# Singular Vectors of Attention Heads Align with Features

**Gabriel Franco · Carson Loughridge · Mark Crovella**

BOSTON UNIVERSITY

ICML International Conference On Machine Learning

**Recent studies infer features from the singular vectors of attention matrices. Why and when do singular vectors align with features?**

**Our answer:** singular vectors **robustly** align with features in a toy model where features can be directly observed; theory shows the top singular vectors are the **covariance-whitened features** of interest, even with anisotropy; and alignment implies a testable prediction, **sparse attention decomposition**, which emerges in real models.

## What is SVF alignment?

An attention head scores a (query $r$, key $s$) token pair with a bilinear form of its weights:

$$logit(r, s) = r^T \Omega s, \quad \Omega = W_Q{}^T W_K = U \Sigma V^T$$

The SVD splits the head into paired directions ($u_k$, $v_k$): each term is a **communication channel**, active only when both tokens project onto it:

$$logit(r, s) = \Sigma_k \, (r \cdot u_k) \, \sigma_k \, (v_k \cdot s)$$

**SVF alignment:** a feature has high cosine similarity to a singular vector of the head's QK matrix. Singular vectors then become **candidate features, read from the weights**.

## Alignment happens robustly in a toy model

We add an attention head to the toy features model of Elhage et al. (2022) and train it to attend when specific **feature pairs** are present (the **features of interest**), so that features and singular vectors can be directly observed.

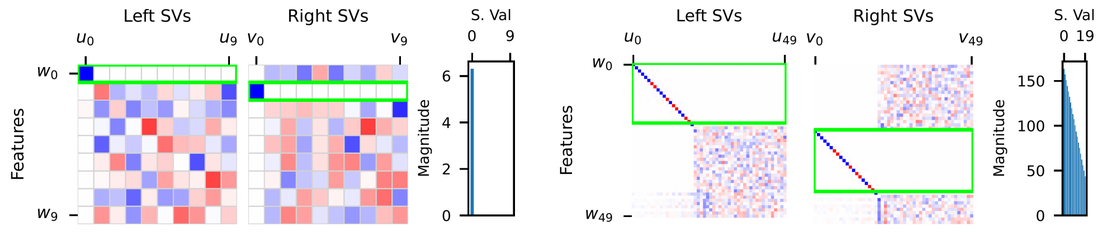



*Fig. 2: singular vectors align with features (green boxes). Left: a single feature pair of interest. Right: 100 features of which $w_0$–$w_{39}$ are of interest (only an initial subset is shown).*

- A single feature pair to be learned by attention → the **top singular vectors** align with it.
- Multiple feature pairs → each aligns with its own singular vector, and features **orthogonalize** to minimize interference.

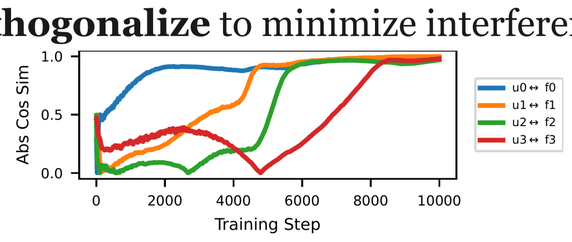
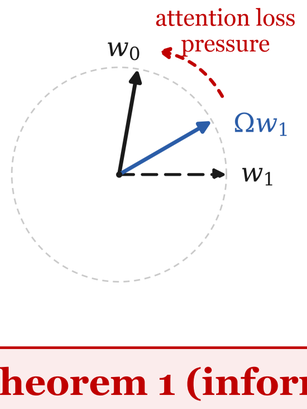



*Fig. 3: alignment of singular vectors with features during training; alignment occurs for highest-logit features first.*

## Why? Theory predicts alignment

**1. Vector-space geometry**




**2. Mathematical consequence**

To maximize the attention score, $\Omega$ must align the features:

$$\Omega w_1 \approx \alpha_o w_o \quad \text{and} \quad \Omega^T w_o \approx \alpha_1 w_1$$

Chaining both directions gives:

$$\Omega^T \Omega w_1 \approx \alpha w_1 \quad \Rightarrow \quad w_1 \approx v_o$$

**Theorem 1 (informal).** *After training, the head's QK matrix is low-rank, and its top singular vectors equal the covariance-whitened features of interest.*

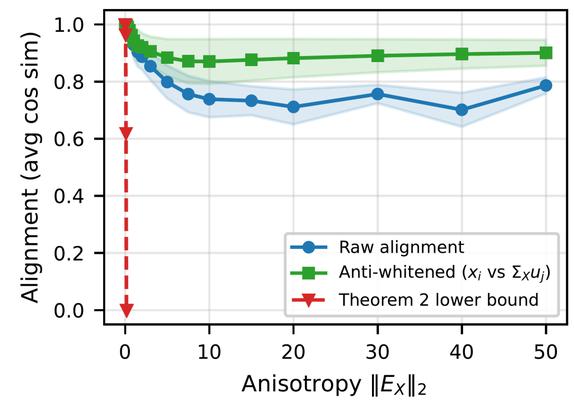
- **Isotropic features → exact alignment** (Corollary 1).
- Anisotropy at the levels found in GPT-2 does not destroy alignment (Thm. 2; Fig. 4).
- Orthogonalizing features of interest minimizes interference (Thm. 3).

## Take-home messages

- Top singular vectors equal the **covariance-whitened features** of interest, and alignment **occurs even with anisotropy** (Thm. 1, Fig. 4).
- Training a head **aligns its singular vectors** with the feature pairs it attends to.
- Top singular vectors align with the feature pair **most important for computing attention** (highest target logit).
- It predicts **sparse attention decomposition**, confirmed in GPT-2 & Pythia.

## Singular vectors are whitened features of attention

The top singular vectors are given by the **covariance-whitened features**: $u_1 \propto \Sigma_X^{-1} x_1$ (Thm. 1). To test this, we measure the **cosine similarity between each true feature and its singular vector**, introducing controlled anisotropy over the range of values found in GPT-2:



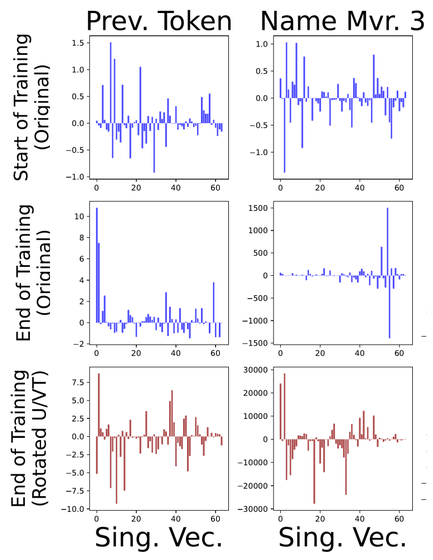

**mean cosine similarity**
**> 0.75**
even at the largest anisotropy seen in GPT-2

**anti-whitening** with $\Sigma_X$ raises it to **≈ 0.9**

*Fig. 4: SVF alignment under anisotropic features, varied over the range of anisotropy observed in GPT-2.*

## The prediction: sparse attention decomposition

In real models, features are not directly observable. But alignment makes a testable prediction: when a head attends **because a few features of interest are present**, its relative attention should concentrate on **a few singular directions**.




Pythia-160M, IOI heads, 130 training checkpoints: attention decomposition **sparsifies during training**, mirroring the toy model.

Randomly rotating $U$ and $V$ preserves the spectrum but eliminates sparsity: SAD is due to **specific singular directions**, not simply a few large singular values.

GPT-2 / IOI: typically only **1–4 singular vectors** reconstruct a head's relative attention (Fig. 9).

*Fig. 8: Pythia decompositions for two IOI heads, early vs. late in training, and with randomly rotated singular vectors, where sparsity does not arise.*

## Why it matters

Singular vectors provide a **comprehensible and tractable search space** for extracting features: decompose activations in the SVD basis, per prompt, in a **single forward pass**. **Far more efficient and scalable than SAEs or linear probes.**

**Code & experiments:** github.com/gaabrielfranco/svf-alignment • Contact: gvfranco@bu.edu • Paper: scan the QR code above