \dots, \mathbf{v}_k$, while the neural network is $f(\mathbf{x}) = \sum_{i=1}^{n} \sigma(\mathbf{w}_i^T \mathbf{x})$ for unit-norm so-called neurons (that are also vectors) $\mathbf{w}_1, \dots, \mathbf{w}_n$. The activation functions $\sigma^*, \sigma : \mathbb{R} \to \mathbb{R}$ may differ but are fixed in both networks. We use the correlation loss as the training objective

$$\min C - \mathbb{E}_{\mathbf{x} \sim \mathcal{N}(0, I_d)} \left[\sum_{i=1}^{n} \sigma(\mathbf{w}_i^T \mathbf{x}) \sum_{j=1}^{k} \sigma^*(\mathbf{v}_j^T \mathbf{x}) \right] \quad (1)$$

where the expectation is taken over the standard Gaussian distribution $\mathcal{N}(0, I_d)$ in $\mathbb{R}^d$. This setting is also known as the teacher-student setup in statistical physics literature (Saad and Solla, 1995; Goldt et al., 2019). Its variations are extensively used to study the ability of neural networks to learn a low-dimensional subspace in high-dimensions (Ba et al., 2022; Bietti et al., 2022; Barak et al., 2022; Damian et al., 2022; Abbe et al., 2023; Berthier et al., 2023; Bietti et al., 2023) but using modified algorithms that may or may not be related to the behavior of gradient flow. Exciting recent work (Glasgow, 2023; Oko et al., 2024; Arous et al., 2024) proved global convergence results for gradient flow and characterized the end-to-end be-

havior of individual neurons. Some of these works and many others have provided sample complexity guarantees in the scenario when gradient flow is successful in the population loss limit. The question of optimal sample complexity in such scenarios is an active area of research and requires a careful selection of learning rate and whether to reuse the data to mitigate the noise due to finite sampling (Damian et al., 2023; Dandi et al., 2024; Lee et al., 2024; Arnaboldi et al., 2024).

In this work, we study the gradient flow of the population loss and characterize whether it succeeds, or fails. This question is studied by Safran and Shamir (2018) through high precision numerical simulations when $\sigma(x) = \sigma^*(x) = \max(0, x)$, showing that gradient flow fails when $n = k \geq 6$ under the mean squared error (MSE) loss. Arjevani and Field (2021); Şimşek et al. (2023) gave algebraic expressions for the fixed points associated with partitions of the symmetry group for the ReLU and erf activation functions, respectively. The combinatorial growth of the number of fixed points for the MSE loss (Şimşek et al., 2021) makes it very challenging to study the end-to-end gradient flow dynamics. Indeed, this is only achieved so far for a single-index model (Xu and Du, 2023) thanks to the absence of the fixed points emerging from the combinations of the index vectors, or for quadratic activation function (Martin et al., 2024) which trades the permutation symmetry with rotational symmetry. In this paper, following Bietti et al. (2023); Arous et al. (2024), we study the correlation loss instead of the MSE loss, consider a broad family of activation functions and learning multi-index models. A key difference is that we do not restrict the weight space to an orthogonal frame (aka Stiefel manifold), which allows studying success or failure of gradient flow when learning index vectors in more general positions.

More concretely, the model (1) features a perfect decoupling of neural network neurons. Thus, the multi-neuron problem reduces to studying the trajectory of a single neuron given the initial condition, and how good the random initialization over multiple neurons is. The evolution of a single neuron is described by the following dynamical system

$$\frac{d}{dt}\mathbf{w} = -(I_d - \mathbf{w}\mathbf{w}^T)\nabla L(\mathbf{w})$$
$$L(\mathbf{w}) = C - \mathbb{E}\left[\sigma(\mathbf{w}^T\mathbf{x})\sum_{j=1}^{k}\sigma^*(\mathbf{v}_j^T\mathbf{x})\right] \qquad (2)$$

where ∇ denotes the standard Euclidean gradient, $(I_d - \mathbf{w}\mathbf{w}^T)$ is the projection onto the tangent space of the unit sphere at $\mathbf{w}$, and expectation is always taken with respect to the standard Gaussian input hence $\mathcal{N}(0, I_d)$ notation is dropped hereafter. Mondelli and Montanari (2019) studied the same model and argued that matching the neurons to the index vectors is computationally hard when the number of index vectors exceeds $d^{3/2}$ under certain complexity-theoretic assumptions on tensor decomposition. We take a purely geometric approach and prove that a special fixed point—one that computes the average of the index vectors—transitions from a saddle point to a minimum when the dot product between distinct index vectors exceeds an explicit threshold. Overall our analysis shows a dichotomy in gradient flow behavior: when the index vectors are orthogonal, we prove that the single neuron converges to the nearest index vector, whereas, when the index vectors are too close or too many with small positive dot products, the single neuron converges to the average of the index vectors, hence failing to match the neurons to the index vectors. We partially explain the latter behavior by proving the saddle-to-minimum transition and providing complementary numerical simulations. In particular, we show

- while our time complexity analysis is similar to recent work following Ben Arous et al. (2021) when considering population gradient flow, it applies to the most general setting when the index vectors are in arbitrary directions (Section 2.1);
- to show convergence to the nearest index vector, we identify a Lyapunov function that has a monotonic behavior over time (Section 2.2), as a result, a mild overparameterization of $k\log(k)$ neurons is sufficient for learning all of the index vectors with gradient flow, starting from random initialization (Section 2.3);
- we then focus on the effect of the geometry of index vectors, prove a saddle-to-minimum transition (Section 3), and numerically show that the gradient flow fails to match the neurons to the index vectors when the index vectors get too close to each other (Section 4);
- finally, we establish an exact correspondence between tensor decomposition and our neural network model (1) which shows the fundamental difficulty of studying gradient flow for arbitrary geometries since most tensor problems are NP hard (Hillar and Lim, 2013).

1.1 Notation

We use bold letters to denote vectors such as $\mathbf{w}, \mathbf{v}, \mathbf{x}$ and capital letters for the matrices such as $W = [\mathbf{w}_1, \ldots, \mathbf{w}_n]$ which has size $d \times n$ and $V = [\mathbf{v}_1, \ldots, \mathbf{v}_k]$ which has size $d \times k$. The matrix of dot products between the index vectors is denoted by $A = V^T V$ which

has size $k \times k$. We assume that the index vectors are linearly independent, hence A is invertible. The set of indices is denoted by $[k] = \{1, \ldots, k\}$. Since the input data distribution is standard Gaussian, it is natural to expand the activation functions $\sigma, \sigma^* : \mathbb{R} \to \mathbb{R}$ using Hermite polynomials. Specifically, the following inner product

$$\langle f_1, f_2 \rangle_\phi := \int_{\mathbb{R}} f_1(x) f_2(x) \phi(x) dx, \quad \phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$$

is used to apply a Gram-Schmidt process to the sequence of monomials $1, x, x^2, x^3, \ldots$ which gives the probabilist's Hermite polynomials. It is convenient to use the normalized sequence

$$h_0 = 1, \;\; h_1(x) = x, \;\; h_2(x) = \frac{x^2 - 1}{\sqrt{2}}, \;\; h_3(x) = \frac{x^3 - 3x}{\sqrt{3!}}, \ldots$$

which ensures $\langle h_p, h_{p'} \rangle_\phi = \delta_{pp'}$. We assume that the activation functions have finite norm $\|\sigma\|_\phi, \|\sigma^*\|_\phi < \infty$ where the norm here is defined with respect to the inner product $\|\sigma\|_\phi = \langle \sigma, \sigma \rangle_\phi^{1/2}$. The activation functions $\sigma, \sigma^* : \mathbb{R} \to \mathbb{R}$ are then expanded as

$$\sigma(x) = \sum_{p=1}^{\infty} a_p h_p(x), \qquad \sigma^*(x) = \sum_{p=1}^{\infty} b_p h_p(x),$$

where the convergence holds with respect to the norm $\|\cdot\|_\phi$. The sequences $(a_p)_{p \geq 1}, (b_p)_{p \geq 1}$ are called the Hermite coefficients. For inner products with respect to two-dimensional Gaussian distribution, we use expressions involving the Hermite coefficients. For the Euclidean ℓ_p norm of vectors, we use $\|\cdot\|_p$.

1.2 Low Dimensional ODE

The ODE (2) describes the movement of d-dimensional vector $\mathbf{w}$ (also called a neuron) on the unit sphere. Due to the rotational invariance of the standard Gaussian, the two-point Gaussian integral can be expressed as

$$g_{\sigma,\sigma^*}(\mathbf{w}^T \mathbf{v}_j) = \mathbb{E}[\sigma(\mathbf{w}^T \mathbf{x}) \sigma^*(\mathbf{v}_j^T \mathbf{x})]$$

where $g_{\sigma,\sigma^*} : [-1, 1] \to \mathbb{R}$ is the natural extension of the so-called dual activation function (Daniely et al., 2016) to the case when σ and σ^* may be non-matching. As a result, one can express the correlation loss in terms of the dot products between $\mathbf{w}$ and $\mathbf{v}_1, \ldots, \mathbf{v}_k$. This enables us to express the surface of the loss $L(\mathbf{w})$ using k variables

$$L_0(V^T \mathbf{w}) = L(\mathbf{w}).$$

The loss L_0 can be viewed as mapping the subspace spanned by the index vectors to the loss values. In effect, it will be sufficient to analyze the behavior of the induced flow on this subspace.

Lemma 1.1. *Assume that $\mathbf{w}(t)$ solves the ODE (2) given an initial condition $\mathbf{w} = \mathbf{w}_0$. Then the vector of dot products $\mathbf{u}(t) = V^T \mathbf{w}(t)$ solves the following ODE*

$$\frac{d}{dt} \mathbf{u} = -(A - \mathbf{u}\mathbf{u}^T) \nabla L_0(\mathbf{u})$$

with an initial condition $\mathbf{u}_0 = V^T \mathbf{w}_0$.

The proof uses a simple chain rule (see Appendix A). Note that when the index vectors are orthogonal, the induced flow is a spherical gradient flow.

Using a nice property of the Hermite polynomials (O'Donnell, 2021, Chapter 11.2), one can expand the dual activation function, hence the loss L_0 is in the following form

$$L_0(\mathbf{u}) = C - \sum_{p=1}^{\infty} a_p b_p \|\mathbf{u}\|_p^p \tag{3}$$

where $\mathbf{u} = V^T \mathbf{w}$. The domain of the loss L_0 is determined by the geometry of the index vectors and is an ellipsoidal ball $D = \{\mathbf{u} : \mathbf{u}^T A^{-1} \mathbf{u} \leq 1\}$. For orthogonal index vectors, note that the domain is a unit ball since $A = I_k$. The derivation of the boundary uses simple linear algebra and is given in Appendix A.

This paper considers the case when $a_p b_p = 0$ for $p = 1$ which implies that $\mathbf{u} = 0$ is a fixed point of the loss since $\nabla L_0(\mathbf{u})\big|_{\mathbf{u}=0} = 0$. The lowest degree Hermite polynomial in the expansion determines the degree of degeneracy of the fixed point $\mathbf{u} = 0$. This important quantity of the target activation function $\sigma^* : \mathbb{R} \to \mathbb{R}$, that is,

$$p^* := \operatorname{argmin}_{p \geq 1} (b_p \neq 0).$$

is called the information exponent in the literature (Dudeja and Hsu, 2018; Ben Arous et al., 2021). We are interested in the degenerate case when $p^* \geq 3$. To ensure the decay of the loss near $\mathbf{u} = 0$, we assume $(a_p b_p)\big|_{p=p^*} > 0$. For technical reasons, we also assume the following throughout the paper.

Assumption 1.1. *The Hermite coefficients of the activation functions σ, σ^* have the same sign:*

$$c_p := a_p b_p \geq 0 \quad \textit{for all} \quad p \geq p^*.$$

Assumption 1.1 is satisfied for example for the matching activation functions $\sigma = \sigma^*$. It may be possible to relax this Assumption to allow some negative coefficients in the higher-order terms as long as there are positive coefficients with larger magnitudes in the lower-order terms.

Assumption 1.2. *We assume that the following series converges*

$$\sum_{p \geq p^*} a_p b_p p \lambda_{\max}(A)^{\frac{p}{2}} < \infty.$$

Assumption 1.2 ensures that the gradient of the correlation loss is well defined.

Assumption 1.3. *For the general case when both σ and σ^* are non-even activation functions, we assume that the index vectors have a positive dot product*

$$\mathbf{v}_j^T \mathbf{v}_{j'} \geq 0 \quad \textit{for all} \quad j, j' \in [k].$$

The assumption is wlog when σ or σ^* is even because one can flip the sign of any index vector without changing the loss.

We need this assumption to avoid situations such as the following worst case, $\mathbf{v}_j = -\mathbf{v}_{j'}$ for a pair of j, j' (in the general case of non-even activations) where the two index vectors would be pulling the single neuron in the opposite directions. In particular, for odd activations, the gradient corresponding to index vector j and j' cancels out hence learning is impossible in this worst case. For general non-odd activations, this situation creates an issue in the analysis hence Assumption 1.3.

2 Learning Index Vectors

2.1 Time Complexity

In this Subsection, we consider the general case when the index vectors are in arbitrary positions. At initialization, a vector $\mathbf{w}$ (neuron) is sampled from the uniform measure on the unit sphere. The dot product between this vector and an unknown vector is small: with high probability, $\mathbf{w}(0)^T \mathbf{v}_0 = \Theta(d^{-1/2})$ when d is large. Hence, at initialization, the vector is not specialized to any direction. Assume $k \ll d$ in this Subsection. The norm of the projection to the subspace of the index vectors has then the same order $\|V^T \mathbf{w}(0)\|_2 = \Theta(d^{-1/2})$.

Following gradient flow, the vector will aggregate the spherical gradient of the correlation loss

$$\nabla^S L(\mathbf{w}) = \sum_{p \geq p^*} a_p b_p p \Big(\sum_{j=1}^{k} (\mathbf{v}_j^T \mathbf{w})^{p-1} \mathbf{v}_j - \sum_{j=1}^{k} (\mathbf{v}_j^T \mathbf{w})^p \mathbf{w} \Big)$$

which points to the k-dimensional subspace spanned by the index vectors when $\mathbf{v}_j^T \mathbf{w}$ is vanishing. How much time T is necessary and sufficient to find the subspace, i.e., to reach non-vanishing projection $\|V^T \mathbf{w}(T)\|_2 = \|\mathbf{u}(T)\|_2 = \Theta(1)$?

Theorem 2.1 (Time complexity)**.** *Assume that d is large, $\mathbf{u}_j(0) = \Theta(d^{-1/2})$ and $\mathbf{u}_j(0) > 0$ for all $j \in [k]$. For $p^* \geq 2$, we show the following time complexities*

- *if $p^* \geq 3$, $T = \Theta(d^{p^*/2-1})$ is necessary and sufficient to reach $\|\mathbf{u}(T)\|_2 = \Theta(1)$,*
- *if $p^* = 2$, $T = \Theta(\log(d))$ is necessary and sufficient to reach $\|\mathbf{u}(T)\|_2 = \Theta(1)$.*

Proof sketch. The key step in the proof is confining the time evolution of $S = \sqrt{\mathbf{u}^T A^{-1} \mathbf{u}}$ which is the generalized Euclidean norm of $\mathbf{u}$ with respect to the metric A^{-1}. In effect, we bound the growth of the quantity S, which in turn gives the growth of ℓ_2 norm of $\mathbf{u}$ thanks to the following bounds

$$\lambda_{\min}(A)^{1/2} S \leq \|\mathbf{u}\|_2 \leq \lambda_{\max}(A)^{1/2} S$$

where $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ are the smallest and biggest eigenvalues of A respectively. (We regard $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ as constants here since $k \ll d$.) Explicitly writing the time derivative of S gives an expression including higher-order powers of $\mathbf{u}$. Using Hölder type inequalities, we give upper and lower bounds for each such term in terms of S. This effectively reduces the system to one dimension, hence S has the same time complexity as the single index case which completes the proof. □

In qualitative terms, the initial escape time is determined by the lowest-order term of the correlation loss: a bigger lowest-order term resulted in gradients with a smaller norm, hence increasing the time complexity of the gradient flow. Theorem 2.1 gives a sharp characterization of the time complexity (matching upper and lower bounds). The proof is given in Appendix B.1.

Comparison to literature The same time complexity guarantee $\Theta(d^{p^*/2-1})$ is obtained for the single index model by Ben Arous et al. (2021) where the authors provided both the lower and upper bounds. Recently, Arous et al. (2024) obtained the same lower bound on the time complexity $\Omega(d^{p^*/2-1})$ for a multi-index model when the index vectors are orthogonal. Our Theorem 2.1 generalizes this result to index vectors in arbitrary directions and provides the upper bound as well as the lower bound on the time complexity.

Generalizations In principle, our proof technique allows us to study the case $k = d^\gamma$ where $\gamma \in (0, 1)$ and d is large. First, at initialization, the norm of the projection to the subspace is $\|V^T \mathbf{w}(0)\|_2 = \Theta(d^{-(1-\gamma)/2})$ with high probability. If the index vectors are orthogonal, it is possible to generalize our time complexity guarantees by recalculating the definite integrals with the new initial condition. If the index vectors are in arbitrary positions (e.g., sampled uniformly from the Haar measure on the unit sphere), $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ may depend on d and change the time complexity in a non-trivial way. Studying this setting would require tools from random matrix theory and is beyond the scope of this paper.

2.2 Directional Convergence

The infinite-time behavior of the gradient flow depends on the geometry of the index vectors in general. We give global convergence guarantees in the orthogonal case, i.e., $\mathbf{v}_i^T\mathbf{v}_j = \delta_{ij}$ in this Subsection.

Proposition 2.1 (Fixed Points $\leftrightarrow$ Eigenvectors)**.** *Assume that $\sigma^* = h_{p^*}$ and $\mathbf{v}_1, \dots, \mathbf{v}_k$ are orthogonal. The fixed points of the ODE (2) are in one-to-one correspondence with the unit eigenvectors of the p^*-th order tensor $T = \sum_{j=1}^k \mathbf{v}_j^{\otimes p^*}$.*

Proof. Let $\mathbf{w}$ be a unit vector. It is an eigenvector of T if and only if

$$T(\mathbf{w}, \dots, \mathbf{w}, I) = \sum_{j=1}^k (\mathbf{v}_j^T\mathbf{w})^{p^*-1}\mathbf{v}_j = \lambda\mathbf{w}$$

where the first equality holds due to the orthogonality of $\mathbf{v}_j$. On the other hand, $\mathbf{w}$ is a fixed point of the dynamical system (2) when the Euclidean gradient is orthogonal to the tangent subspace of the unit sphere, that is

$$\nabla L(\mathbf{w}) = \lambda\mathbf{w} \quad \Leftrightarrow \quad \sum_{j=1}^k (\mathbf{v}_j^T\mathbf{w})^{p^*-1}\mathbf{v}_j = \lambda\mathbf{w}.$$

Note that the two conditions are the same. □

In words, finding the complete set of fixed points of the ODE (2) is equivalent to finding the complete set of eigenvectors of the orthogonal tensor decomposition problem. Let us observe that $\mathbf{v}_j$'s are eigenvectors of the p^*-th order tensor due to orthogonality. Anandkumar et al. (2014) noted that $\mathbf{w} = \mathbf{v}_j + \mathbf{v}_{j'}$ is also an eigenvector, unlike the matrix case.

Complete set of fixed points Finding the complete set of fixed points is a challenge in non-convex optimization. Fortunately, Robeva (2016, Theorem 2.3) characterized the complete set of eigenvectors of the orthogonal tensor decomposition problem, which are (informally stated)

$$\sum_{j\in S} \xi_j\mathbf{v}_j \quad \text{or,} \quad \mathbf{v}^\perp \ \text{ such that } \mathbf{v}_j^T\mathbf{v}^\perp = 0 \text{ for all } j \in [k]$$

where S is a non-empty subset of $[k]$ and $\xi_j = 1$ if p^* is odd, $\xi_j \in \{\pm 1\}$ if p^* is even. By Proposition 2.1, we get that normalizing those eigenvectors to be unit norm gives the complete set of fixed points of the dynamical system (2). Hence, there are exactly $2^k - 1$ fixed points in the subspace spanned by the index vectors when p^* is odd. We call $\{\xi_1\mathbf{v}_1, \dots, \xi_k\mathbf{v}_k\}$ 'pure' fixed points, in comparison to 'mixed' fixed points which compute the normalized average of more than one index vector.

Stiefel manifold The same correlation loss function is studied for more general multi-index functions by Bietti et al. (2023). However, the authors restrict the domain of the loss to the Stiefel manifold, which forces the network neurons to be orthogonal during training. The complete set of fixed points in their setting corresponds to picking a subset of singular values to be zero. This similarity between the structure of the fixed points suggests that the loss landscape may not change dramatically when restricted to the Stiefel manifold.

Linear component The goal of this paper is to study target activation functions with information exponent $p^* \geq 3$, that is, when the origin is a degenerate saddle point, following a large body of work in literature. Such activation functions require polynomial time complexity in the initial search phase dedicated to finding the index vector directions (search phase). When the target activation function has a linear component—i.e., $p^* = 1$—then $\mathbf{u} = 0$ is not a fixed point, hence time complexity is not relevant. For $\sigma = \sigma^*$, Arjevani and Field (2021); Şimşek et al. (2023) characterized families of fixed points for ReLU and erf activation functions respectively for the MSE loss. The fixed point structure of the correlation loss when $p^* = 1$ remains an interesting open question.

Among the fixed points, it is unclear a priori which one the gradient flow converges to.

Proposition 2.2 (Directional Convergence)**.** *Let $\ell \in [k]$ and $I = [k] \setminus [\ell]$. Assume that $\mathbf{u}_j(0) > 0$ for all $j \in [k]$ and wlog $\mathbf{u}_1(0) = \cdots = \mathbf{u}_\ell(0) > \max_{j\in I} \mathbf{u}_j(0)$. For any σ^* with information exponent $p^* \geq 3$, the dynamics converge to*

$$\lim_{t\to\infty} \mathbf{u}_j(t) = \frac{1}{\sqrt{\ell}} \quad \text{for} \quad j \in [\ell].$$

If σ^* is even and with information exponent $p^* \geq 4$, assume that wlog $|\mathbf{u}_1(0)| = \cdots = |\mathbf{u}_\ell(0)| > \max_{j\in I} |\mathbf{u}_j(0)|$, the dynamics converge to

$$\lim_{t\to\infty} \mathbf{u}_j(t) = \frac{\operatorname{sgn}(\mathbf{u}_j(0))}{\sqrt{\ell}} \quad \text{for} \quad j \in [\ell].$$

This implies $\lim_{t\to\infty} \mathbf{u}_j(t) = 0$ for $j \in I$ for both cases.

A random initialization breaks ties between any pair of dot products almost surely hence the single neuron converges to the nearest index vector. As a byproduct, Proposition 2.2 shows that the pure and mixed fixed points described for the tensor decomposition problem are fixed points for the dynamical system even when σ^* is not restricted to have a single Hermite component.

Proof sketch. In a nutshell, for the general (non-even) σ^*, we identify the following Lyapunov functions

$$\Delta_{1j}(t) = \mathbf{w}(t)^T\mathbf{v}_1 - \mathbf{w}(t)^T\mathbf{v}_j$$

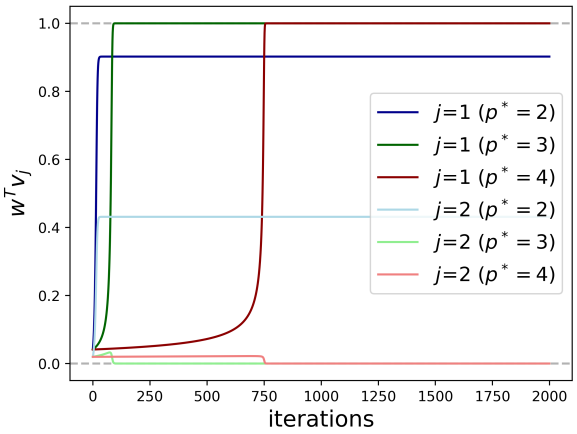


Figure 1: *Dot products during training;* $n = 1, d = 1000$. We run gradient descent by updating the unit norm vector with the spherical gradient using a learning rate $\eta = 0.1$ and normalizing the vector after the update. The target activation is $\sigma^* = h_{p^*}$ and $k = 2$. For $p^* \in \{3, 4\}$, the unit vector (neuron) converges in the direction of the nearest index vector at initialization (here, that is $j = 1$). For $p^* = 2$, a linear combination of the two directions is learned instead. The maximum dot product reaches a non-vanishing value in a longer timescale when the information exponent is bigger.

where $j \neq 1$ and $\mathbf{u}_1(0) = \max_j \mathbf{u}_j(0)$ wlog and show

$$\frac{d}{dt}\Delta_{1j}(t) \geq 0 \quad \text{given} \quad \Delta_{1j}(0) > 0 \tag{4}$$

using some inequalities. This invariance 4 implies that $\mathbf{u}_1(t) > \mathbf{u}_j(t)$ for all times. The only fixed point satisfying this condition is $\mathbf{w} = \mathbf{v}_1$ hence we conclude by the stable manifold theorem. Convergence to the mixed fixed points are obtained similarly due to $\frac{d}{dt}\Delta_{1j}(t) = 0$ if $\Delta_{1j}(0) = 0$. □

The full proof is given in the Appendix B.2.

Note 2.1 ($p^* = 2$). *Proposition 2.2 also applies when* $p^* = 2$ *as long as there is another Hermite polynomial in the expansion without a non-zero coefficient, i.e.,* $a_p b_p > 0$ *for some* $p > 2$. *The special case* $\sigma^*(x) = h_2(x)$ *gives the loss function* $L_0(\mathbf{u}) = C - a_2 b_2 \|\mathbf{u}\|_2^2$ *which is rotationally invariant and minimized for any* $\mathbf{u}$ *with unit norm. This implies that every unit vector that is in the subspace of the index vectors, i.e.,* $\mathbf{u}_1\mathbf{v}_1 + \cdots + \mathbf{u}_k\mathbf{v}_k$ *with* $\|\mathbf{u}\| = 1$ *is a global minimum.*

Tensor decomposition Some remarks are in order to compare the behavior of gradient flow to tensor problems. Recall that the p^*-th order tensor is given by

$$T = \sum_{j=1}^{k} \mathbf{v}_j^{\otimes p^*}.$$

For the correspondence to tensor problems, we should consider the case $\sigma^* = h_{p^*}$. Anandkumar et al. (2014, Theorem 4.1) showed that the tensor power iteration algorithm (which is also local search) converges to the set $\{\mathbf{v}_1, \ldots, \mathbf{v}_k\}$ almost surely, escaping the other eigenvectors corresponding to partial averages of the index vectors. Hence, the tensor power iteration algorithm succeeds in finding the unique decomposition of the orthogonal tensor. More recently, Arous et al. (2024) showed that gradient flow matches neurons to the index vectors of the tensor PCA according to a greedy selection mechanism when the domain of the correlation loss is a Stiefel manifold. Our Proposition 2.2 applies when the activation function is an arbitrary mix of Hermite polynomials (but not restricted to be a polynomial itself), and when the neurons move independently. Decoupling of the neurons allows us to study whether the random initialization gives a good enough coverage to match neurons to the index vectors in the multi-neuron case (see Subsection 2.3).

2.3 Gradient Flow Learns with Mild Overparameterization

We now turn to the study of neural networks with an arbitrary number of neurons. Each neuron of the neural network evolves following the dynamical system (2) and does not interact with other neurons in our model. At initialization, we would expect half of the correlations to be negative due to the rotational symmetry of the initialization and orthogonality of the index vectors. As a first step, we give the information of the direction to the learner by flipping the sign of the correlation if it is negative.

What remains is to ensure at least one neuron is allocated to each one of the index vectors. We use Proposition 2.2 and a classic matching argument to make sure overparameterization works. It suffices to ensure that all of the k directions $\mathbf{v}_1, \ldots, \mathbf{v}_k$ are "collected" (in the coupon-collecting sense) by the n student neurons at initialization. Say $\mathbf{v}_j$ is *collected* by $\mathbf{w}_i$ (at initialization) if

$$\mathbf{w}_i(0)^T \mathbf{v}_j > \max_{j' \neq j} \mathbf{w}_i(0)^T \mathbf{v}_{j'}.$$

By Proposition 2.2, if $\mathbf{v}_j$ is collected by $\mathbf{w}_i$ at initialization, then $\mathbf{w}_i(t) \to \mathbf{v}_j$ as $t \to \infty$. By symmetry of the random initialization, for any $i \in [n]$ and $j \in [k]$,

$$P_{\text{collect}} := \Pr[\mathbf{v}_j \text{ is collected by } \mathbf{w}_i] = 1/k.$$

Therefore, by independence of the initialization and a union bound,

$$\Pr[\exists j \in [k] \text{ s.t. } \mathbf{v}_j \text{ is not collected by any } \{\mathbf{w}_i\}_{i\in[n]}] \leq k(1-P_{\text{collect}})^n \leq k\exp(-P_{\text{collect}}n). \quad (5)$$

This failure probability bound is less than $1/k^\epsilon$ when $n \geq (1+\epsilon)k\ln k$ for any $\epsilon > 0$. Hence, a mild overparameterization of $\log(k)$ factor is sufficient for matching the neurons to the index vectors perfectly when the index vectors are orthogonal to each other.

We can also easily get the following lower-bound

$$\Pr[\exists j \in [k] \text{ s.t. } \mathbf{v}_j \text{ is not collected by any } \{\mathbf{w}_i\}_{i\in[n]}] \geq (1-P_{\text{collect}})^n \to \exp(-\gamma) \text{ as } k\to\infty,\ \gamma = n/k. \quad (6)$$

The failure probability is more than $\exp(-\gamma)$ when $n = \gamma k$ for any constant γ in the proportional limit, hence using the correlation loss does not guarantee matching neurons to the index vectors when using a constant factor of overparameterization. Using only a constant factor of overparameterization may be sufficient when using the MSE loss (discussed informally in Section 4).

3 Saddle-to-Minimum Transition

In this section, we study the effect of index vectors moving from an orthogonal frame towards each other. For analytic tractability, we consider the following dot product matrix

$$A = \begin{bmatrix} 1 & \beta & \dots & \beta \\ \beta & 1 & \dots & \beta \\ \vdots & \vdots & \ddots & \vdots \\ \beta & \beta & \dots & 1 \end{bmatrix} \quad (7)$$

where $\beta \in [0,1]$. In other words, the index vectors form an equiangular set. The matrix A is invertible for $\beta < 1$ (see the eigenvalue calculations in the Appendix A). $\beta = 1$ implies $\mathbf{v}_1 = \cdots = \mathbf{v}_k$, that is, the single-index model. Hence A is not necessarily invertible in this Section. Specifically, we focus on the local geometry of the point that computes the average of the index vectors

$$\bar{\mathbf{w}} = \frac{1}{\|\sum_{j=1}^k \mathbf{v}_j\|_2}\sum_{j=1}^k \mathbf{v}_j,$$

which we call the 'average' fixed point.

Lemma 3.1. *$\bar{\mathbf{w}}$ is a fixed point of the dynamical system (2) under the symmetric dot product structure (7).*

Proof. $\mathbf{w}$ is a fixed point if and only if $\nabla L(\mathbf{w}) = \lambda \mathbf{w}$. Let us write the Euclidean gradient explicitly using the expression (3):

$$\nabla L(\mathbf{w}) = -\sum_{p\geq p^*} a_p b_p \sum_{j=1}^k (\mathbf{w}^T\mathbf{v}_j)^{p-1}\mathbf{v}_j,$$

and plug in $\bar{\mathbf{w}}$:

$$\nabla L(\mathbf{w})\big|_{\mathbf{w}=\bar{\mathbf{w}}} = -\sum_{p\geq p^*} a_p b_p \sum_{j=1}^k (\bar{\mathbf{w}}^T\mathbf{v}_j)^{p-1}\mathbf{v}_j.$$

Observe that $\bar{\mathbf{w}}^T\mathbf{v}_j$ is the same for all j due to the dot product structure (7). Hence we can push the constant term $(\bar{\mathbf{w}}^T\mathbf{v}_j)^{p-1}$ outside of the summation and conclude that $\nabla L(\bar{\mathbf{w}})$ is parallel to $\bar{\mathbf{w}}$. □

Intuitively, the correlation between the predictor $f(\mathbf{x}) = \sigma(\bar{\mathbf{w}}^T\mathbf{x})$ and the multi-index function increases as the index vectors approach each other, as $\bar{\mathbf{w}}$ would get closer to the index vectors. To formalize this, we need to make β explicit in the loss function. Indeed, the evaluation of the loss at $\bar{\mathbf{w}}$ is given by the two instances (i) orthogonal, $\beta = 0$, (ii) single-index, $\beta = 1$ as follows

$$L_\beta(\bar{\mathbf{w}})\big|_{\beta=0} = C - \sum_{p\geq p^*} c_p\Big(\frac{1}{k}\Big)^{p/2}k, \quad L_\beta(\bar{\mathbf{w}})\big|_{\beta=1} = C - \sum_{p\geq p^*} c_p k$$

where $L_\beta(\bar{\mathbf{w}})$ is $L(\bar{\mathbf{w}})$ under the dot product structure (7). Indeed $L_\beta(\bar{\mathbf{w}})\big|_{\beta=0} > L_\beta(\bar{\mathbf{w}})\big|_{\beta=1}$, and we show that $L_\beta(\bar{\mathbf{w}})$ is decreasing as β increases (see Appendix for the calculations).

In the proportional limit $k = cd$ with $c \in (0,1)$ and as $d\to\infty$, $L_\beta(\bar{\mathbf{w}})$ approaches C, hence $\bar{\mathbf{w}}$ is as bad as not learning anything. Whereas for $\beta = 1$, we have $\bar{\mathbf{w}} = \mathbf{v}_1 = \cdots = \mathbf{v}_k$, all fixed points collapse on each other, and $\bar{\mathbf{w}}$ is the optimal. It is important to characterize whether $\bar{\mathbf{w}}$ is a strict saddle or a local minimum. To do so, one needs to study how the curvature of $L_\beta(\bar{\mathbf{w}})$ changes as β increases. In the next Theorem, we sharply characterize the sign change in the curvature for polynomial target activation σ^* of degree P and inf. exponent p^*.

Theorem 3.2 (Saddle-to-Minimum)**.** *Assume $d > k$. $\bar{\mathbf{w}}$ is a strict saddle when the dot product is upper bounded by*

$$\beta < \frac{p^*-2}{k+p^*-2},$$

whereas $\bar{\mathbf{w}}$ is a local minimum when the dot product is lower bounded by

$$\frac{P-2}{k+P-2} < \beta.$$

Therefore, for the case $\sigma^* = h_{p^*}$

$$\beta_c = \frac{p^*-2}{k+p^*-2}$$

is the sharp threshold characterizing the transition from a strict saddle to a local minimum.

Muller et al. (2022) studied the tensor decomposition problem when the component vectors form an equiangular set, focusing on the regime when the component vectors are robust eigenvectors under the tensor power iteration algorithm. Our Theorem 3.2 shows that the average fixed point turns into a local minimum when β exceeds a certain threshold. This suggests that there may be another threshold β_f such that the average fixed point is attractive for gradient descent with a small learning rate (see Fig (4)). Our result indicates a hard regime for tensor decomposition as well as neural networks–which was not studied before to our knowledge.

4 Simulations

We give simulations in three scenarios:

1. orthogonal index vectors, compare the behavior of correlation loss and MSE loss, in terms of the benefit of MSE loss in decreasing the failure probability for insufficiently overparameterized neural networks or unlucky initializations (Fig (2)),
2. orthogonal index vectors, compare the behavior of correlation loss and MSE loss, in terms of the loss curves and visual inspection of gradient flow trajectories for different activation functions (Fig (3)),
3. non-orthogonal index vectors for $k = 2$, study the effect of increasing the dot product β in the behavior of gradient flow (Fig (4)).

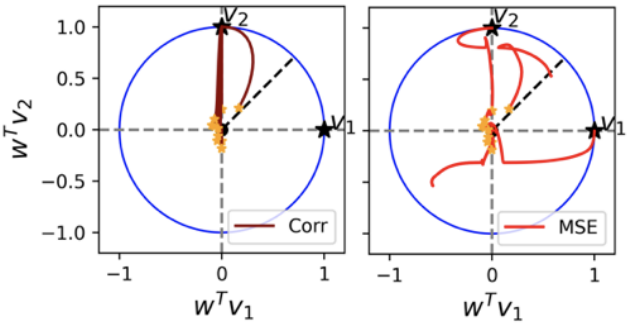
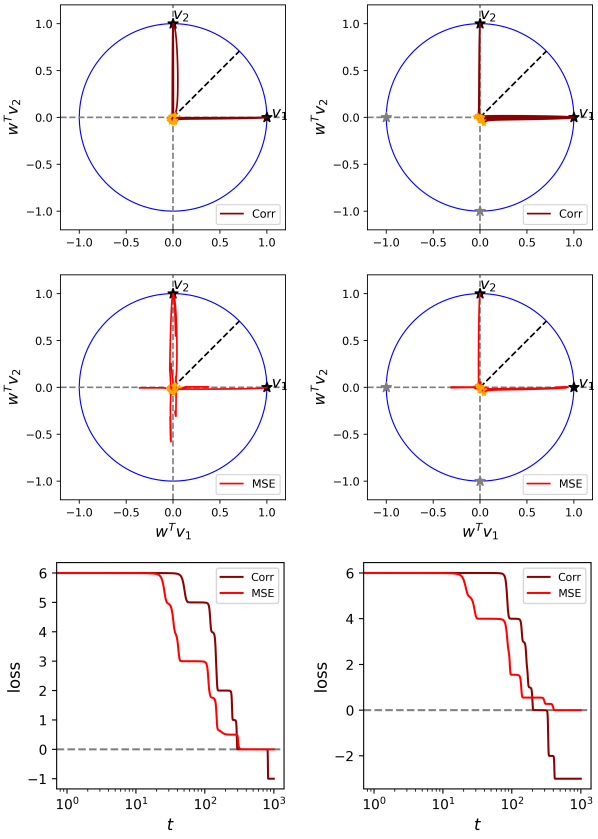


Figure 2: *MSE loss helps with neuron allocation; fixed initialization in both figures,* $k=2$, $n=10$. If no neuron at initialization is closest to one of the index vectors, gradient flow fails to find it when using the correlation loss (*left panel*) whereas the MSE loss fixes this issue thanks to the repulsion between neurons (*right panel*).

In particular, using a factor $\gamma = 4$ of overparameterization is recommendable to get a very small lower bound, that is $\exp(-4) \approx 0.0183$, on the probability of the failing initialization for the correlation loss. Curiously, this recommendation is consistent with the numerical simulations of Martinelli et al. (2023) for learning two-layer neural networks with biases using gradient flow under the MSE loss, for a large family of toy problems.



(a) $\sigma^* = h_3$ (b) $\sigma^* = h_4$

Figure 3: *Gradient flow trajectories projected to the subspace of orthogonal index vectors for MSE and correlation losses and loss curves; odd activation (a), even activation (b).* Initialization and the number of neurons are fixed; $k = 2$, $n = 10$, $d = 1000$. The winning neurons move toward the closest index vectors also for the MSE loss, but the other neurons move non-trivially due to interactions between them. Adding the repulsion term (MSE loss) virtually decreases the time complexity (bottom row), however, the improvement in the time complexity may be only up to a constant factor.

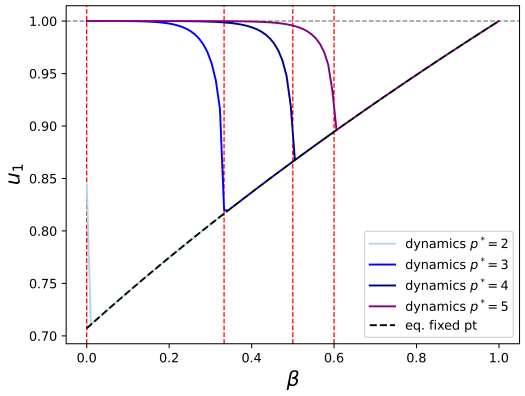


Figure 4: *Non-orthogonal index vectors with dot product structure* (7)*, bifurcation diagram;* $k = 2$*,* $\sigma^* = h_{p^*}$*.* The infinite-time behavior of the student vector abruptly changes from monotonic convergence to the nearest direction to convergence to the average of directions (non-monotonically) at a critical value $\beta_f \in (0, 1)$. The red dashed line indicates the saddle-to-minimum threshold β_c given in Theorem 3.2. Observe the small gap between β_c and β_f.

Acknowledgements

BS appreciates many discussions with Loucas Pillaud-Vivien and Alberto Bietti in the early phases of this work, and the valuable feedback from Denny Wu and Ludovic Stephan on the manuscript. DH acknowledges support from the ONR under grant N00014-24-1-2700.

References

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

Sebastian Goldt, Madhu Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. *Advances in neural information processing systems*, 32, 2019.

Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

Alberto Bietti, Joan Bruna, Clayton Sanford, and Min Jae Song. Learning single-index models with shallow neural networks. *Advances in Neural Information Processing Systems*, 35:9768–9783, 2022.

Boaz Barak, Benjamin Edelman, Surbhi Goel, Sham Kakade, Eran Malach, and Cyril Zhang. Hidden progress in deep learning: Sgd learns parities near the computational limit. *Advances in Neural Information Processing Systems*, 35:21750–21764, 2022.

Alexandru Damian, Jason Lee, and Mahdi Soltanolkotabi. Neural networks can learn representations with gradient descent. In *Conference on Learning Theory*, pages 5413–5452. PMLR, 2022.

Emmanuel Abbe, Enric Boix Adsera, and Theodor Misiakiewicz. Sgd learning on neural networks: leap complexity and saddle-to-saddle dynamics. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 2552–2623. PMLR, 2023.

Raphaël Berthier, Andrea Montanari, and Kangjie Zhou. Learning time-scales in two-layers neural networks. *arXiv preprint arXiv:2303.00055*, 2023.

Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

Margalit Glasgow. Sgd finds then tunes features in two-layer neural networks with near-optimal sample complexity: A case study in the xor problem. *arXiv preprint arXiv:2309.15111*, 2023.

Kazusato Oko, Yujin Song, Taiji Suzuki, and Denny Wu. Learning sum of diverse features: computational hardness and efficient gradient-based training for ridge combinations. *arXiv preprint arXiv:2406.11828*, 2024.

Gérard Ben Arous, Cédric Gerbelot, and Vanessa Piccolo. High-dimensional optimization for multi-

spiked tensor pca. *arXiv preprint arXiv:2408.06401*, 2024.

Alex Damian, Eshaan Nichani, Rong Ge, and Jason D Lee. Smoothing the landscape boosts the signal for sgd: Optimal sample complexity for learning single index models. *arXiv preprint arXiv:2305.10633*, 2023.

Yatin Dandi, Emanuele Troiani, Luca Arnaboldi, Luca Pesce, Lenka Zdeborová, and Florent Krzakala. The benefits of reusing batches for gradient descent in two-layer networks: Breaking the curse of information and leap exponents. *arXiv preprint arXiv:2402.03220*, 2024.

Jason D Lee, Kazusato Oko, Taiji Suzuki, and Denny Wu. Neural network learns low-dimensional polynomials with sgd near the information-theoretic limit. *arXiv preprint arXiv:2406.01581*, 2024.

Luca Arnaboldi, Yatin Dandi, Florent Krzakala, Luca Pesce, and Ludovic Stephan. Repetita iuvant: Data repetition allows sgd to learn high-dimensional multi-index functions. *arXiv preprint arXiv:2405.15459*, 2024.

Itay Safran and Ohad Shamir. Spurious local minima are common in two-layer relu neural networks. In *International conference on machine learning*, pages 4433–4441. PMLR, 2018.

Yossi Arjevani and Michael Field. Analytic study of families of spurious minima in two-layer relu neural networks: a tale of symmetry ii. *Advances in Neural Information Processing Systems*, 34:15162–15174, 2021.

Berfin Şimşek, Amire Bendjeddou, Wulfram Gerstner, and Johanni Brea. Should under-parameterized student networks copy or average teacher weights? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Berfin Şimşek, François Ged, Arthur Jacot, Francesco Spadaro, Clément Hongler, Wulfram Gerstner, and Johanni Brea. Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. In *International Conference on Machine Learning*, pages 9722–9732. PMLR, 2021.

Weihang Xu and Simon Du. Over-parameterization exponentially slows down gradient descent for learning a single neuron. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1155–1198. PMLR, 2023.

Simon Martin, Francis Bach, and Giulio Biroli. On the impact of overparameterization on the training of a shallow neural network in high dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 3655–3663. PMLR, 2024.

Marco Mondelli and Andrea Montanari. On the connection between learning two-layer neural networks and tensor decomposition. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1051–1060. PMLR, 2019.

Gerard Ben Arous, Reza Gheissari, and Aukosh Jagannath. Online stochastic gradient descent on non-convex losses from high-dimensional inference. *The Journal of Machine Learning Research*, 22(1):4788–4838, 2021.

Christopher J Hillar and Lek-Heng Lim. Most tensor problems are np-hard. *Journal of the ACM (JACM)*, 60(6):1–39, 2013.

Amit Daniely, Roy Frostig, and Yoram Singer. Toward deeper understanding of neural networks: The power of initialization and a dual view on expressivity. *Advances in neural information processing systems*, 29, 2016.

Ryan O'Donnell. Analysis of boolean functions. *arXiv preprint arXiv:2105.10386*, 2021.

Rishabh Dudeja and Daniel Hsu. Learning single-index models in gaussian space. In *Conference On Learning Theory*, pages 1887–1930. PMLR, 2018.

Animashree Anandkumar, Rong Ge, Daniel J Hsu, Sham M Kakade, Matus Telgarsky, et al. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832, 2014.

Elina Robeva. Orthogonal decomposition of symmetric tensors. *SIAM Journal on Matrix Analysis and Applications*, 37(1):86–102, 2016.

Tommi Muller, Elina Robeva, and Konstantin Usevich. Robust eigenvectors of symmetric tensors. *SIAM Journal on Matrix Analysis and Applications*, 43(4):1784–1805, 2022.

Flavio Martinelli, Berfin Simsek, Wulfram Gerstner, and Johanni Brea. Expand-and-cluster: Parameter recovery of neural networks. *arXiv preprint arXiv:2304.12794*, 2023.

A Reparameterization of the Loss

Recall that the correlation loss expands in terms of the neurons $\mathbf{w}$ and index vectors $\mathbf{v}_j$ as

$$L(\mathbf{w}) = C - \sum_{j=1}^{k} \mathbb{E}[\sigma(\mathbf{w}^T\mathbf{x})\sigma^*(\mathbf{v}_j^T\mathbf{x})].$$

Due to the rotational symmetry of the standard Gaussian distribution, each term can be expressed as

$$\mathbb{E}[\sigma(\mathbf{w}^T\mathbf{x})\sigma^*(\mathbf{v}_j^T\mathbf{x})] = \mathbb{E}_{(x,y)\sim\mathcal{N}(0,C(u))}[\sigma(x)\sigma^*(y)]$$

where $C(u)$ is a 2×2 covariance matrix with entries $C(u)_{11}=1, C(u)_{22}=1, C(u)_{12}=u, C(u)_{21}=u$ where $u=\mathbf{w}^T\mathbf{v}_j$. Hence one can express the above expectation using the variable u only, and the dual activation is well-defined. This observation also allows us to express the loss in terms of the dot products $\mathbf{w}^T\mathbf{v}_j$ hence

$$L_0(V^T\mathbf{w}) = L(\mathbf{w}). \tag{8}$$

We will use this expression to derive Lemma 1.1 of the main which is restated below.

Lemma A.1. *Assume that $\mathbf{w}(t)$ solves the ODE (2) given an initial condition $\mathbf{w}=\mathbf{w}_0$. Then the vector of dot products $\mathbf{u}(t)=V^T\mathbf{w}(t)$ solves the following ODE*

$$\frac{d}{dt}\mathbf{u} = -(A-\mathbf{u}\mathbf{u}^T)\nabla L_0(\mathbf{u})$$

with an initial condition $\mathbf{u}_0 = V^T\mathbf{w}_0$.

Proof. The chain rule allows us to develop the time derivative as follows

$$\frac{d}{dt}V^T\mathbf{w} = V^T\frac{d}{dt}\mathbf{w} = -V^T(I_d-\mathbf{w}\mathbf{w}^T)\nabla L(\mathbf{w}).$$

The Euclidean gradient of L can be expressed as follows using Eq. (8)

$$\nabla L(\mathbf{w}) = V\nabla L_0(V^T\mathbf{w}).$$

Plugging in the Euclidean gradient of L, we get

$$\begin{aligned}\frac{d}{dt}V^T\mathbf{w} &= -V^T(I_d-\mathbf{w}\mathbf{w}^T)V\nabla L_0(V^T\mathbf{w})\\ &= -(V^TV-(V^T\mathbf{w})(V^T\mathbf{w})^T)\nabla L_0(V^T\mathbf{w})\end{aligned}$$

which completes the derivation. Substituting $\mathbf{u}=V^T\mathbf{w}$ and $A=V^TV$ gives the expression stated in the Lemma. □

Derivation of the Boundary In order to explicitly describe the domain of L_0, we need to understand how the domain of $L(\mathbf{w})$, i.e., the unit sphere, changes under the linear projection $V^T:\mathbb{R}^d\to\mathbb{R}^k$. This is achieved below, where we characterize the image of the linear projection $V^T:\mathbb{S}^d\to D$ when restricted to the unit sphere.

Let us express $\mathbf{w}$ as a linear combination of $\mathbf{v}_j$'s and an orthogonal component to the span of $\mathbf{v}_j$'s

$$\mathbf{w} = \sum_{j=1}^{k}\boldsymbol{\alpha}_j\mathbf{v}_j + \mathbf{v}^\perp.$$

Since $\mathbf{w}$ is on the unit sphere, the $\boldsymbol{\alpha}_j$'s should satisfy the following constraint

$$\|\mathbf{w}\|_2^2 = \boldsymbol{\alpha}^T(V^TV)\boldsymbol{\alpha} + \|\mathbf{v}^\perp\|_2^2 = 1 \quad\Rightarrow\quad \boldsymbol{\alpha}^T(V^TV)\boldsymbol{\alpha}\le 1.$$

The dot products are given by $\mathbf{u}_j = \mathbf{w}^T\mathbf{v}_j$. Plugging in the expansion of $\mathbf{w}$ in the basis of $\mathbf{v}_j$, we get

$$\mathbf{u}_j = \sum_{j'=1}^{k} \boldsymbol{\alpha}_{j'}\mathbf{v}_j^T\mathbf{v}_{j'} \quad \Leftrightarrow \quad \boldsymbol{\alpha} = (V^TV)^{-1}\mathbf{u}.$$

Using the constraint on $\boldsymbol{\alpha}$ allows us to derive the constraint on $\mathbf{u}$ using the above identity as follows

$$\boldsymbol{\alpha}^T(V^TV)\boldsymbol{\alpha} \le 1 \quad \Leftrightarrow \quad \mathbf{u}^T(V^TV)^{-1}(V^TV)(V^TV)^{-1}\mathbf{u} \le 1 \quad \Leftrightarrow \quad \mathbf{u}^T(V^TV)^{-1}\mathbf{u} \le 1.$$

Hence the space of dot products is an ellipsoidal ball $D = \{\mathbf{u} : \mathbf{u}^TA^{-1}\mathbf{u} \le 1\}$ where the shape is given by A^{-1}.

Eigenvalue and Eigenvector analysis Let us denote with $\mathbf{1}_k$ and $\mathbf{I}_k$ the all-ones vector and the identity matrix of size k. We want to determine the eigenvalues and eigenvectors of the matrix

$$A = \beta\mathbf{1}_k\mathbf{1}_k^T + (1-\beta)\mathbf{I}_k.$$

First, note that A and $\mathbf{1}_k\mathbf{1}_k^T$ have the same eigenvectors. If $\mathbf{v}$ is an eigenvector of $\mathbf{1}_k\mathbf{1}_k^T$ with eigenvalue λ, then

$$\begin{aligned} A\mathbf{v} &= \beta\mathbf{1}_k\mathbf{1}_k^T\mathbf{v} + (1-\beta)\mathbf{I}_k\mathbf{v} \\ &= \beta\lambda\mathbf{v} + (1-\beta)\mathbf{v} = (\beta\lambda + (1-\beta))\mathbf{v}, \end{aligned}$$

and $\mathbf{v}$ is an eigenvector of A with eigenvalue $(\beta\lambda + (1-\beta))$. Since $\mathbf{1}_k\mathbf{1}_k^T$ has rank one, it has exactly one non-zero eigenvalue given by $\|\mathbf{1}_k\|^2 = k$ with eigenspace $\mathrm{span}(\mathbf{1}_k)$ and $k-1$ zero eigenvalues with eigenspace $\mathrm{span}(\mathbf{1}_k)^\perp = \mathrm{span}\{\mathbf{e}_1 - \mathbf{e}_2, \mathbf{e}_1 - \mathbf{e}_3, \dots, \mathbf{e}_1 - \mathbf{e}_k\}$, where $\mathbf{e}_i$ denotes the ith standard unit vector. This implies that A has eigenvalues $1+\beta(k-1)$ and $(1-\beta)$ with eigenspaces $\mathrm{span}(\mathbf{1}_k)$ and $\mathrm{span}\{\mathbf{e}_1-\mathbf{e}_2, \mathbf{e}_1-\mathbf{e}_3, \dots, \mathbf{e}_1-\mathbf{e}_k\}$ respectively. In particular, this shows that A is invertible for $0 \le \beta < 1$.

For subsequent calculations, we need to relate the eigenvalues and eigenvectors of $A = V^TV$ to the eigenvalues and eigenvectors of VV^T. This is achieved in the following Lemma.

Lemma A.2. *The matrices V^TV and VV^T satisfy the following properties.*

(i) V^TV and VV^T have the same nonzero eigenvalues.

(ii) Every eigenvector $\mathbf{v}$ of VV^T with nonzero eigenvalue $\lambda_{\mathbf{v}}$ satisfies $\mathbf{v} = V\mathbf{w}$, where $\mathbf{w}$ is an eigenvector of V^TV with the same eigenvalue.

Proof. Both assertions follow from the following observation. Let $\mathbf{w}$ be an eigenvector of V^TV with eigenvalue $\lambda_{\mathbf{w}} \neq 0$, then

$$V^TV\mathbf{w} = \lambda_{\mathbf{w}}\mathbf{w} \implies V\Big(V^TV\Big)\mathbf{w} = \lambda_{\mathbf{w}}V\mathbf{w} \implies \Big(VV^T\Big)V\mathbf{w} = \lambda_{\mathbf{w}}V\mathbf{w},$$

and $V\mathbf{w} \neq 0$, since $\lambda_{\mathbf{w}} \neq 0$. In other words, $V\mathbf{w}$ is an eigenvector of VV^T with the same nonzero eigenvalue $\lambda_{\mathbf{w}}$. Repeating the argument with VV^T instead of V^TV also shows that $V^T\mathbf{w}$ is an eigenvector of V^TV with eigenvalue $\lambda_{\mathbf{w}}$, whenever $\mathbf{w}$ is an eigenvector of VV^T with $\lambda_{\mathbf{w}} \neq 0$. □

B Learning Index Vectors

The low dimensional dynamical system is expressed in terms of the Euclidean gradient of the loss in Lemma A.1. We will use the concrete version where each coordinate of $\mathbf{u}$ is made explicit in some proofs in the following Subsections. Let us make the expression for the Euclidean gradient of L_0 explicit using Eq. (3)

$$\nabla L_0(\mathbf{u}) = -\sum_{p \ge p^*} c_p p\ \mathbf{u}^{\odot p-1} \quad \text{where} \quad c_p = a_pb_p.$$

The time derivative of $\mathbf{u}$ is expressed as a matrix vector product in Lemma A.1. Carrying out the matrix vector product explicitly, one gets the following expression for the ODE in terms of the Hermite coefficients

$$\frac{d}{dt}\mathbf{u}_j = \sum_{p \ge p^*} c_p p\big(\mathbf{u}_j^{p-1} + \sum_{j' \neq j} A_{jj'}\mathbf{u}_{j'}^{p-1} - \mathbf{u}_j\sum_{j'}\mathbf{u}_{j'}^p\big). \qquad \text{(general index vectors)} \tag{9}$$

Note that $a_p b_p \geq 0$ by Assumption 1.1 and $A_{jj'} = \mathbf{v}_j^T \mathbf{v}_{j'} \geq 0$ by Assumption 1.3. We analyze the dynamical system in the positive quadrant, i.e., $\mathbf{u}_j \geq 0$ for all $j \in [k]$. Hence, we can give the following interpretation for the low dimensional ODE (9) in terms of the three terms: (i) a self-reinforcing term promoting the self-growth of $\mathbf{u}_j$, (ii) the reinforcing term promoting the growth of $\mathbf{u}_j$ given the other dot products, and (iii) a damping term decreasing the growth proportional to $\mathbf{u}_j$ and the ℓ_p-norm of the vector of dot products.

Trajectories remain in the positive quadrant For the above interpretation to be valid, we need to show that the trajectories are confined to the positive quadrant. More formally, that is, the set $S = [0,1]^k \cap D$ is invariant. To show that trajectories do not leave the set S, it suffices to analyze the boundaries. When $\mathbf{u}_j = 0$ and $\mathbf{u}_{j'} \geq 0$ for $j' \geq j$, we have the following

$$\frac{d}{dt}\mathbf{u}_j = \sum_{p \geq p^*} c_p p \sum_{j' \neq j} A_{jj'} \mathbf{u}_{j'}^{p-1} \geq 0, \tag{10}$$

due to Assumption 1.1 and Assumption 1.3. Hence, $\mathbf{u}_j$ does not decrease and become negative; in other words, the trajectories do not leave the set S.

The assumption on the positivity of the dot products at initialization In most of our analysis, we assume that the dot products are positive at initialization. With random initialization, half of the dot products would be positive on average. However, negative dot products at initialization require a more refined analysis than we present here (except for the even activation case, where the proofs can be extended to negative dot products using a sign-flipping argument). A reasonable guess is that as long as one of the dot products at initialization is positive, the single neuron dynamics exhibits the same time complexity and converges to the nearest index vector. However, the proof needs refinements for the relaxed assumption on the initialization to be sufficient (i.e., when only one dot product is positive) which is beyond the scope of this work.

B.1 Time complexity

Theorem B.1 (Time complexity)**.** *Assume that d is large, $\mathbf{u}_j(0) = \Theta(d^{-1/2})$ and $\mathbf{u}_j(0) > 0$ for all $j \in [k]$. For $p^* \geq 2$, we show the following time complexities*

- *if $p^* \geq 3$, $T = \Theta(d^{p^*/2-1})$ is necessary and sufficient to reach $\|\mathbf{u}(T)\|_2 = \Theta(1)$,*
- *if $p^* = 2$, $T = \Theta(\log(d))$ is necessary and sufficient to reach $\|\mathbf{u}(T)\|_2 = \Theta(1)$.*

Proof. Since the dot products are assumed to be positive at initialization, they remain non-negative at all times (see Eq. 10 and the comments therein). That is, we have $\mathbf{u}_j(t) \geq 0$ for all $j \in [k]$ and for all $t \geq 0$.

Let us define $s_2 = \mathbf{u}^T A^{-1} \mathbf{u}$, the square of the norm induced by A^{-1}. The time derivative of s_2 is given by

$$\begin{aligned}
\frac{1}{2}\frac{d}{dt}s_2 &= \Big(\sum_{p \geq p^*} c_p p (A - \mathbf{u}\mathbf{u}^T)\mathbf{u}^{\odot p-1}\Big)^T A^{-1}\mathbf{u} && (11)\\
&= \sum_{p \geq p^*} c_p p \Big((\mathbf{u}^{\odot p-1})^T \Big(A^T - (\mathbf{u}\mathbf{u}^T)^T\Big) A^{-1}\mathbf{u}\Big)\\
&= \sum_{p \geq p^*} c_p p \Big((\mathbf{u}^{\odot p-1})^T \Big(A - \mathbf{u}\mathbf{u}^T\Big) A^{-1}\mathbf{u}\Big)\\
&= \sum_{p \geq p^*} c_p p \Big((\mathbf{u}^{\odot p-1})^T \mathbf{u}(1 - \mathbf{u}^T A^{-1}\mathbf{u})\Big)\\
&= \sum_{p \geq p^*} c_p p \Big((\mathbf{u}^{\odot p-1})^T \mathbf{u}(1 - s_2)\Big), && (12)
\end{aligned}$$

where in equation (11) we used that

$$\frac{1}{2}\frac{d}{dt}(\mathbf{u}^T A^{-1}\mathbf{u}) = (\frac{d}{dt}\mathbf{u})^T A^{-1}\mathbf{u}.$$

Since the correlations are positive at all times, we have $\|\mathbf{u}\|_p^p = (\mathbf{u}^{\odot p-1})^T\mathbf{u}$ at all times. Furthermore, since we initialize $\mathbf{u}$ inside the ellipsoidal ball $D = \{\mathbf{u} : \mathbf{u}^T A^{-1}\mathbf{u} \leq 1\}$ (which comes from the fact that $\mathbf{u} = V^T\mathbf{w}$ for a unit vector $\mathbf{w}$), we have that $1 - s_2 \geq 0$ holds at all times. This implies that s_2 is increasing until we hit the boundary of D. We can control the ℓ_p norm by using an inequality between norms

$$r \leq p \quad \Rightarrow \quad \|\mathbf{u}\|_p \leq \|\mathbf{u}\|_r \leq k^{\frac{1}{r}-\frac{1}{p}}\|\mathbf{u}\|_p,$$

which is a consequence of Hölder's inequality. For $r = 2$, this is equivalent to

$$k^{1-\frac{p}{2}}\|\mathbf{u}\|_2^p \leq \|\mathbf{u}\|_p^p \leq \|\mathbf{u}\|_2^p. \tag{13}$$

Using the following estimate

$$\lambda_{\min}(A)\mathbf{u}^T A^{-1}\mathbf{u} \leq \|\mathbf{u}\|_2^2 \leq \lambda_{\max}(A)\mathbf{u}^T A^{-1}\mathbf{u}, \tag{14}$$

and Hölder's inequality (13), we can bound the ℓ_p norm in terms of s_2

$$k^{1-\frac{p}{2}}\lambda_{\min}(A)^{\frac{p}{2}}(\mathbf{u}^T A^{-1}\mathbf{u})^{\frac{p}{2}} \leq \|\mathbf{u}\|_p^p \leq \lambda_{\max}(A)^{\frac{p}{2}}(\mathbf{u}^T A^{-1}\mathbf{u})^{\frac{p}{2}}.$$

Let us denote $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ with λ_1 and λ_k respectively. For $p^* \geq 2$, the time derivative of s_2 can be bounded as

$$\sum_{p\geq p^*} 2c_p p\lambda_k^{\frac{p}{2}}k^{1-\frac{p}{2}}s_2^{\frac{p}{2}}(1-s_2) \leq \frac{d}{dt}s_2 \leq \sum_{p\geq p^*} 2c_p p\lambda_1^{\frac{p}{2}}s_2^{\frac{p}{2}}(1-s_2)$$
$$\sum_{p\geq p^*} 2C_p\lambda_k^{\frac{p}{2}}s_2^{\frac{p}{2}}(1-s_2) \leq \frac{d}{dt}s_2 \leq \sum_{p\geq p^*} 2\bar{C}_p\lambda_1^{\frac{p}{2}}s_2^{\frac{p}{2}}(1-s_2) \quad \text{where} \quad C_p = c_p p k^{1-\frac{p}{2}}, \bar{C}_p = c_p p.$$

Since the c_p's are all positive, we can drop the higher order terms and get the following lower bound

$$2C_{p^*}\lambda_k^{\frac{p^*}{2}}s_2^{\frac{p^*}{2}}(1-s_2) \leq \frac{d}{dt}s_2 \leq 2\big(\sum_{p\geq p^*}\bar{C}_p\lambda_1^{\frac{p}{2}}\big)s_2^{\frac{p^*}{2}}(1-s_2),$$

where the upper bound comes from observing that s_2 does not exceed one ($s_2 \leq 1$) for all times due to the problem geometry. To simplify, let us study the dynamics for the interval $s_2 \in [b_0, 1/2]$. Up to a change in the constant of the lower bound, we get the following lower and upper bounds

$$C_{p^*}s_2^{\frac{p^*}{2}} \leq \frac{d}{dt}s_2 \leq 2\big(\sum_{p\geq p^*}\bar{C}_p\lambda_1^{\frac{p}{2}}\big)s_2^{\frac{p^*}{2}}.$$

Using separation of variables, we can integrate the sandwiched part

$$\int_{b_0}^{b_1} s_2^{\frac{-p^*}{2}}ds_2$$

which is lower and upper bounded by Ct and $\bar{C}t$ for some constants C and $\bar{C}$. Let us do the integration first for $p^* \geq 3$

$$\frac{s_2^{\frac{-p^*}{2}+1}}{\frac{-p^*}{2}+1}\Big|_{b_0}^{b_1} = \frac{1}{\frac{p^*}{2}-1}\big(-b_1^{\frac{-p^*}{2}+1} + b_0^{\frac{-p^*}{2}+1}\big) = \Theta(d^{\frac{p^*}{2}-1}),$$

since $b_1 = \Theta(1)$ and $b_0 = \Theta(d^{-1})$ (because every correlation is initialized with $\Theta(d^{-1/2})$). It remains to do the integration for $p^* = 2$ which is simply

$$\int_{b_0}^{b_1} s_2^{-1}ds_2 = \log(s_2)\Big|_{b_0}^{b_1} = \log(b_1) - \log(b_0) = \Theta(\log(d)).$$

Finally, the estimate in (14) implies that the time needed for the ℓ_2 norm of $\mathbf{u}$ to reach a non-vanishing value is of the same order as the time needed for s_2. □

B.2 Directional Convergence

This subsection applies to orthogonal index vectors. In this case $A_{jj'} = 0$ for $j \neq j'$, hence the ODE (9) simplifies

$$\frac{d}{dt}\mathbf{u}_j = \sum_{p \geq p^*} c_p p\big(\mathbf{u}_j^{p-1} - \mathbf{u}_j \sum_{j'} \mathbf{u}_{j'}^p\big). \qquad \text{(orthogonal index vectors)} \tag{15}$$

First, we make a helper lemma for the invariance.

Lemma B.2 (Symmetry breaking happens at initialization)**.** *Wlog assume that* $|\mathbf{u}_1(0)| = \max_{j\in\{1,\dots,k\}} |\mathbf{u}_j(0)|$. *Assume that* σ^* *has an information exponent* $p^* \geq 2$. *We show the following*

(i) if all correlations are positive at initialization, then the maximum correlation is preserved, i.e. $\mathbf{u}_1(t) = \max_{j\in\{1,\dots,k\}} \mathbf{u}_j(t)$ *for any* σ^*,

(ii) if σ^* *is even, then* $|\mathbf{u}_1(t)| = \max_{j\in\{1,\dots,k\}} |\mathbf{u}_j(t)|$ *for any initialization.*

Lemma B.2 gives us that if all correlations are positive at initialization, $j^* = \text{argmax}_{j\in\{1,\dots,k\}} \mathbf{w}_i(t)^T \mathbf{v}_j$ is fixed over time. The same argument holds up to a sign change for even activation functions and arbitrary initializations.

Proof. Let us start with the simplest case (i) where the dot products are assumed to be positive at initialization. Since the dot products are assumed to be positive at initialization, they remain non-negative at all times (see Eq. 10 and the comments therein). That is, we have $\mathbf{u}_j(t) \geq 0$ for all $j \in [k]$ and for all $t \geq 0$.

Since $\mathbf{u}_1$ is assumed to be the leading correlation at initialization, this implies that

$$\mathbf{u}_1(0) = \max_{j\in\{1,\dots,k\}} \mathbf{u}_j(0) > 0.$$

We want to show

$$\mathbf{u}_1(t) = \max_{j\in\{1,\dots,k\}} \mathbf{u}_j(t) \quad \text{for all times } t.$$

Our strategy is to show that $\frac{d}{dt}(\mathbf{u}_1 - \mathbf{u}_j)$ is non-negative if $\mathbf{u}_1 - \mathbf{u}_j$ is positive. We will use that $\mathbf{u}_1 - \mathbf{u}_{j'}$ is non-negative for all j'. Taking the difference between the time derivative of $\mathbf{u}_1$ and $\mathbf{u}_j$ gives

$$\frac{d}{dt}(\mathbf{u}_1 - \mathbf{u}_j) = \sum_{p\geq p^*} c_p p\big((\mathbf{u}_1^{p-1} - \mathbf{u}_j^{p-1}) - (\mathbf{u}_1 - \mathbf{u}_j)\sum_{j'=1}^{k} \mathbf{u}_{j'}^p\big). \tag{16}$$

The following chain of inequalities completes the argument

$$\begin{aligned}
\mathbf{u}_1^{p-1} - \mathbf{u}_j^{p-1} &\geq (\mathbf{u}_1^{p-1} - \mathbf{u}_j^{p-1})(\sum_{j'=1}^{k} \mathbf{u}_{j'}^2) && \text{since } \textstyle\sum \mathbf{u}_j^2 \leq 1 \text{ and } \mathbf{u}_1 \geq \mathbf{u}_j \\
&= (\mathbf{u}_1 - \mathbf{u}_j)(\mathbf{u}_1^{p-2} + \mathbf{u}_1^{p-3}\mathbf{u}_j + \dots + \mathbf{u}_j^{p-2})(\sum_{j'=1}^{k} \mathbf{u}_{j'}^2) \\
&\geq (\mathbf{u}_1 - \mathbf{u}_j)\mathbf{u}_1^{p-2}(\sum_{j'=1}^{k} \mathbf{u}_{j'}^2) && \text{since } \mathbf{u}_j \geq 0 \\
&\geq (\mathbf{u}_1 - \mathbf{u}_j)(\sum_{j'=1}^{k} \mathbf{u}_{j'}^p) && \text{since } \mathbf{u}_1 \geq \mathbf{u}_{j'} \text{ for all } j'
\end{aligned}$$

There is a remaining step to complete the proof of the case (i): the leading correlation might not be unique at initialization. That is the case $\mathbf{u}_1(0) = \dots = \mathbf{u}_\ell(0) > \max_{j\in\{\ell+1,\dots,k\}} \mathbf{u}_j(0)$. Observe that Eq. 16 ensures that $\mathbf{u}_1(t) = \dots = \mathbf{u}_\ell(t)$ for all times and the argument above applies to the differences $\mathbf{u}_1 - \mathbf{u}_j$ for $j \in \{\ell+1, \dots, k\}$. This completes the proof of case (i).

We will handle the case (ii) very similarly by using a sign flip argument. First, let us show that the odd Hermite coefficients of σ^* vanish to zero when σ^* is even.

Using Rodrigues' formula, we get that h_p is an odd function when p is odd (see the Wikipedia page). Hence the integrand below is odd and integrating it from $-\infty$ to ∞ gives a zero, i.e.,

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}\sigma^*(x)h_p(x)\exp(-\frac{x^2}{2})dx=0.$$

Therefore, in this case, for the ODE in Eq. 15, the terms corresponding to the Hermite modes p is zero when p is odd. Let us compute the time derivative of $-\mathbf{u}_j$

$$\begin{aligned}\frac{d}{dt}(-\mathbf{u}_j)&=(-\mathbf{u}_j)\sum_{p\text{ even}}c_p p(\mathbf{u}_j^{p-2}-\sum_{j'=1}^{k}\mathbf{u}_{j'}^p),\\&=(-\mathbf{u}_j)\sum_{p\text{ even}}c_p p((-\mathbf{u}_j)^{p-2}-\sum_{j'=1}^{k}(-\mathbf{u}_{j'})^p).\end{aligned}$$

The second equality holds since all Hermite modes are even. Hence, if $(\mathbf{u}_1(t),...,\mathbf{u}_k(t))$ is a solution of this ODE with an initial condition $(\mathbf{u}_1(0),...,\mathbf{u}_k(0))$, then $(\xi_1\mathbf{u}_1(t),...,\xi_k\mathbf{u}_k(t))$ is also a solution with the initial condition $(\xi_1\mathbf{u}_1(0),...,\xi_k\mathbf{u}_k(0))$ where $\xi_i\in\{\pm1\}$.

Note that for any initialization, we can flip the sign of the negative correlations. Applying the result of case (i), we get that the maximum correlation $\xi_1\mathbf{u}_1(t)$ is preserved for all times. This completes the proof of case (ii). □

Proposition B.1 (Directional Convergence)**.** *Let $\ell\in[k]$ and $I=[k]\setminus[\ell]$. Assume that $\mathbf{u}_j(0)>0$ for all $j\in[k]$ and wlog $\mathbf{u}_1(0)=\cdots=\mathbf{u}_\ell(0)>\max_{j\in I}\mathbf{u}_j(0)$. For any σ^* with information exponent $p^*\geq3$, the dynamics converge to*

$$\lim_{t\to\infty}\mathbf{u}_j(t)=\frac{1}{\sqrt{\ell}}\quad\textit{for}\quad j\in[\ell].$$

Moreover, if σ^* is even and with information exponent $p^*\geq4$, assume that wlog $|\mathbf{u}_1(0)|=\cdots=|\mathbf{u}_\ell(0)|>\max_{j\in I}|\mathbf{u}_j(0)|$, the dynamics converge to

$$\lim_{t\to\infty}\mathbf{u}_j(t)=\frac{\operatorname{sgn}(\mathbf{u}_j(0))}{\sqrt{\ell}}\quad\textit{for}\quad j\in[\ell].$$

This implies $\lim_{t\to\infty}\mathbf{u}_j(t)=0$ for $j\in I$ for both cases.

Proof. Since the dot products are assumed to be positive at initialization, they remain non-negative at all times (see Eq. 10 and the comments therein). That is, we have $\mathbf{u}_j(t)\geq0$ for all $j\in[k]$ and for all $t\geq0$.

Let us recall the evolution of the total alignment from the proof of Theorem B.1

$$\frac{1}{2}\frac{d}{dt}s_2=\sum_{p\geq p^*}c_p p(\sum_{j=1}^{k}\mathbf{u}_j^p)(1-s_2).$$

For all $s_2\in(0,1)$, the time derivative is positive and s_2 increases over time until it reaches $s_2=1$. Once $s_2=1$ is satisfied, the time derivative vanishes and s_2 remains constant. In other words, once $\mathbf{w}$ reaches the subspace of the index vectors, it remains in that subspace for all times.

To specify in which direction this convergence happens, we analyze the fixed points of the ODE. The fixed points correspond to the solution of the following non-linear system of equations with k variables

$$\sum_{p\geq p^*}c_p p\big(\mathbf{u}_j^{p-1}-\mathbf{u}_j\sum_{j'=1}^{k}\mathbf{u}_{j'}^p\big)=0\quad\text{for}\quad j\in\{1,...,k\}.$$

From the proof of Lemma B.2, the following holds for all times

$$\mathbf{u}_1(t)=...=\mathbf{u}_\ell(t)>\max_{j\in\{\ell+1,...,k\}}\mathbf{u}_j(t)\geq0.\tag{17}$$

Moreover, note that for $j \in \{1, ..., \ell\}$, and for $p \geq 3$

$$\mathbf{u}_j^{p-1} \geq \mathbf{u}_j^{p-1}(\sum_{j=1}^{k} \mathbf{u}_j^2) \geq \mathbf{u}_j \sum_{j'=1}^{k} \mathbf{u}_{j'}^p \tag{18}$$

where the first inequality is due to $\sum \mathbf{u}_j^2 \leq 1$ and the second one is due to $\mathbf{u}_j > \mathbf{u}_{j'}$ for $j' \in \{\ell+1, ..., k\}$, $\mathbf{u}_j > 0$, and $\mathbf{u}_{j'} \geq 0$ for all times (see Eq. 17).

For configurations for which the above inequality (Eq. 18) is strict, we have

$$\sum_{p \geq p^*} c_p p\big(\mathbf{u}_j^{p-1} - \mathbf{u}_j \sum_{j'} \mathbf{u}_{j'}^p\big) > 0.$$

For configurations that are fixed points, the above inequality (Eq. 18) must be tight which implies the following

$$\sum_{j=1}^{k} \mathbf{u}_j^2 = 1 \quad \text{and} \quad \mathbf{u}_{j'} = 0 \quad \text{for all} \;\; j' \in \{\ell+1, ..., k\}.$$

Finally, setting $\mathbf{u}_j = \alpha$ for $j \in \{1, ..., \ell\}$ and $\mathbf{u}_{j'} = 0$ for $j' \in \{\ell+1, ..., k\}$ gives $\alpha = \frac{1}{\sqrt{\ell}}$ which completes the proof for the first case.

We can handle the second case similarly by using a sign flip argument as done in the proof of Lemma B.2. In particular, consider the dynamics of $(\xi_1 \mathbf{u}_1, ..., \xi_k \mathbf{u}_k)$ with an initial condition $(\xi_1 \mathbf{u}_1(0), ..., \xi_k \mathbf{u}_k(0))$ where $\xi_i = \text{sgn}(\mathbf{u}_i(0))$. The first case then shows that the dynamics converge to

$$\lim_{t \to \infty} \xi_i \mathbf{u}_i(t) = \frac{1}{\sqrt{\ell}}$$

which is equivalent to the statement in the second case up to a multiplication of both sides with ξ_i.

□

C Saddle-to-Minimum Transition

First, we will evaluate the loss function at the 'average' fixed point given by

$$\bar{\mathbf{w}} = \frac{\mathbf{v}_1 + \cdots + \mathbf{v}_k}{\|\mathbf{v}_1 + \cdots + \mathbf{v}_k\|},$$

where

$$\|\mathbf{v}_1 + \cdots + \mathbf{v}_k\|^2 = k + k(k-1)\beta, \quad \beta \in [0, 1].$$

The dot product between $\bar{\mathbf{w}}$ and the index vector $\mathbf{v}_j$ is given by

$$\bar{\mathbf{w}} \cdot \mathbf{v}_j = \Big(\frac{1 + (k-1)\beta}{k}\Big)^{1/2}.$$

Using the expression in equation (3), the loss function evaluated at $\bar{\mathbf{w}}$ can be computed as follows

$$\begin{aligned} L_\beta(\bar{\mathbf{w}}) &= C - \sum_{p \geq p^*} c_p \sum_{j=1}^{k} (\bar{\mathbf{w}}^T \mathbf{v}_j)^p \\ &= C - \sum_{p \geq p^*} c_p k \Big(\frac{1 + (k-1)\beta}{k}\Big)^{p/2}. \end{aligned}$$

In particular, evaluating the loss function at $\beta = 0$ and $\beta = 1$ implies

$$L_\beta(\bar{\mathbf{w}})\big|_{\beta=0} = C - \sum_{p \geq p^*} c_p \Big(\frac{1}{k}\Big)^{p/2} k, \quad L_\beta(\bar{\mathbf{w}})\big|_{\beta=1} = C - \sum_{p \geq p^*} c_p k.$$

Moreover, the derivative with respect to β satisfies

$$\frac{d}{d\beta}L_\beta(\bar{\mathbf{w}}) = -\sum_{p\geq p^*}\frac{c_p p}{2}\Big(\frac{1+(k-1)\beta}{k}\Big)^{p/2-1}(k-1) < 0.$$

Intuitively, the correlation between $f(\mathbf{x}) = \sigma(\bar{\mathbf{w}}^T\mathbf{x})$ and the multi-index model increases as the target vectors approach each other, i.e., β increases, hence decreasing the correlation loss.

The manifold geometry of the unit sphere makes it complicated to compute the Hessian. One may need to change the coordinate system to polar coordinates, which makes computations cumbersome for $d \geq 3$.

We consider all paths on the unit sphere passing through $\bar{\mathbf{w}}$. In particular, we consider the unit circle spanned by $\bar{\mathbf{w}}$ and $\mathbf{v}$ such that $\bar{\mathbf{w}}^T\mathbf{v} = 0$, that is

$$\mathbf{w}(\theta) = \bar{\mathbf{w}}\cos(\theta) + \mathbf{v}\sin(\theta).$$

The loss on the circle is then one-dimensional

$$L_\beta(\theta) = C - \sum_{p=p^*}^{P} c_p \sum_{j=1}^{k}((\bar{\mathbf{w}}\cos(\theta) + \mathbf{v}\sin(\theta))^T\mathbf{v}_j)^p$$

which can be viewed as a periodic function with period 2π. This allows us to compute derivatives without worrying about the reparameterization of the unit sphere. Note that we truncated the series at the P-th term due to technical reasons. This corresponds to choosing a polynomial target activation function σ^* with degree P and information exponent p^*. Let us restate Theorem 3.2 below and then give the proof.

Theorem C.1 (Saddle-to-Minimum)**.** *Assume $d > k$. $\bar{\mathbf{w}}$ is a strict saddle when the dot product is upper bounded by*

$$\beta < \frac{p^*-2}{k+p^*-2},$$

whereas $\bar{\mathbf{w}}$ is a local minimum when the dot product is lower bounded by

$$\frac{P-2}{k+P-2} < \beta.$$

Therefore, for the case $\sigma^* = h_{p^*}$

$$\beta_c = \frac{p^*-2}{k+p^*-2}$$

is the sharp threshold characterizing the transition from a strict saddle to a local minimum.

Proof. The first derivative is given by

$$\frac{d}{d\theta}L_\beta(\theta) = -\sum_{p=p^*}^{P} c_p p \sum_{j=1}^{k}((\bar{\mathbf{w}}\cos(\theta) + \mathbf{v}\sin(\theta))^T\mathbf{v}_j)^{p-1}(-\bar{\mathbf{w}}\sin(\theta) + \mathbf{v}\cos(\theta))^T\mathbf{v}_j.$$

Evaluating the first derivative at zero, we get

$$\frac{d}{d\theta}L_\beta(\theta)\Big|_{\theta=0} = -\sum_{p=p^*}^{P} c_p p \sum_{j=1}^{k}(\bar{\mathbf{w}}^T\mathbf{v}_j)^{p-1}(\mathbf{v}^T\mathbf{v}_j).$$

The inner sum can be viewed as a dot product between $[\alpha, ..., \alpha]$ and $[\mathbf{v}^T\mathbf{v}_1, ..., \mathbf{v}^T\mathbf{v}_k]$. Observe that the evaluation of the derivative at $\theta = 0$ vanishes, since $\sum_{j=1}^{k}\mathbf{v}^T\mathbf{v}_j = c\mathbf{v}^T\bar{\mathbf{w}} = 0$, which implies that $\bar{\mathbf{w}}$ is a fixed point. To study the curvature, we need to compute the second derivatives.

The second derivative is given by

$$\frac{d^2}{d\theta^2}L_\beta(\theta) = -\sum_{p=p^*}^{P} c_p p \sum_{j=1}^{k}\Big((p-1)(\bar{\mathbf{w}}\cos(\theta)+\mathbf{v}\sin(\theta))^T\mathbf{v}_j)^{p-2}((-\bar{\mathbf{w}}\sin(\theta)+\mathbf{v}\cos(\theta))^T\mathbf{v}_j)^2 \\ + ((\bar{\mathbf{w}}\cos(\theta)+\mathbf{v}\sin(\theta))^T\mathbf{v}_j)^{p-1}(-\bar{\mathbf{w}}\cos(\theta)-\mathbf{v}\sin(\theta))^T\mathbf{v}_j\Big).$$

Evaluating the second derivative at zero, we get

$$\frac{d^2}{d\theta^2}L_\beta(\theta)\Big|_{\theta=0} = \sum_{p=p^*}^{P} c_p p \sum_{j=1}^{k}\big(-(p-1)(\bar{\mathbf{w}}^T\mathbf{v}_j)^{p-2}(\mathbf{v}^T\mathbf{v}_j)^2 + (\bar{\mathbf{w}}^T\mathbf{v}_j)^p\big). \tag{19}$$

The curvature at $\bar{\mathbf{w}}$ is positive in the direction of $\mathbf{v}$ if the following holds for all $p \geq p^*$

$$\sum_{j=1}^{k}(p-1)(\bar{\mathbf{w}}^T\mathbf{v}_j)^{p-2}(\mathbf{v}^T\mathbf{v}_j)^2 < \sum_{j=1}^{k}(\bar{\mathbf{w}}^T\mathbf{v}_j)^p \quad \Leftrightarrow$$
$$(p-1)\big(\frac{1+\beta(k-1)}{k}\big)^{p/2-1}\sum_{j=1}^{k}(\mathbf{v}^T\mathbf{v}_j)^2 < k\big(\frac{1+\beta(k-1)}{k}\big)^{p/2} \quad \Leftrightarrow$$
$$(p-1)\sum_{j=1}^{k}(\mathbf{v}^T\mathbf{v}_j)^2 < (1+\beta(k-1)).$$

If $\mathbf{v} \perp \mathbf{V}$ where $\mathbf{V} = \text{Span}(\mathbf{v}_1, ..., \mathbf{v}_k)$, the above inequality is trivially true. We will study the case $\mathbf{v} \in \mathbf{V}$ in what follows by giving an upper bound on the LHS term. By Lemma (A.2) we know that the leading eigenvector of VV^T is $\bar{\mathbf{w}}$ and the orthogonal space to this direction in the subspace $\mathbf{V} = \text{Span}(\mathbf{v}_1, ..., \mathbf{v}_k)$ has eigenvalues $1-\beta$. This implies that

$$\max_{\substack{\mathbf{v}^T\bar{\mathbf{w}}=0 \\ \mathbf{v}\in\mathbf{V}}} \left\| \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_k^T \end{bmatrix} \mathbf{v} \right\|_2^2 = \max_{\substack{\mathbf{v}^T\bar{\mathbf{w}}=0 \\ \mathbf{v}\in\mathbf{V}}} \mathbf{v}^T \begin{bmatrix} \mathbf{v}_1 & \dots & \mathbf{v}_k \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_k^T \end{bmatrix} \mathbf{v} = \lambda_{\max,2}(VV^T) = \lambda_{\max,2}(V^TV) = 1-\beta$$

where $\lambda_{\max,2}$ denotes the second largest eigenvalue. The following condition gives us the desired upper bound on the LHS

$$(p-1)(1-\beta) < 1+\beta(k-1) \quad \Leftrightarrow \quad \beta > \frac{p-2}{k+p-2}.$$

We want this to be satisfied for all $P \geq p \geq p^*$, and the necessary and sufficient condition for this is

$$\beta > \frac{P-2}{k+P-2}. \tag{20}$$

This argument applies to any $\mathbf{v} \perp \bar{\mathbf{w}}$. Wlog, consider $\bar{\mathbf{w}}$ as the north pole, and an arbitrary $\mathbf{v} \perp \bar{\mathbf{w}}$ located on the equator. For every path trough $\bar{\mathbf{w}}$ we can find a $\mathbf{v} \perp \bar{\mathbf{w}}$ on the equator, such that the circle through $\mathbf{v}$ and $\bar{\mathbf{w}}$ coincides with the given path in a neighborhood of $\bar{\mathbf{w}}$. Hence we conclude that $\bar{\mathbf{w}}$ is a local minimum if β is bigger than the lower bound (20).

The curvature at $\bar{\mathbf{w}}$ is negative in the direction of $\mathbf{v}$ if the following holds for all $p \geq p^*$

$$(1+\beta(k-1)) < (p-1)\sum_{j=1}^{k}(\mathbf{v}^T\mathbf{v}_j)^2.$$

This calls for a lower bound on the RHS term. Let us first consider the case when $\mathbf{v} \in \mathbf{V}$. Such a vector $\mathbf{v}$ must lie in the orthogonal complement of the kernel of V^T, since the image of V is orthogonal to the kernel of V^T. In

particular, $V^T\mathbf{v} \neq 0$ and hence $\|V^T\mathbf{v}\|_2^2 > 0$. This allows us to argue that

$$\min_{\substack{\mathbf{v}^T\bar{\mathbf{w}}=0 \\ \mathbf{v}\in\mathbf{V}}} \left\| \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_k^T \end{bmatrix} \mathbf{v} \right\|_2^2 = \min_{\substack{\mathbf{v}^T\bar{\mathbf{w}}=0 \\ \mathbf{v}\in\mathbf{V}}} \mathbf{v}^T \begin{bmatrix} \mathbf{v}_1 & \dots & \mathbf{v}_k \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_k^T \end{bmatrix} \mathbf{v} = \lambda_{\min\neq 0}(VV^T) = \lambda_{\min}(V^TV) = 1-\beta,$$

where $\lambda_{\min\neq 0}(VV^T)$ is the smallest nonzero eigenvalue of VV^T. The following condition gives us the desired lower bound on the RHS.

$$(1+\beta(k-1)) < (p-1)(1-\beta) \quad \Leftrightarrow \quad \beta < \frac{p-2}{k+p-2}$$

for all $p \geq p^*$. Then the necessary and sufficient condition for this is

$$\beta < \frac{p^*-2}{k+p^*-2}. \tag{21}$$

In particular, if $d = k$, the curvature is negative in every direction since $\mathbb{S}^{d-1} \subset \mathbf{V}$. This implies that $\bar{\mathbf{w}}$ is a local maximum and not a saddle. However, when $k < d$, we can compute the curvature in a direction $\mathbf{v}^\perp$ that is orthogonal to the span of $\{\mathbf{v}_1, ..., \mathbf{v}_k\}$ and also orthogonal to $\bar{\mathbf{w}}$ as a result. Evaluating Eq. (19), observe that the first term is zero and the second term is positive, hence yielding a positive second derivative in this direction.

Therefore, under condition (21), $\bar{\mathbf{w}}$ is a saddle point and it has index k.

□

One may wonder whether $\bar{\mathbf{w}}$ is a fixed point for more general dot product matrices. However, this situation is specific to the equiangular frame. For example, consider $k = 3$ and the following dot product matrix

$$\begin{bmatrix} 1 & \beta_1 & \beta_2 \\ \beta_1 & 1 & \beta_2 \\ \beta_1 & \beta_2 & 1 \end{bmatrix} \qquad \text{where} \quad \beta_1 \neq \beta_2.$$

The average point $\bar{\mathbf{w}}$ is not a fixed point in this case as $\nabla L(\bar{\mathbf{w}})$ is not parallel to $\bar{\mathbf{w}}$.

D Supplementary Simulations

D.1 Bifurcation as k increases

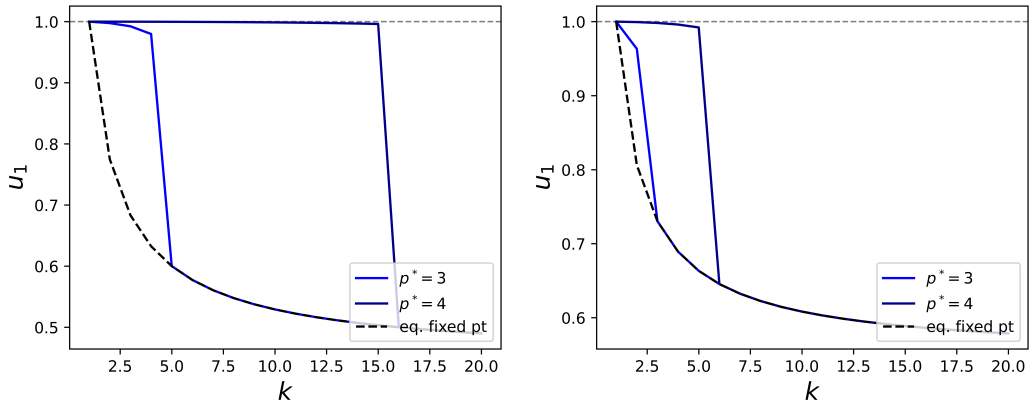


Figure 5: *Maximum dot product at convergence as the number of index vectors k increases from* 1 *to* 20. (left) $\beta = 0.2$ and (right) $\beta = 0.3$. Observe that increasing the number of index vectors pulls the flow away from preferring one of the index vectors to the average of the index vectors as indicated by the black dashed line.

D.2 Phase Portraits

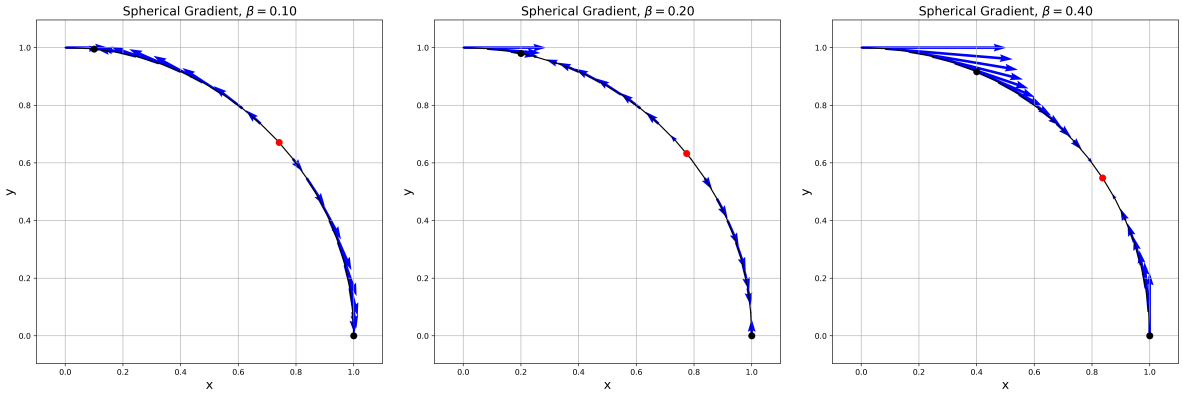


Figure 6: *The spherical gradient flow vector field, as the dot product between the two vectors increases from left to right;* $d = k = 2$. The two index vectors are shown as black dots. $\beta \in \{0.1, 0.2, 0.4\}$ as shown in the title. The average fixed point $\bar{\mathbf{w}}$ (red dot) turns from a maximum (for $\beta \in \{0.1, 0.2\}$) to a minimum for $\beta = 0.4$. The activation function here is h_3 hence the saddle-to-minimum transition happens at $\beta_c = 1/3$.