清華大學 Tsinghua University

Toy Models of Superposition

Zhu Liu

2024.06.19

Toy Models of Superposition

AUTHORS

Nelson Elhage*, Tristan Hume*, Catherine Olsson*, Nicholas Schiefer*, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg*, Christopher Olah‡

AFFILIATIONS

Anthropic, Harvard

PUBLISHED

Sept 14, 2022

* Core Research Contributor; ‡ Correspondence to colah@anthropic.com; Author contributions statement below.

https://transformer-circuits.pub/2022/toy_model/index.html

Superposition

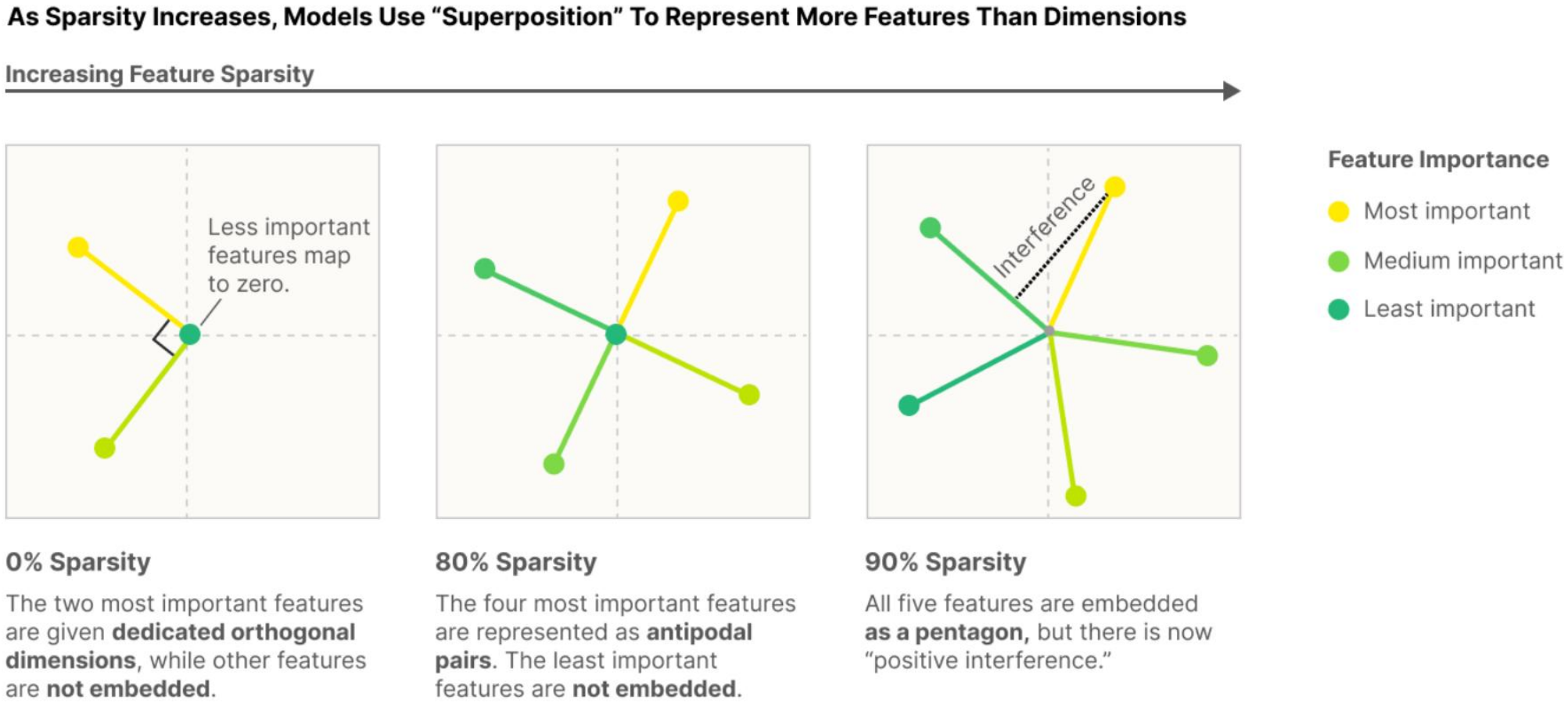
- 叠加 [量子系统在被测量之前**同时**处于**多种状态**的能力；数学...]
- Individual neurals represents unrelated concepts/features.
- Vision: One neural can represent color red, a left-facing curve or a dog noise
- [Similar case in a word: homonymy (and polysemy??)]
- In models: sometimes one-to-one, sometimes not.

Superposition

- Why is it that neurons sometimes align with features and sometimes don't? Why do some models and tasks have many of these clean neurons, while they're vanishingly rare in others?
- Toy model: small ReLU networks trained on synthetic data with sparse input features
- To investigate how and when models represent more features than they have dimensions (number of neurons)?

Feature Sparity increases...

n_features = 5
m_neurons = 2



Key results and broader examples

- Superposition is a real, observed phenomenon
- Both monosemantic and polysemantic neurons can form
- At least some kinds of computation can be performed in superposition
- Whether features are stored in superposition is governed by a phase change
- Superposition organizes features into geometric structures such as digons, trangles, pentagons, tetrahedrons (geometry).
- Adversarial examples and grokking, MoE, training dynamics, larger models....

Definitions

- Linear representation hypothesis of neural networks
  - **Decompositionality**: Network representations can be described in terms of *independently understandable* features.
  - **Linearity**: Features are represented by direction.

- Linear Structure: “word embeddings have a gender direction”
- A property for Mechanism Interpretability:

  e.g. to identify the individual features within a representations

Empirical Phenomena

- Word Embeddings: directions
- Latent Spaces: “vector arithmetic”
- Interpretable Neurons
- Universality
- Polysemantic Neurons

What are features? (Working Definitions)

- Features as arbitrary functions.
- Features as **interpretable** properties.
- Neurons in Sufficiently Large Models

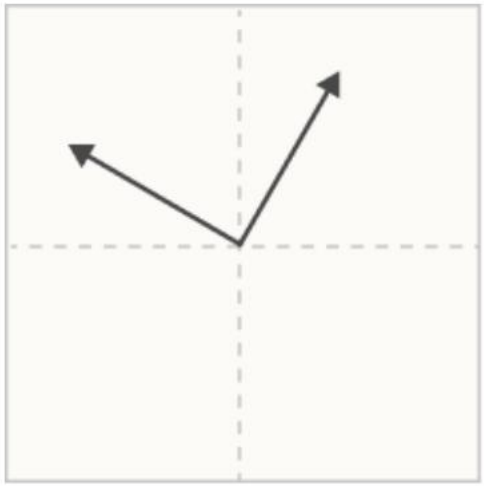
- Features as directions: directions as W_i, and activated values on that direction x_i

$$h = x_1 W_1 + x_2 W_2 + \dots$$

Definitions

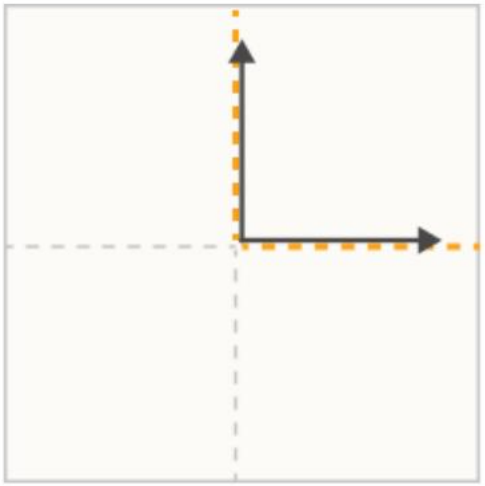
- Privileged Basis: Only some representations have a privileged basis which encourages features to align with basis directions (i.e. to correspond to neurons) .
- Superposition: Linear representations can represent more features than dimensions

Priviledged basis vs. non-priviledged



In a **non-privileged basis**, features can be embedded in any direction. There is no reason to expect basis dimensions to be special.

Examples: word embeddings, transformer residual stream



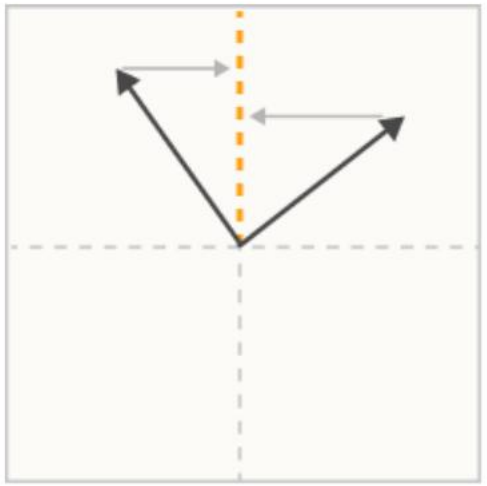
In a **privileged basis**, there is an incentive for features to align with basis dimensions. This doesn't necessarily mean they will.

Examples: conv net neurons, transformer MLPs

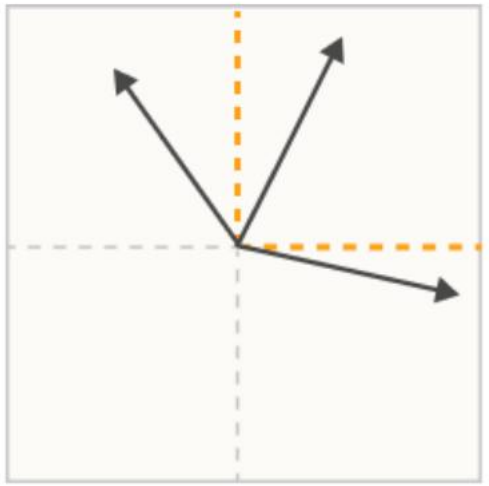
- Whether the certain basis corresponds to certain feature direction
- Priviledge basis -> basis aligned (one hot as feature directions)

Superposition Hypothesis

- Linear representations can represent more features than dimensions



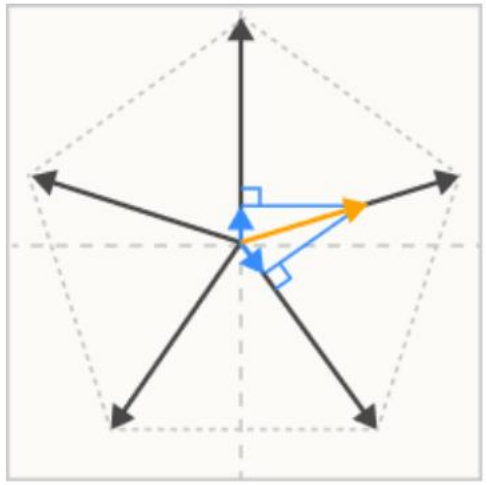
Polysemanticity is what we'd expect to observe if features were not aligned with a neuron, despite incentives to align with the privileged basis.



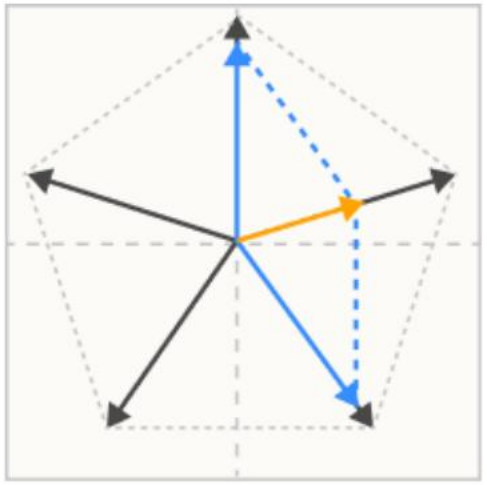
In the **superposition hypothesis**, features can't align with the basis because the model embeds more features than there are neurons. Polysemanticity is inevitable if this happens.

Superposition Hypothesis

- features are represented as almost-orthogonal directions in the vector space of neuron outputs
- Interference cost could be alleviated with highly sparse features

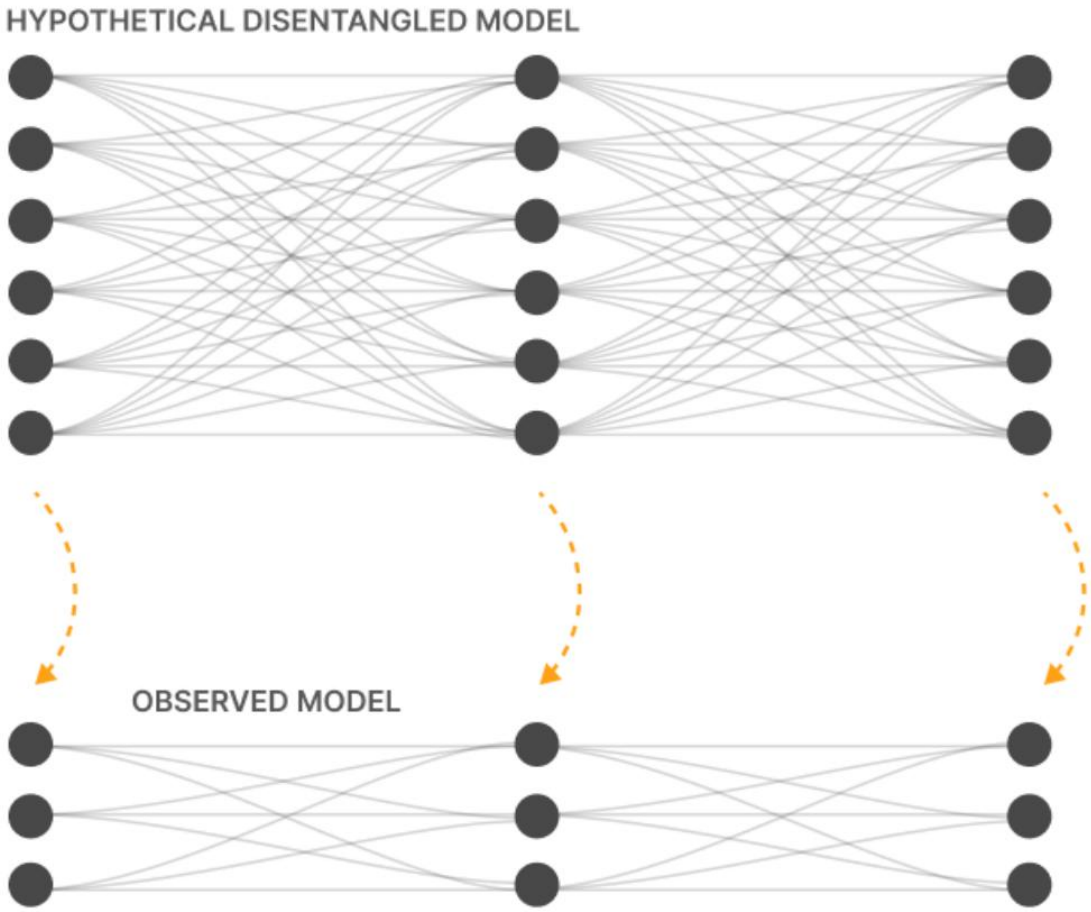


Even if only **one sparse feature** is active, using linear dot product projection on the superposition leads to **interference** which the model must tolerate or filter.



If the features aren't as sparse as a superposition is expecting, **multiple present features** can additively interfere such that there are multiple possible nonlinear reconstructions of an **activation vector**.

Superposition Hypothesis - Simulation



Under the superposition hypothesis, the neural networks we observe are **simulations of larger networks** where every neuron is a disentangled feature.

These idealized neurons are **projected** on to the actual network as “almost orthogonal” vectors over the neurons.

The network we observe is a **low-dimensional projection** of the larger network. From the perspective of individual neurons, this presents as polysemanticity.

Summary: A Hierarchy of Feature Properties

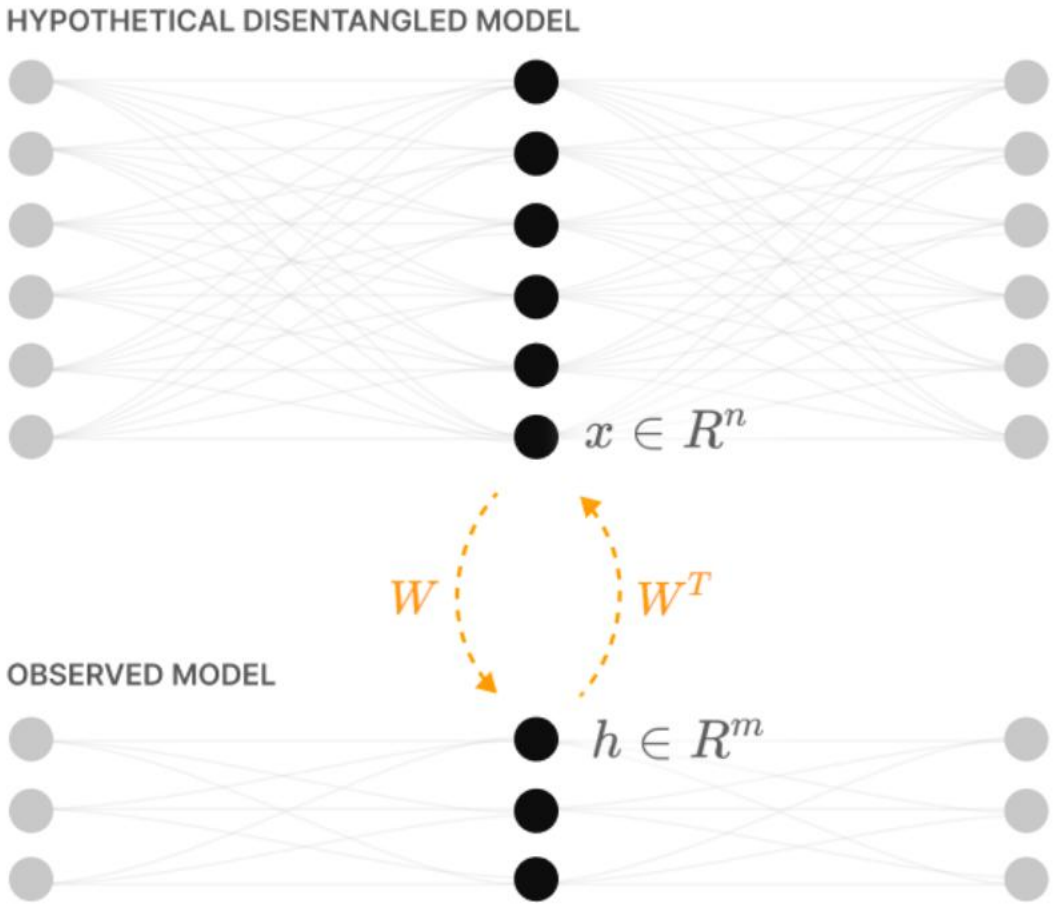
- Decomposability: from representation to features
- Linearity: from features to representation in a linear way
- Superposition vs Non-Superposition
- Basis-Aligned: all W_i are one-hot basis vectors

- Superposition has to have priviledged basis
- Basis-Aligned -> non-superposition

Anything to discuss?

Experiments

- Whether a neural network can project a high dimensional vector x (n) into a lower dimensional vector h (m) and recover it?
- Store and recover (Autoencoder with a bottleneck, reverse of sparse dictionary)
- Synthetic data for x with n features with properties:
  - Sparsity: $p(x \quad = 0) = S$
    $p(x \sim U(0,1)) = 1 - S$
  - Importance: I_i
  - n > m



Experiments

$$x \in R^n \quad h \in R^m \quad b \in R^n$$

$$W \in R^{m \times n} \quad W^T \in R^{n \times m} \quad W^T W \in R^{n \times n}$$

- Two mapping models

Linear Model

$$h = Wx$$
$$x' = W^T h + b$$

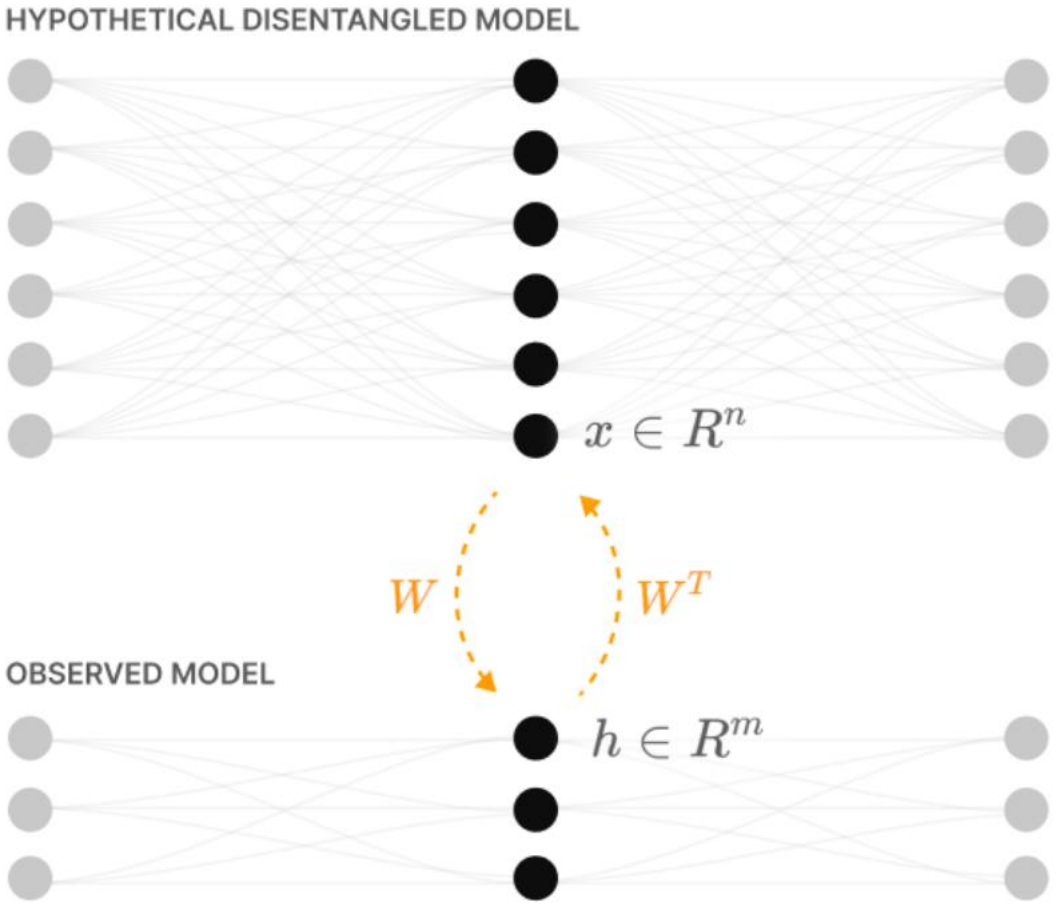
$$x' = W^T W x + b$$

ReLU Output Model

$$h = Wx$$
$$x' = \mathrm{ReLU}(W^T h + b)$$

$$x' = \mathrm{ReLU}(W^T W x + b)$$

$$L = \sum_x \sum_i I_i (x_i - x'_i)^2$$



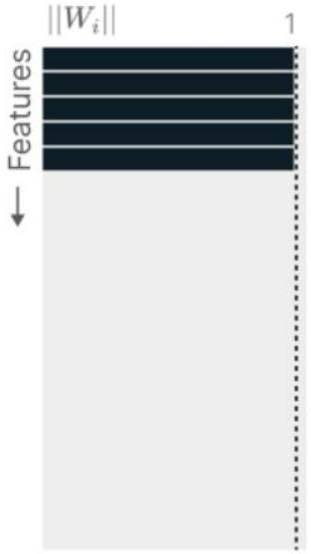
Visualization

It tends to be easier to visualize W^TW than W.

Here we see that W^TW is an **identity matrix** for the most important features and **0** for less important ones.

We can also look at the bias, b.

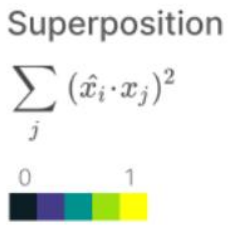
The bias is **zero** for features learned to pass though, and the **expected value** (a positive number) for others.



We want to understand which features the model chooses to represent in its hidden representation, and whether they're orthogonal to each other.

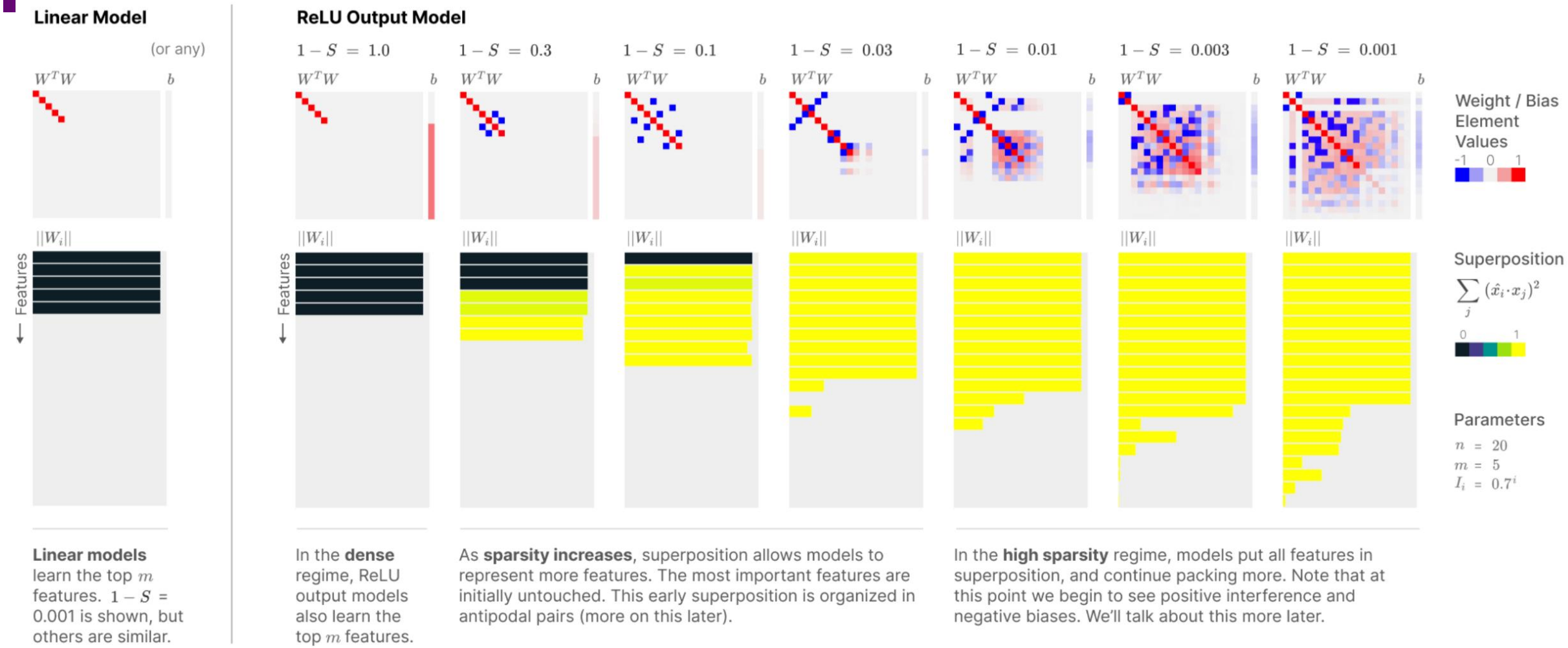
To do this, we visualize the norm of each feature's direction vector, $||W_i||$. This will be ~1 if a feature is fully represented, and zero if it is not. For each feature, we also use color to visualize whether it is orthogonal to other features (i.e. in superposition).

This model simply dedicates one dimension to each of the most important features, representing them orthogonally.



- Feature Representation vs Interference (Superposition/Orthology)
- W^TW and b
- $\|W_i\|$ and $\sum_{j \neq i} (W_i \bullet W_j)^2$

Results (n=20, m=5, I decreases)



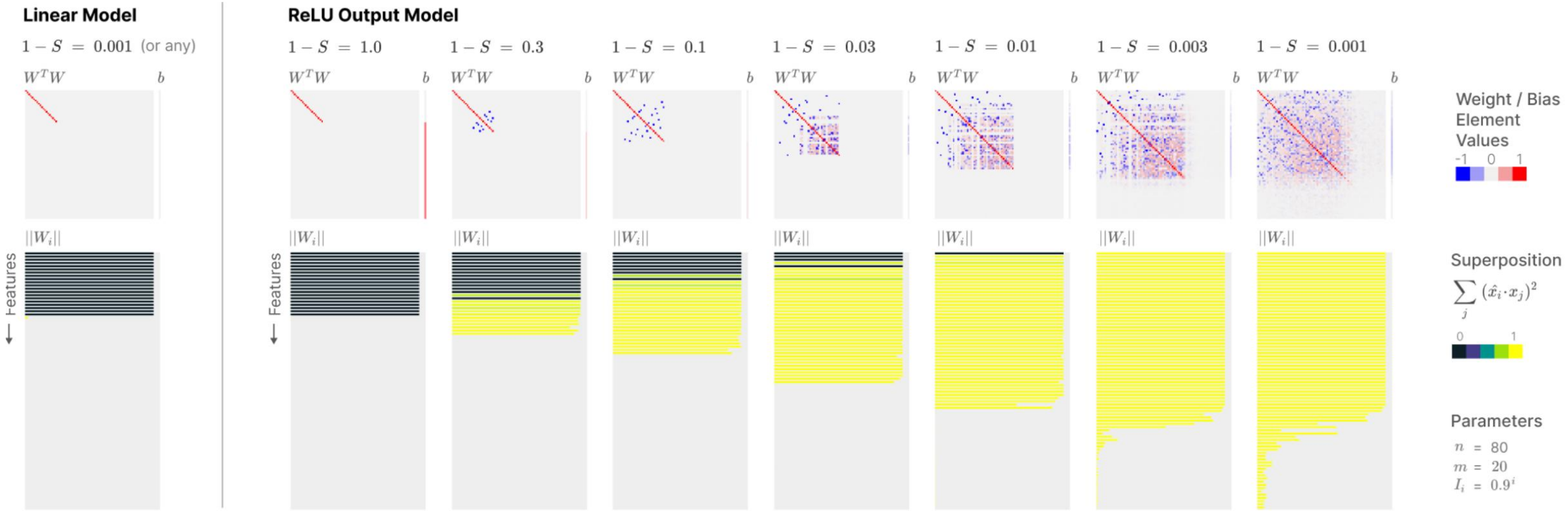
Linear models learn the top m features. $1-S$ = 0.001 is shown, but others are similar.

In the **dense** regime, ReLU output models also learn the top m features.

As **sparsity increases**, superposition allows models to represent more features. The most important features are initially untouched. This early superposition is organized in antipodal pairs (more on this later).

In the **high sparsity** regime, models put all features in superposition, and continue packing more. Note that at this point we begin to see positive interference and negative biases. We'll talk about this more later.

More features and more dimensions



Mathematical Understanding

- Empirical result: adding a ReLU to the output of the model allowed a radically different solution - superposition
- Can we analytically understand why superposition is occuring?
- Inspiration from PCA
- feature representation vs interference

feature benefit vs interference

$$x' = W^T W x + b$$

$$L = \sum_x \sum_i I_i (x_i - x_i')^2$$

$$L \sim \sum_i I_i (1 - ||W_i||^2)^2 + \sum_{i \neq j} I_j (W_j \cdot W_i)^2$$

Feature benefit is the value a model attains from representing a feature. In a real neural network, this would be analagous to the potential of a feature to improve predictions if represented accurately.

Interference betwen x_i and x_j occurs when two features are embedded non-orthogonally and, as a result, affect each other's predictions. This prevents superposition in linear models.

(Linear model)

- The deduction can be refered to *Exact solutions to the nonlinear dynamics of learning in deep linear neural networksA.M. Saxe, J.L. McClelland, S. Ganguli. 2014.*

feature benefit vs interference

$$x' = \mathrm{ReLU}(W^T W x + b)$$

$$L = \sum_x \sum_i I_i (x_i - x'_i)^2$$

$$L = \int_x ||I(x - \mathrm{ReLU}(W^T W x + b))||^2 d\mathbf{p}(x)$$

$$L = (1-S)^n L_n + \ldots + (1-S) S^{n-1} L_1 + S^n L_0,$$

$$L_1 = \sum_i \int_{0 \le x_i \le 1} I_i (x_i - \mathrm{ReLU}(||W_i||^2 x_i + b_i))^2 + \sum_{i \neq j} \int_{0 \le x_i \le 1} I_j \mathrm{ReLU}(W_j \cdot W_i x_i + b_j)^2$$

If we focus on the case $x_i = 1$, we get something which looks even more analagous to the linear case:

$$= \sum_i I_i (1 - \mathrm{ReLU}(||W_i||^2 + b_i))^2 + \sum_{i \neq j} I_j \mathrm{ReLU}(W_j \cdot W_i + b_j)^2$$

Feature benefit is similar to before. Note that ReLU never makes things worse, and that the bias can help when the model doesn't represent a feature by taking on the expected value.

Interference is similar to before but ReLU means that negative interference, or interference where a negative bias pushes it below zero, is "free" in the 1-sparse case.

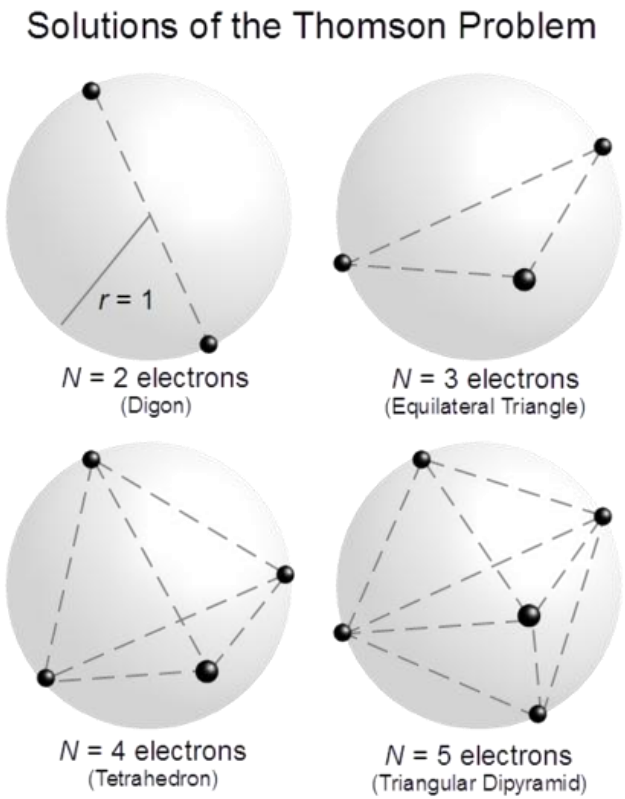
(Non-Linear model)

Similar to famous Thomson problem

$$\sum_i I_i(1 - \mathrm{ReLU}(||W_i||^2 + b_i))^2 \quad + \quad \sum_{i \neq j} I_j \mathrm{ReLU}(W_j \cdot W_i + b_j)^2$$

Feature benefit is similar to before. Note that ReLU never makes things worse, and that the bias can help when the model doesn't represent a feature by taking on the expected value.

Interference is similar to before but ReLU means that negative interference, or interference where a negative bias pushes it below zero, is "free" in the 1-sparse case.

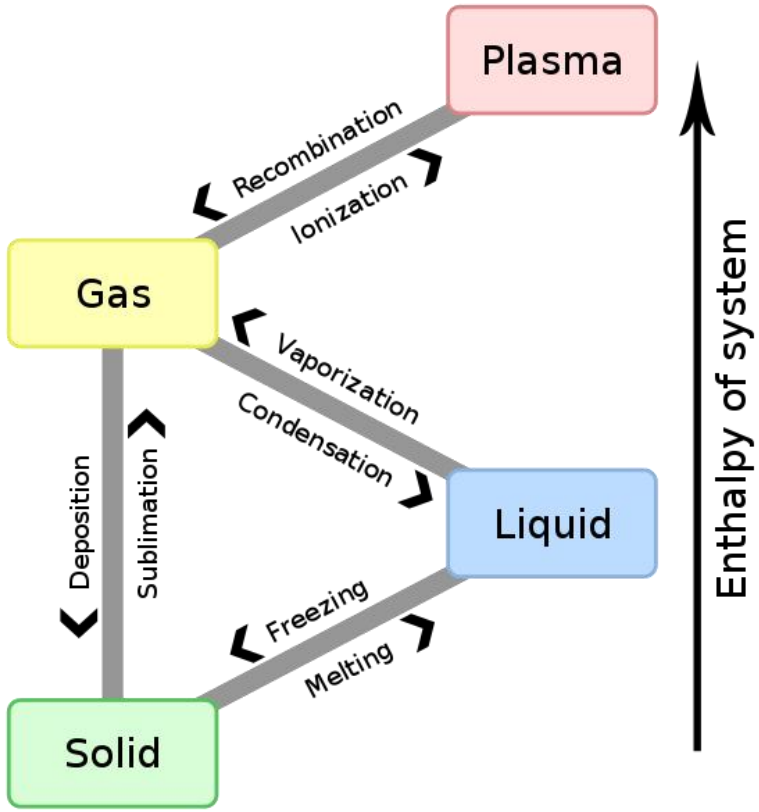


- Uniform importance with a fixed number of features with 1 norm and others with 0 norms => only interference term remains
- points (features); space (hidden dimension)

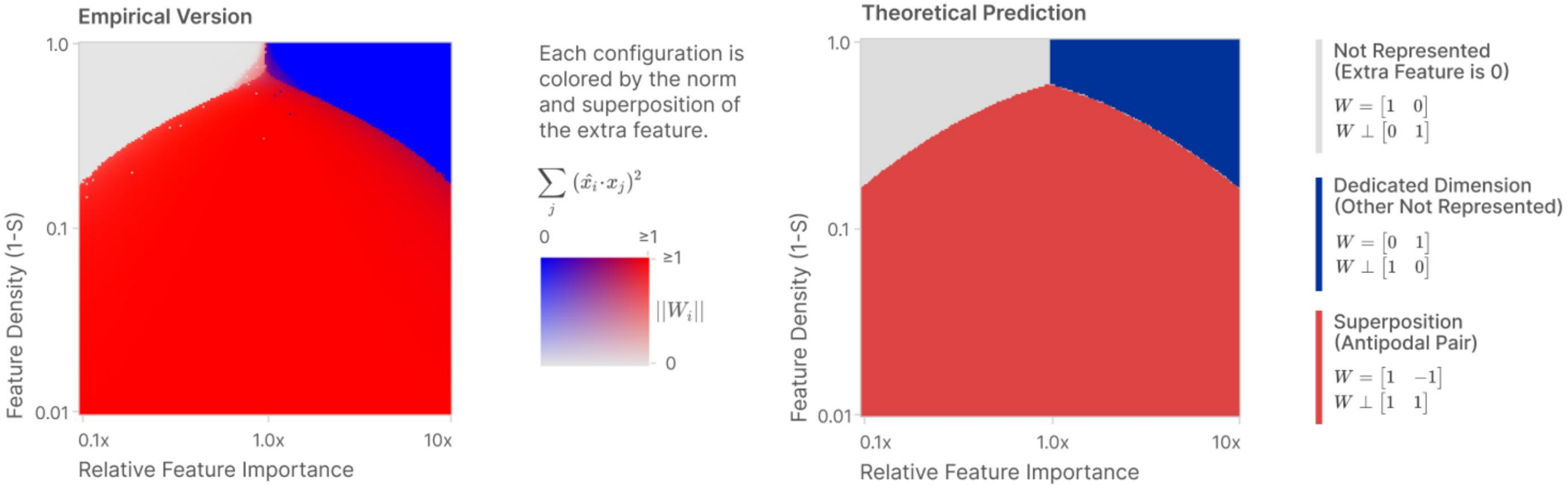
Anything to discuss?

Superposition as a Phase Change

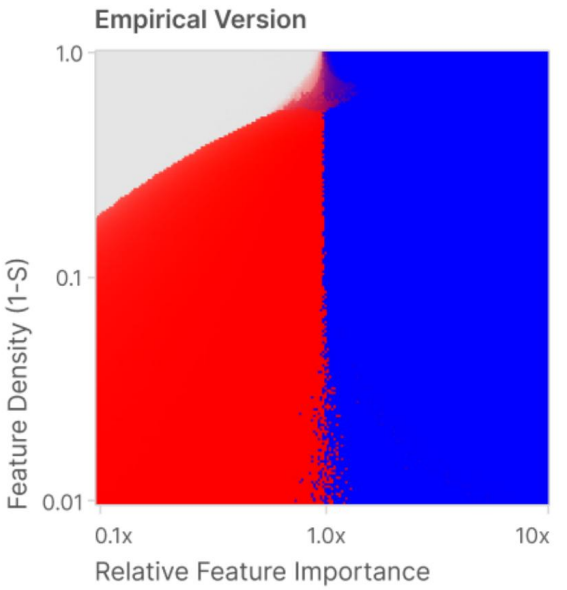
- Three outcomes/phase for a feature
  - simply not be learned
  - learned but represented in superposition
  - represented with a dedicated dimension
- Some kind of phase change
- Experiments
  - To isolate the effects
  - 2 features with 1 dimension
  - ReLU after the linear output
  - Importance: 1 for the first, 0.1-10 for the second (, which is our focus)
  - Sparsity: 1.0 - 0.01



Results (Empirical vs. Theoretical, n=2 m=1)



Results (n=3, m=2)



Each configuration is colored by the norm and superposition of the extra feature.

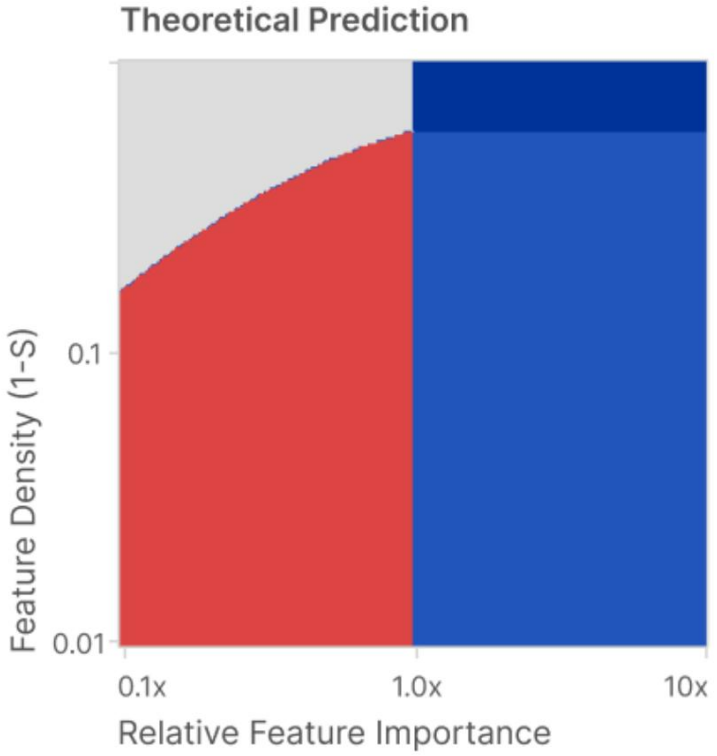
$$\sum_j (\hat{x}_i \cdot x_j)^2$$

0 ≥1

≥1

$||W_i||$

0



Not Represented

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$$

$$W \perp \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

Dedicated Dimension - Other Not Represented

$$W = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$W \perp \begin{bmatrix} 1 & 0 & 0 \end{bmatrix}$$

Superposition

$$W = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -1 \end{bmatrix}$$

$$W \perp \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$$

Dedicated Dimension - Others in Superposition

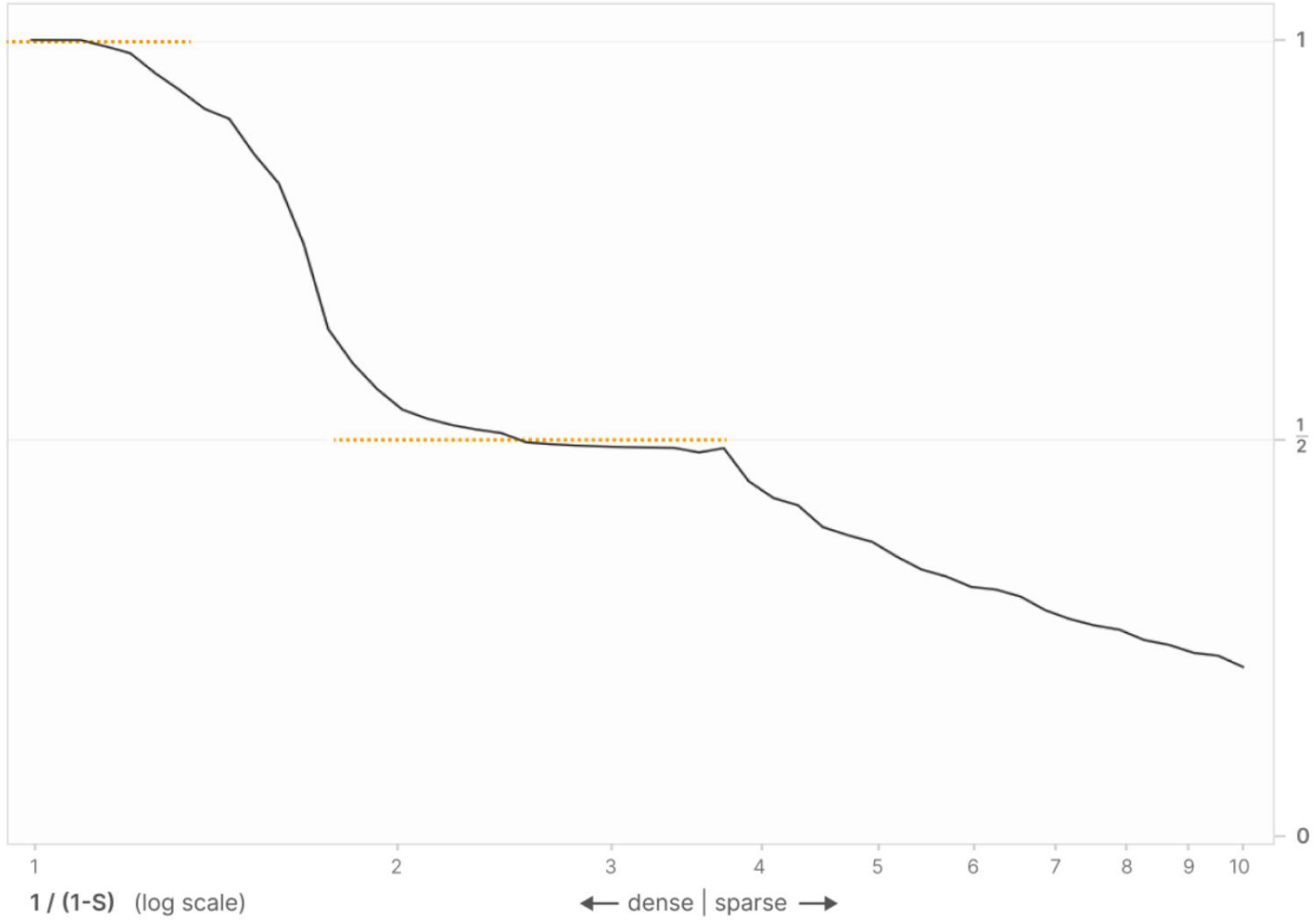
$$W = \begin{bmatrix} 1 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$W \perp \begin{bmatrix} 1 & 1 & 0 \end{bmatrix}$$

The Geometry of Superposition

- Inspired by Thomson problem
- Uniform Superposition: all features have the same importance and sparsity
- n=400, m=30 (n >> m is just okay)
- Number of represented feature: $||W||_F^2$
- “Dimensions per feature”: $D^* = m/||W||_F^2$

The Geometry of Superposition



Number of Hidden Dimensions per Embedded Feature

$m/||W||_F^2$

Unexpectedly, we find that the number of dimensions per feature is "sticky" at 1 and 1/2. We denote this with ····· .

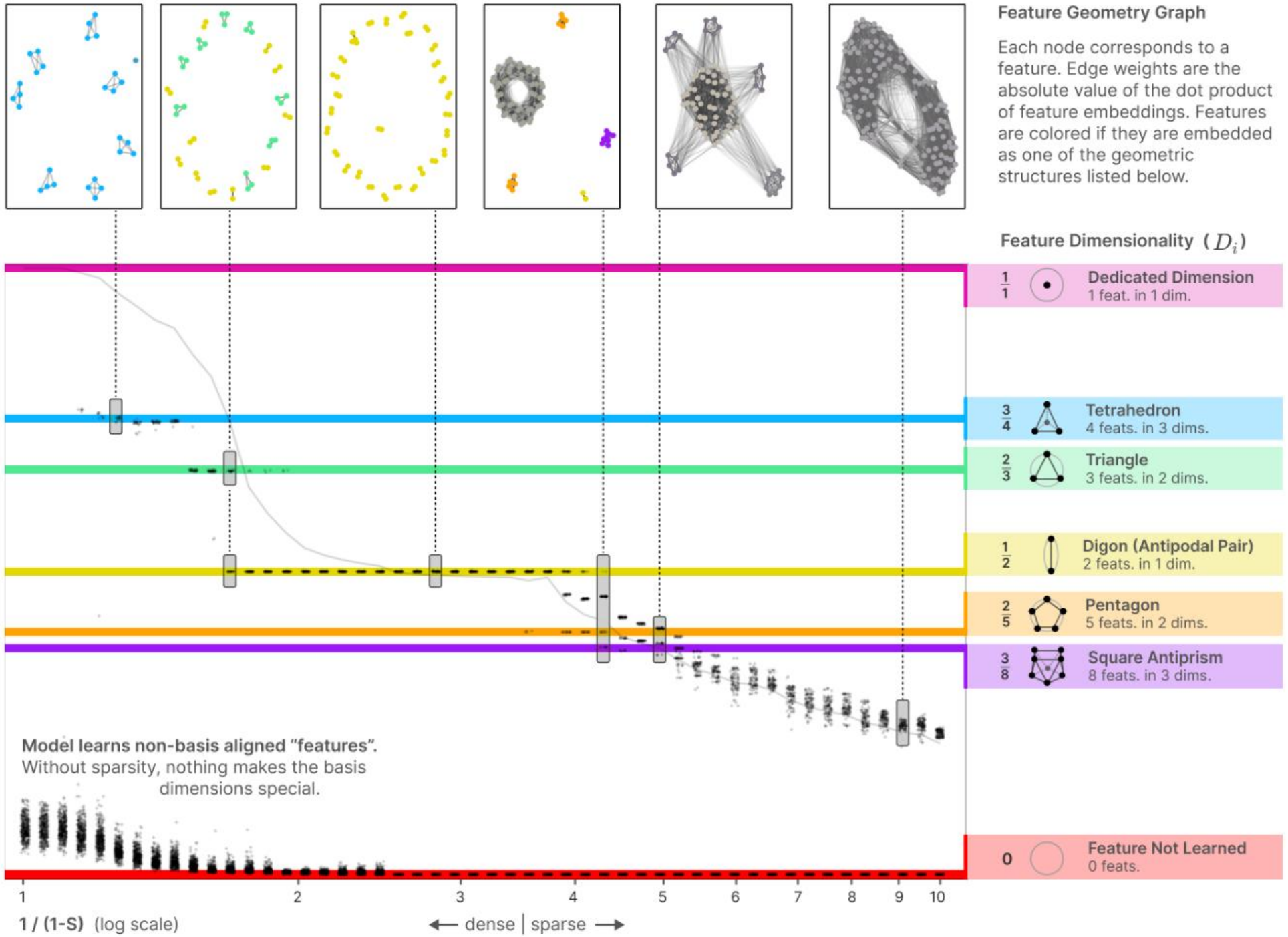
Dimensionality

- Dimensionality for a feature i
- an antipodar pair [1, -1] has 1/2 of D for each feature
- D=0 for feature not learned
- D of all features sum up to 1.

$$D_i = \frac{||W_i||^2}{\sum_j (\hat{W}_i \cdot W_j)^2}$$

Geometry (many sticky points)

- We start with the line plot we had in the previous section.
- We overlay this with a scatter plot of the individual feature dimensionalities for each feature in the models at each sparsity level.
- The feature dimensionalities cluster at certain fractions, so we draw lines for those. (It turns out that each fraction corresponds to a specific weight geometry – we'll discuss this shortly.)
- We visualize the weight geometries for a few models with a "feature geometry graph" where each feature is a node and edge weights are based on the absolute value of the dot product feature embedding vectors. So features are connected if they aren't orthogonal.

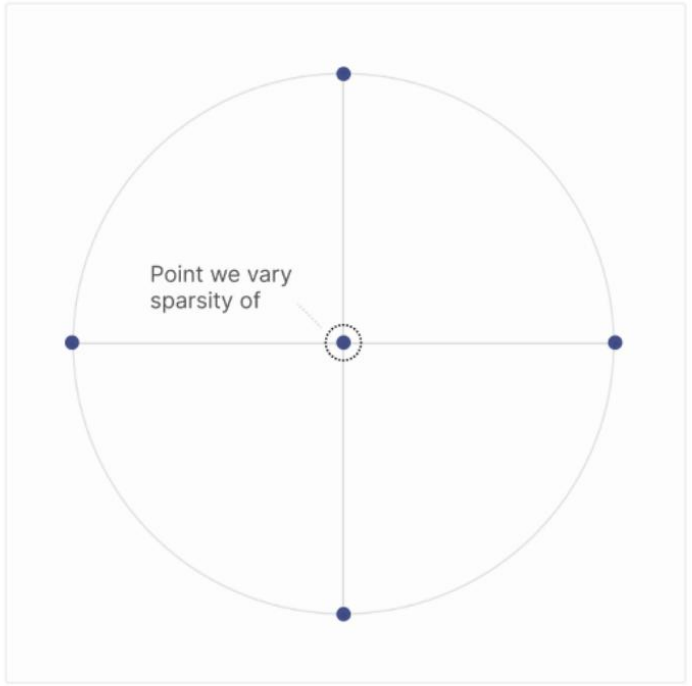

Feature Geometry Graph
Each node corresponds to a feature. Edge weights are the absolute value of the dot product of feature embeddings. Features are colored if they are embedded as one of the geometric structures listed below.
Feature Dimensionality (D_i)
1/1 Dedicated Dimension
1 feat. in 1 dim.
3/4 Tetrahedron
4 feats. in 3 dims.
2/3 Triangle
3 feats. in 2 dims.
1/2 Digon (Antipodal Pair)
2 feats. in 1 dim.
2/5 Pentagon
5 feats. in 2 dims.
3/8 Square Antiprism
8 feats. in 3 dims.
0 Feature Not Learned
0 feats.
Model learns non-basis aligned "features".
Without sparsity, nothing makes the basis dimensions special.
1
2
3
4
5
6
7
8
9
10
1 / (1-S) (log scale)
← dense | sparse →

Non-Uniform Superposition

- Features varying in importance or sparsity
- Correlated features
- Anti-correlated features

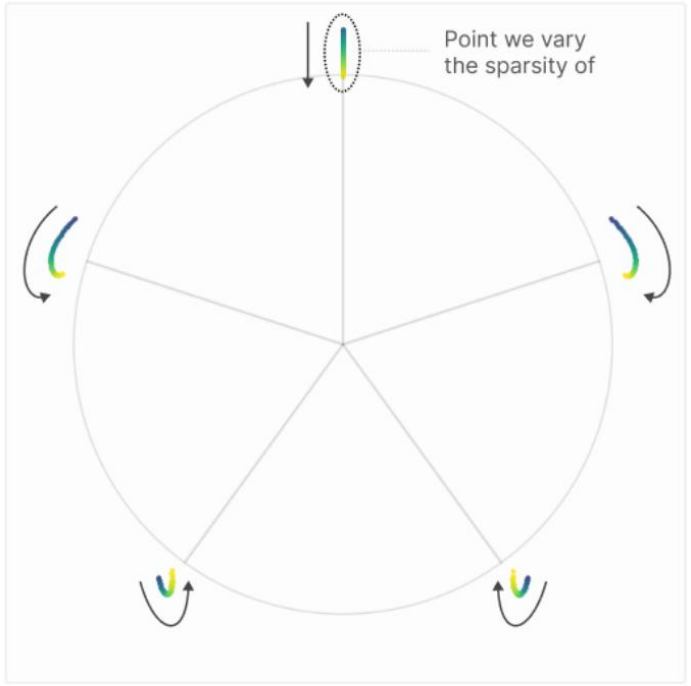
Features varying in sparsity

Digon (Square) Solutions



When the sparsity of the varied point falls below a certain critical threshold (~2.5x less than others) the pentagon solution changes to two digons.

Pentagon Solutions



Note how vertices shift as sparsity changes

To study non-uniform sparsity, we consider models with five features, varying the sparsity of a single feature and observing how the resulting solutions change. We observe a mixture of continuous deformation and sharp phase changes.

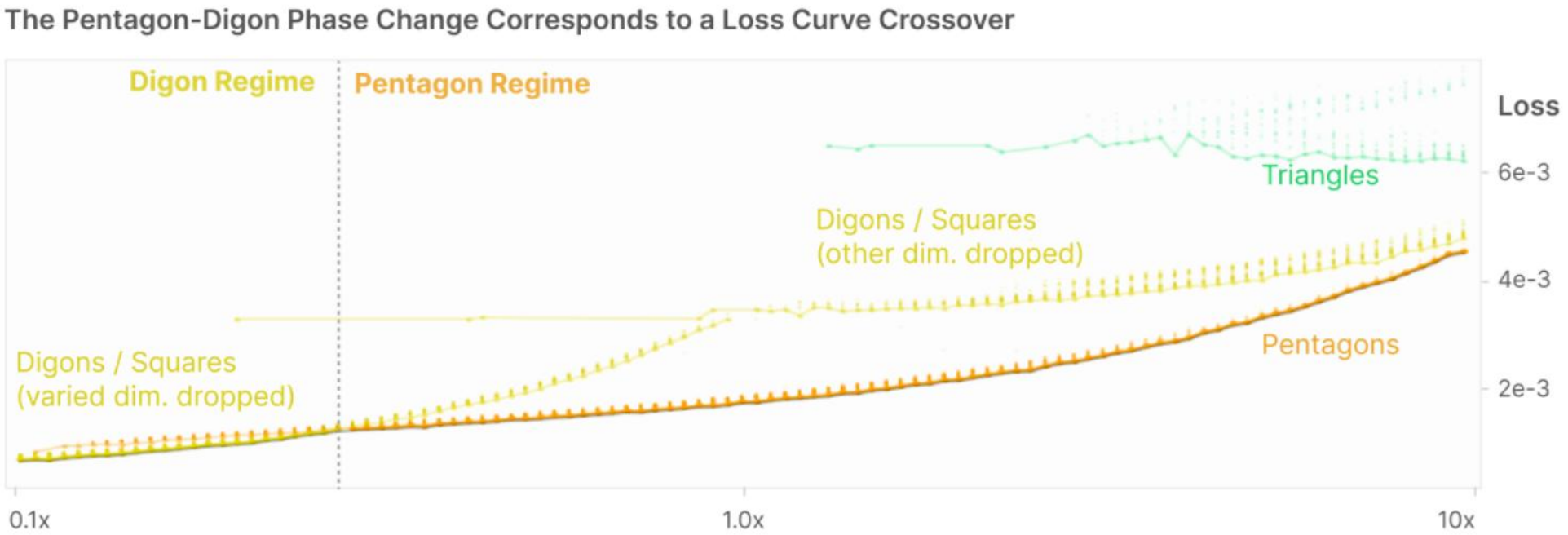
Parameters

n = 5
m = 2
I_i = 1
$1\text{-}S$ = 0.05 (baseline)

Relative Feature Density (1-S)

0.1x 1.0x 10x
sparser denser

Pentagon-Digon Phase Change

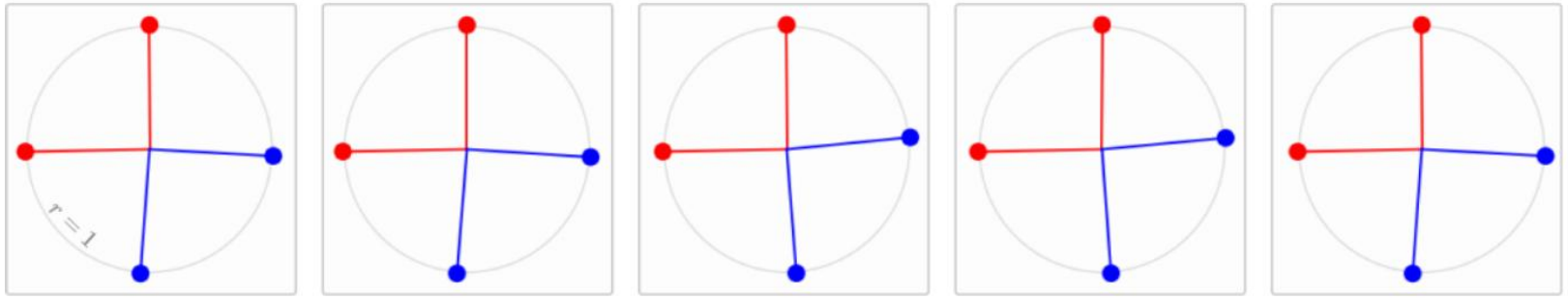


Gradient descent has trouble moving between solutions associated with different geometries. As a result, fitting the model will often produce non-optimal solutions. By characterizing and plotting these, we can see that each geometry creates a different loss curve, and that the pentagon-digon phase change corresponds to a cross over between the curves.

Correlated and Anticorrelated Features

Models prefer to represent correlated features in orthogonal dimensions.

We train several models with 2 sets of 2 correlated features (n=4 total) and a m=2 hidden dimensions. We then visualize the weight column for each feature. For ease of comparison, we rotate and flip solutions to have a consistent orientation.

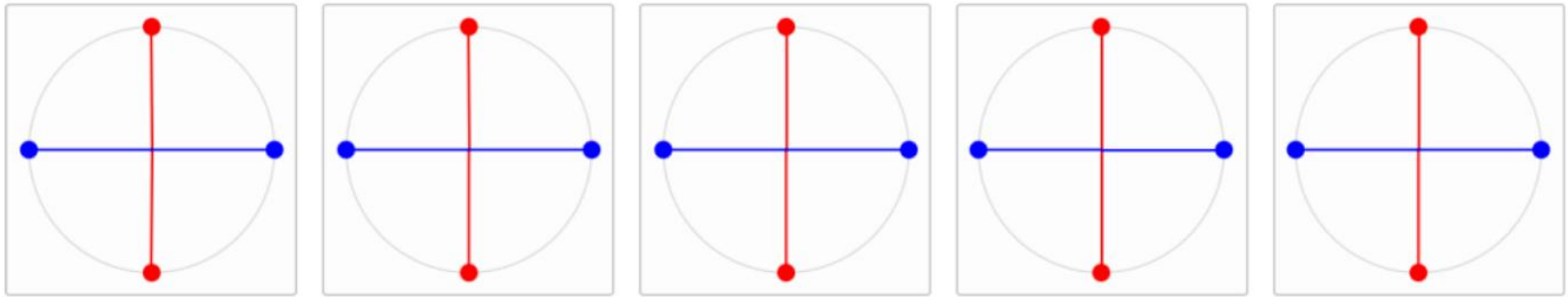


●● and ●● denote **correlated** feature sets.

Correlated feature sets are constructed by having them always co-occur (ie. be zero or not) at the same time.

Models prefer to represent anticorrelated features in opposite directions.

We train several models with 2 sets of 2 anticorrelated features (n=4 total) and a m=2 hidden dimensions. We then visualize the weight column for each feature. For ease of comparison, we rotate and flip solutions to have a consistent orientation.



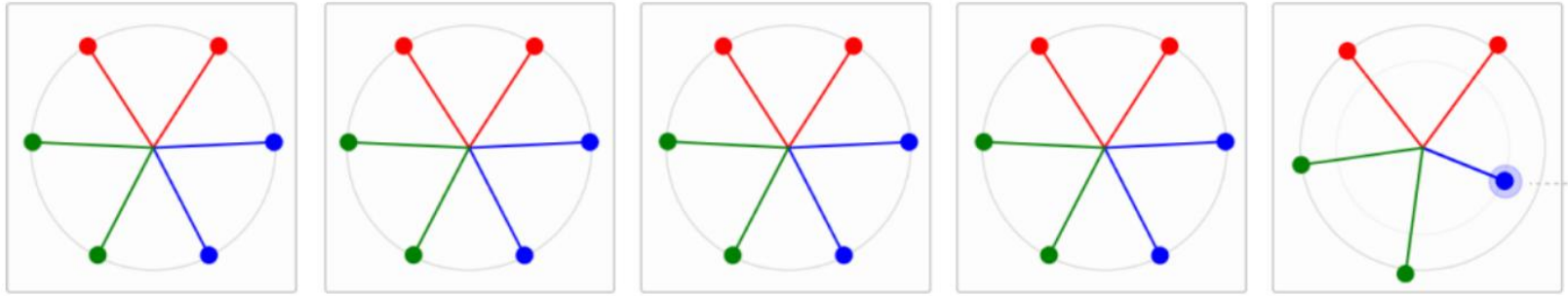
●● and ●● denote **anticorrelated** feature sets.

Anticorrelated feature sets are constructed by having them never co-occur (ie. be zero or not) at the same time.

Correlated and Anticorrelated Features

Models prefer to arrange correlated features side by side if they can't be orthogonal.

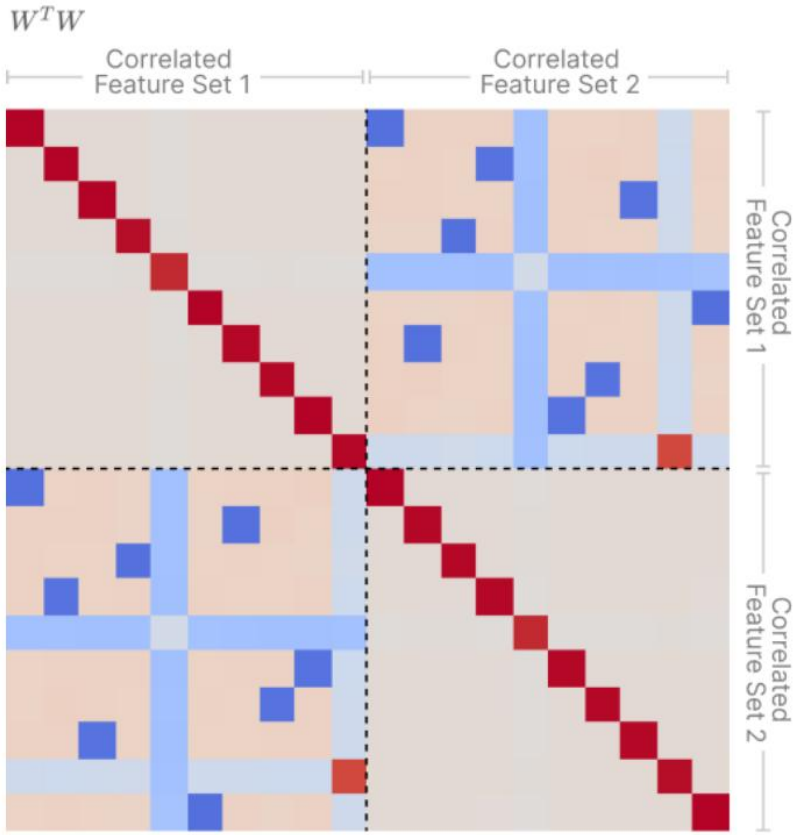
We train several models with 3 sets of 2 correlated features (n=6 total) and a m=2 hidden dimensions. We then visualize the weight column for each feature. For ease of comparison, we rotate and flip solutions to have a consistent orientation. (Note that models will not embed 6 independent features as a hexagon like this.)



●●, ●●, and ●● denote **correlated** feature sets.

Sometimes correlated feature sets "collapse". In this case it's an optimization failure, but we'll return to it shortly as an important phennomenon.

Local Almost-orthogonal bases



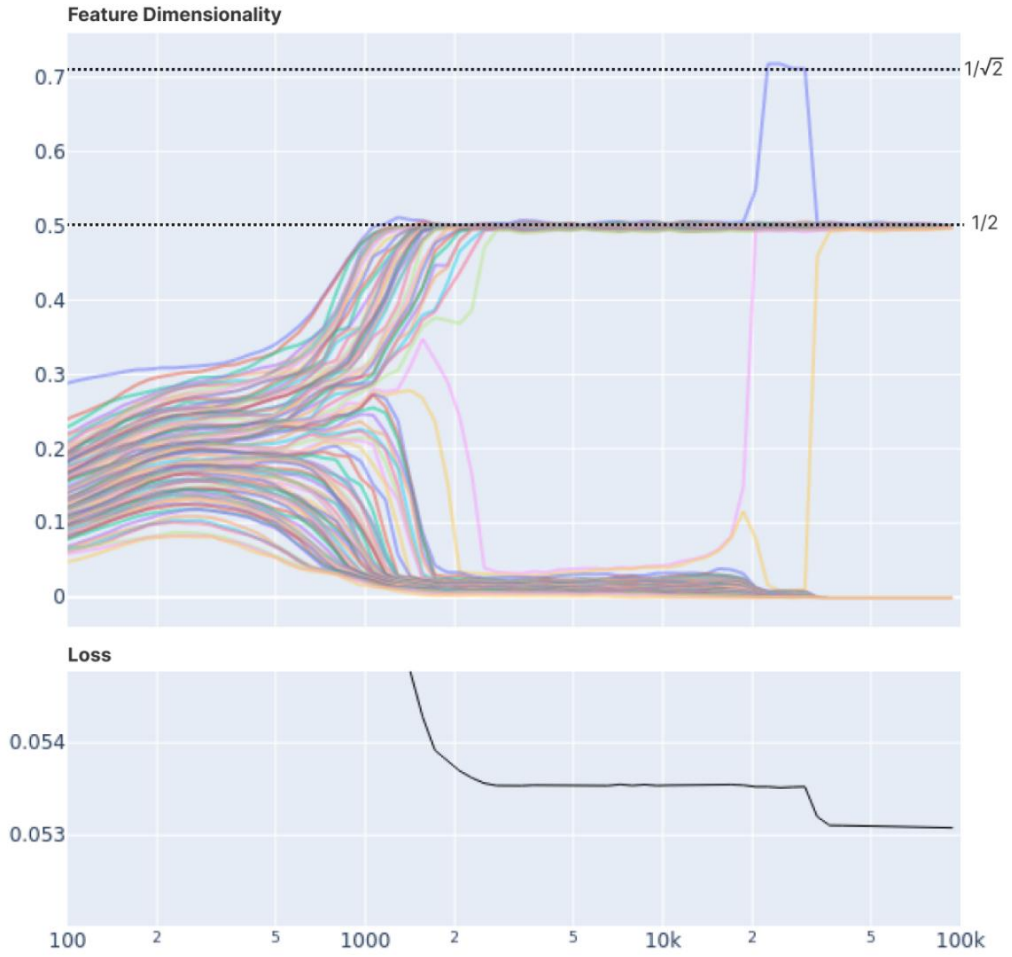
Models prefer to represent correlated features in orthogonal dimensions, creating "local orthogonal bases".

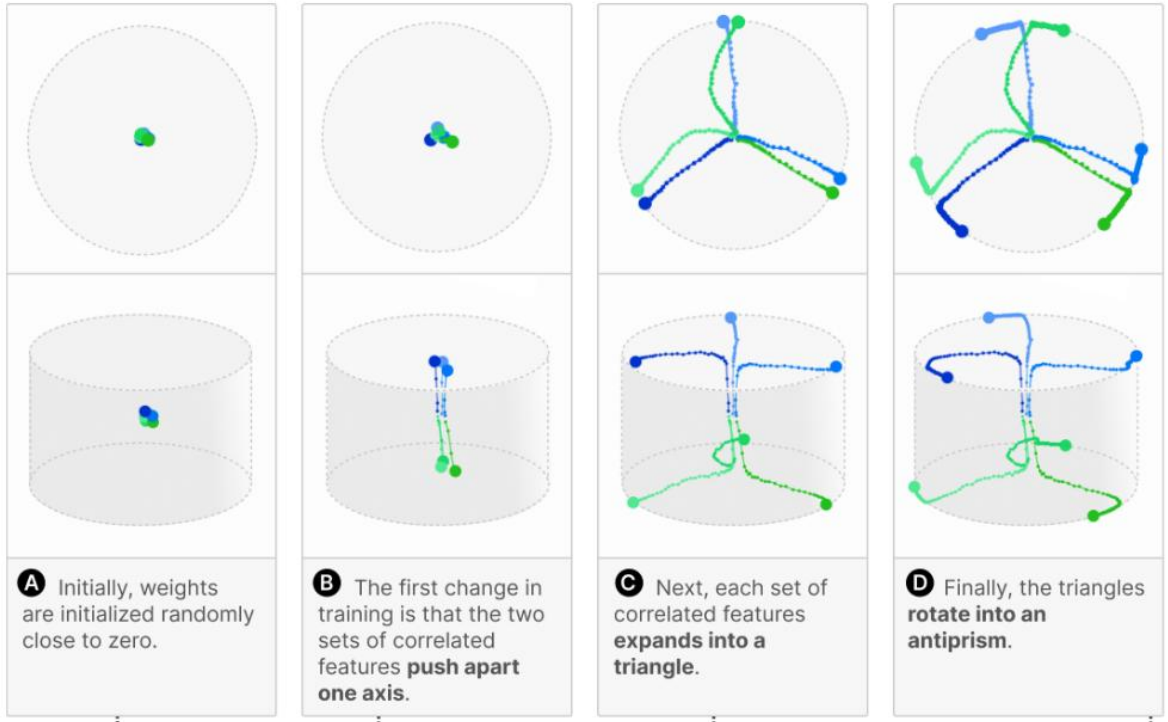
We train a model with 2 sets of 10 correlated features (n=20 total) with m=10 hidden dimensions.

Within each set of correlated featuers, the model creates a *local orthogonal basis*, having each feature be represented orthogonally.

Collapsing of correlated features

Superposition and Learning Dynamics





Feature Weight Trajectories (top and 3D perspecitve)

●●● and ●●● denote correlated feature sets.

Note that the resulting triangular antiprism is equivelant to a octahedron, with features forming antipodal pairs with features from a different correlated feature set.

Loss Curve

The loss curve goes through several distinct regimes corresponding to different geometric transformations of the weights (as seen above).

Other parts...

- Relationship to Adversarial Robustness
- Superposition in a priviledged basis
- Computation in Superposition

Conclusions

- Definitions of related concepts
- Superposition is observed in non-linear models with sparse features (tradeoff between feature benefit and interference)
- Mathematical Understanding
- Visualization: A phase change
- Geometry of feature directions (uniform vs. non-uniform)
- Omitted parts

清華大學
Tsinghua University

Q & A

THANK YOU

Note