

A Mechanism for the Emergence of Superposition in Toy Models

Wu Shuin Jian, Choong Kai Zhe, Harshank Shrotriya

Abstract

This work connects the emergence of superposition in neural networks to Nonlinear Matrix Decomposition (NMD). Superposition occurs when networks encode more features than neurons by overlapping representations in lower-dimensional spaces. We demonstrate that the Toy Model of Superposition (TMS) and NMD produce identical anticorrelated feature patterns when processing sparse, nonnegative data through low-rank approximations with ReLU nonlinearities. We conjecture that NMD is the underlying mechanism driving superposition, with networks iteratively discovering higher-rank representations during training. This would reframe superposition as an efficient compression strategy rather than pathological behavior, potentially advancing interpretability research.

1 Introduction

Understanding how neural networks represent and process information has become a central challenge in machine learning, particularly as models grow in scale and complexity. Interpretability methods aim to reverse-engineer the inner workings of these models to reveal the structure of their learned representations. Yet, progress in this area is hampered by a persistent obstacle: neural networks rarely learn features that are cleanly separated. Instead, features are distributed across the network in unintuitive, overlapping ways.

A key hypothesis emerging from recent interpretability work is that neural networks can represent many more features than they have neurons by using a strategy known as **superposition** [1]. In superposition, multiple sparse features could be encoded in overlapping directions within the same representational subspace, rather than being assigned orthogonal, axis-aligned neurons. This poses a fundamental challenge for interpretability, as traditional interpretations often rely on the assumption that each neuron corresponds to a distinct and interpretable concept — an assumption that breaks down under superposition. Sparse autoencoders [2] are an alternative technique that has been developed to tackle superposition, but this approach has ran into challenges regarding precision [3], which await resolution through further theoretical progress.

Despite its importance, the emergence of superposition in the learning dynamics of neural networks remains poorly understood. When does superposition

arise during training? Are features learnt sequentially or simultaneously? Most importantly, how does superposition affect our ability to recover the features a network has learned? These questions are critical for the development of robust interpretability tools, yet remain largely unanswered.

This report argues that superposition is not a pathological behavior, but a *structural consequence* of efficient representation under resource constraints. We emphasize that interpretability research must grapple directly with the dynamics and geometry of superposition if it is to scale to modern models. In doing so, we aim to clarify the central role of superposition in neural representations, and lay the groundwork for approaches that treat superposition not as noise, but as signal — a key to understanding what neural networks know.

2 Iterative mechanism for emergence of superposition

In this section we briefly recap the toy model considered in [1], on which we build our arguments.

2.1 Toy Model of Superposition (TMS)

Input vector: Input with n sparse features:

$$x \in \mathbb{R}^n,$$

where x_i is 0 with probability S_i (sparsity), and otherwise uniformly distributed over $[0, 1]$.

Linear autoencoder:

$$h = Wx,$$

where $W \in \mathbb{R}^{d \times n}$, $d < n$.

Decoder followed by nonlinearity

$$\hat{x} = \text{ReLU}(W^\top h)$$

where $\text{ReLU}(x) = \max(0, x)$ is applied element-wise.

Reconstruction loss (mean squared error)

$$\mathcal{L} = \mathbb{E}_x [\|\hat{x} - x\|^2] = \mathbb{E}_x [\|\text{ReLU}(W^\top Wx) - x\|^2]$$

Loss term The loss term considered here is an expectation over input samples, weighted by feature importance;

$$\mathcal{L} = \mathbb{E}_x \left[\sum_{i=1}^n \alpha_i (\hat{x}_i - x_i)^2 \right]$$

Superposition: When $d < n$, the model reuses directions in $\mathbb{R}^d$ to represent more features than it has dimensions. This introduces interference and requires nonlinearity in the model to make distinction between features.

2.2 Nonlinear Matrix Decomposition

Nonlinear Matrix Decomposition (NMD), introduced in [4], is an iterative method to obtain a low-rank approximation of a sparse non-negative matrix, a problem that frequently shows up in data science. However, we find a striking similarity in the proposed setup for NMD and the toy model considered here.

Let $\mathbf{X}$ be a sparse nonnegative matrix of dimensions $d \times n$. The goal of NMD is to estimate a matrix of lower rank r than $\mathbf{X}$, which we will refer to as $\mathbf{\Theta}$, such that $\mathbf{X} \approx \max(0, \mathbf{\Theta})$. To achieve this, an auxiliary matrix $\mathbf{Z}$ that can be of any rank is introduced and the following minimization problem is formulated thus:

$$\min_{\mathbf{Z}, \mathbf{\Theta}} \|\mathbf{Z} - \mathbf{\Theta}\|_F^2 \quad \text{s.t.} \quad \begin{cases} \text{rank}(\mathbf{\Theta}) = r \\ \max(0, \mathbf{Z}) = \mathbf{X} \end{cases} \quad (1)$$

where the subscript F refers to the Frobenius norm. Note that the condition $\max(0, \mathbf{Z}) = \mathbf{X}$ means that the elements of $\mathbf{X}$ can be recovered perfectly from $\mathbf{Z}$

The minimization is achieved by performing an alternating optimization over $\mathbf{Z}$ and $\mathbf{\Theta}$, i.e. keeping one constant while adjusting the other. When keeping $\mathbf{\Theta}$ fixed, $\mathbf{Z}$ is optimized by:

$$\mathbf{Z}_{ij} = \begin{cases} \mathbf{X}_{ij} & \text{if } \mathbf{X}_{ij} > 0 \\ \min(0, \mathbf{\Theta}_{ij}) & \text{if } \mathbf{X}_{ij} = 0 \end{cases} \quad (2)$$

On the other hand, when $\mathbf{Z}$ is held constant, $\mathbf{\Theta}$ is optimized by performing a truncated SVD to obtain the best matrix approximation of rank r to $\mathbf{Z}$. Note that the choice of rank r is arbitrary here, and in general the larger r is the better the approximation.

One choice of initial values for $\mathbf{Z}$ and $\mathbf{\Theta}$ is derived from the mean, $\bar{x}$, and variance, σ^2 of the elements of $\mathbf{X}$, defined as:

$$\bar{x} = \frac{1}{dn} \sum_{i,j} X_{ij}, \quad \text{and} \quad (3)$$

$$\sigma^2 = \frac{1}{dn} \sum_{i,j} (X_{ij} - \bar{x})^2 \quad (4)$$

From there, the initial choice for $\mathbf{Z}$ and $\mathbf{\Theta}$ is as follows:

$$\mathbf{\Theta}_{ij} = \bar{x} \quad \text{and} \quad (5)$$

$$\mathbf{Z}_{ij} = \mathcal{N}(\mathbf{\Theta}_{ij}, \sigma^2) \quad (6)$$

where $\mathcal{N}(\mu, \sigma^2)$ represents a Gaussian distribution with mean μ and variance σ^2 . Alternatively, one could initialize with $\mathbf{Z} = \mathbf{X}$ and $\mathbf{\Theta} = \mathbf{0}$.

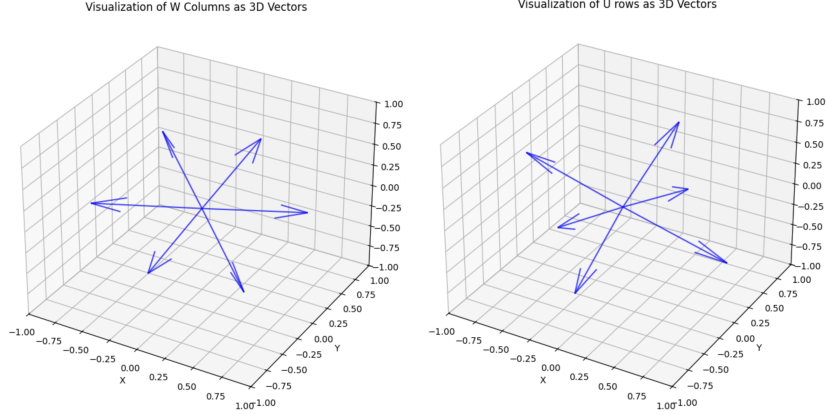


Figure 1: Left: a plot of the feature directions learned by a single-layer ReLU network with hidden dimension 3, when trained on a set $\mathbf{X}$ of sparse features of dimension 6. Right: the basis vectors of Θ of rank 3, when estimating the same $\mathbf{X}$. Both plots display sets of anticorrelated features, a hallmark of superposition.

2.3 Experiment 1 (NMD producing superposition)

One of the central observations in the toy model is the emergence of *superposition*—the phenomenon where more features are encoded than there are available neurons (dimensions). As the autoencoder is trained on sparse, nonnegative input data using a bottleneck layer (i.e., a representation space with fewer dimensions than input features), it learns to *compress* and *reconstruct* the input by reusing basis directions. This leads to overlapping (or superposed) representations of features in the lower-dimensional latent space.

A useful lens to understand this behavior is provided by NMD, which introduces a method for approximating a *sparse nonnegative matrix* $X \in \mathbb{R}^{d \times n}$ (with columns representing sparse data vectors) using a *nonlinear transformation* of a *low-rank matrix* Θ . Their model seeks an approximation of the form:

$$X \approx f(\Theta).$$

The resemblance to the autoencoder setup from [1] is striking:

- The input data is similarly sparse and nonnegative.
- The encoder computes a latent code $h = Wx$, and the decoder reconstructs $\hat{x} = \text{ReLU}(W^\top h)$.
- The training objective minimizes the weighted reconstruction error $\mathbb{E}_x[\alpha_i(x_i - \hat{x}_i)^2]$ effectively learning a *nonlinear low-rank approximation* of the input space.

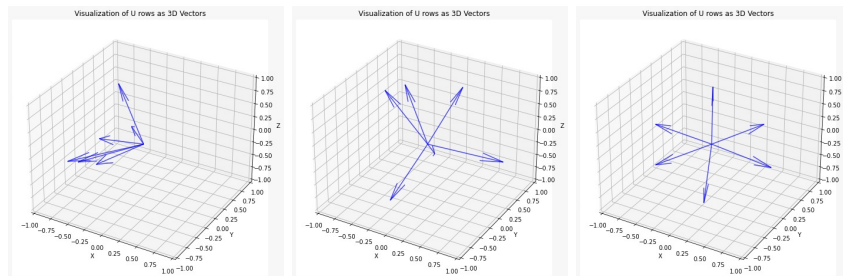


Figure 2: The basis vectors of Θ of ranks 1 (left), 2 (middle), and 3 (right), when estimating $\mathbf{X}$ of rank 6.

Both settings rely on *nonlinear filtering* (via ReLU) applied to a *low-rank representation* to approximate sparse high-dimensional data. In this context, the phenomenon of superposition can be seen as a *consequence of attempting to preserve information about many sparse features through a low-rank nonlinear embedding*. Just as in Saul’s formulation [4], where sparsity is retained through nonlinear transformations despite dimensionality reduction, the toy model encodes multiple sparse features into overlapping neuron activations.

This perspective also reinforces the intuition behind the *phase change* observed in the toy model. As the number of active features per input increases (i.e., lower sparsity), the network must resort to denser superposition to represent them in the limited latent space—mirroring the increasing approximation error observed in Saul’s model when compressing less sparse data.

We applied NMD to the same data set used in the TMS to investigate if superposition also occurs. As seen in Figure 1, the basis vectors discovered by NMD show remarkable similarities to those found by the TMS. In both cases, sets of anticorrelated features are clearly visible, strongly pointing to the presence of superposition.

Thus, *superposition emerges naturally as a nonlinear strategy for low-rank approximation of sparse high-dimensional data*, enabling models to compress more features than dimensions by sharing directions in the latent space and exploiting sparsity and nonlinearity for selective decoding.

3 The Dynamics of Superposition

As has been previously noted in [5], a slight modification to the TMS model, with the ReLU function replaced with a linear function, produces a system that can be minimized through principle component analysis. The learning dynamics for these systems are empirically characterized by a loss that resembles a stepwise function, jumping between saddle points in a loss landscape. Previous work on Alternating Gradient Flows (AGF) [6] has sought to explain the origin of such dynamics via alternating regimes of *utility maximization*, which precipitates

a change in the active set of neurons, and *cost minimization*, where a newly established set of neurons learns a stable configuration.

In a similar manner, the reconstruction loss during the learning process of the toy models in [1] goes down in a step-like fashion. Additionally, each step was correlated with the model increasing the dimensionality of its feature embedding within its hidden dimension. Borrowing the framework in [6], we conjecture that each these steps correspond to a period of utility maximization, after which a new hidden dimension (or neuron) is exploited to differentiate the sparse input features further.

We also conjecture that NMD is the mechanism behind the occurrence of superposition in the TMS. In particular, the step like loss curve for TMS optimization make use of a hierarchy of increasing rank NMD optimizations to solve for feature geometry (of course the rank could only increase up to the number of neurons in the model). Beginning with a rank 1 NMD solution, the first step-wise drop in loss occurs when a rank 2 NMD solution is found, and as higher rank NMD solutions are found the loss begins to taper off. As seen in Figure 2, performing NMD with matrix Θ of ranks 1, 2 and 3 yields basis vectors that are increasingly spread out and orthogonal, eventually matching the form produced by the TMS.

4 Conclusion and further work

We conjecture that superposition is a key feature of neural networks that allows them to efficiently model high dimensional data with limited resources, and understanding how this process occurs could lead to tools that aid with interpretability. In particular, NMD seems to provide insight into superposition as the neural network steadily discovering higher rank representations of the input data set. A deeper exploration of NMD, and connection to AGF, is needed to test this idea, but our initial experiment shows early promise.

5 Acknowledgements

We are highly grateful to Daniel Kunin for his mentorship during the course of this hackathon.

References

- [1] Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022.

- [2] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. 2023.
- [3] David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. 2025.
- [4] Lawrence K. Saul. A nonlinear matrix decomposition for mining the zeros of sparse data. *SIAM Journal on Mathematics of Data Science*, 4(2):431–463, 2022.
- [5] Zhongtian Chen, Edmund Lau, Jake Mendel, Susan Wei, and Daniel Murfet. Dynamical versus bayesian phase transitions in a toy model of superposition. 2023.
- [6] Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James B. Simon, Michael R. DeWeese, Surya Ganguli, and Nina Miolane. Alternating gradient flows: A theory of feature learning in two-layer neural networks. 2025.