On the Limits of Sparse Autoencoders: A Theoretical Framework and Reweighted Remedy

Jingyi Cui [1] Qi Zhang [1] Yifei Wang [2] Yisen Wang [1]
1 Peking University; 2 Amazon AGI SF Lab
Correspondence to yisen.wang@pku.edu.cn

Background

- Sparse autoencoders (SAEs) are widely used to interpret internal features of LLMs and VLMs.
- Under the superposition hypothesis, observed polysemantic features are linear mixtures of latent monosemantic features.
- **Key question**: *when does reconstructing polysemantic inputs actually recover the ground-truth monosemantic features?*

Our Contributions

- First closed-form theoretical analysis of SAE feature recovery under the superposition hypothesis.
- Standard SAEs generally fail via feature shrinking and feature vanishing unless the true features are extremely sparse.
- Propose a reWeighted SAE (WSAE) and derive a principled weight-selection rule that improves ground-truth reconstruction.
- Validate the theory on synthetic data and on pretrained language and vision models.

Mathematical Formulations

Notations. x: ground-truth monosemantic feature; x_p: superposed polysemantic feature; x_m: SAE latent. We assume $n > n_p$ and $n_m > n_p$

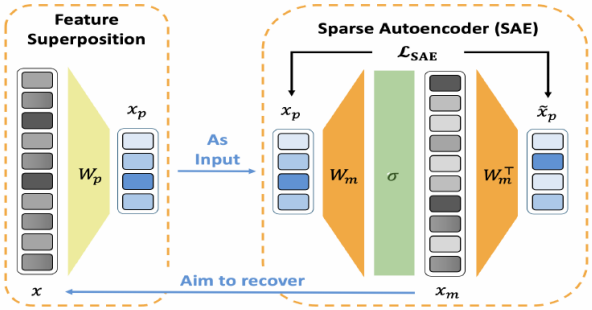
Superposition of features $x_p = W_p x$

Sparse autoencoder $x_p = W_p x \qquad \tilde{x}_p = W_m^\top x_m$

SAE reconstruction $\mathcal{L}_{\text{SAE}}(W_m; x_p) = \mathbb{E}_{x_p}\|x_p - \tilde{x}_p\|^2 \quad W_m^* = \arg\min_{W_m} \mathcal{L}_{\text{SAE}}(W_m; x_p)$

Feature recovery $x_m \sim x$

$\sim$: equivalence in sense of reordering and zero-padding



Sparsity assumption. Each dimension $x_i > 0$ with probability $1 - S$ and $x_i > 0$ with probability S. Extreme sparsity indicates $S \to 1$.

SAE FAILS UNLESS THE GROUND TRUTH IS EXTREMELY SPARSE

Main Theoretical Results

Theorem 1 (Closed-Form Solution to SAEs). *If $n_m \geq n$ and the columns of W_p within the superposed dimensions form digons/polygons, then we have*

$$W_m^* = I^*(W_p, \mathbf{0})^\top \in \arg\min_{W_m} \mathcal{L}_{\text{SAE}}(W_m; x_p).$$

✅ SAE has closed-form solution $W_m^* \sim W_p^\top$

Theorem 2 (Optimality under extreme sparsity). *For $n_m \geq n$, and the columns of W_p have non-positive interferences, if $S \to 1$, then for arbitrary x, we have*

$$x_m \sim x$$

✅ SAE recovers ground truth x when it is extremely sparse ($S \to 1$)
❌ but SAE fails in general cases

Failure Example 1 (Feature shrinking)

$$\begin{bmatrix}0.5\\1.0\\0.8\end{bmatrix} \xrightarrow{\text{superposition}} \begin{bmatrix}0.5\\0.2\end{bmatrix} \xrightarrow{\text{SAE}} \begin{bmatrix}0.5\\0.2\\0\end{bmatrix} \qquad \left(W_p = \begin{bmatrix}1 & 0 & 0\\0 & 1 & -1\end{bmatrix}\right)$$

❌ The superposed feature dimensions shrink after SAE recovery.
❌ The top activated dimension changes because of feature shrinking.

Failure Example 2 (Feature vanishing)

$$\begin{bmatrix}0.7\\0.5\\0.3\end{bmatrix} \xrightarrow{\text{superposition}} \begin{bmatrix}0.1\sqrt{3}\\0.3\end{bmatrix} \xrightarrow{\text{SAE}} \begin{bmatrix}0.3\\0\\0\end{bmatrix} \qquad \left(W_p = \begin{bmatrix}0 & \sqrt{3}/2 & -\sqrt{3}/2\\1 & -1/2 & -1/2\end{bmatrix}\right)$$

❌ When feature shrinking is severe enough, some features even completely vanish.

Reconstruction Gap

Ideally, the goal is to optimize the ground truth reconstruction loss

$$\mathcal{L}_{\text{GT}}(W_m; x) = \mathbb{E}_x\|x - x_m\|^2 = \mathbb{E}_x\|x - \sigma(W_m W_p x)\|^2 = \mathbb{E}_x\|x - \sigma(W_m W_p x)\|^2.$$

There is a gap between SAE reconstruction and GT reconstruction

Theorem 4 (Gap between $\mathcal{L}_{\text{SAE}}$ and $\mathcal{L}_{\text{GT}}$). *When $W_m = W_p^\top$, we have*

$$\mathcal{L}_{\text{SAE}}(W_m; x_p) - \mathcal{L}_{\text{GT}}(W_m; x) = [x - \sigma(W_p^\top W_p x)]^\top (W_p^\top W_p - I_{n\times n})[x - \sigma(W_p^\top W_p x)].$$

Reweighted Remedy (WSAE)

Narrow the reconstruction gap by adding weights.

$$\mathcal{L}_{\text{WSAE}}(W_m; x_p) = \mathbb{E}_{x_p}\|\Gamma[x_p - W_m^\top \text{ReLU}(W_m x_p)]\|_2^2, \quad \Gamma = \text{diag}(\gamma_1, \ldots, \gamma_{n_p}).$$

Theorem 5 (Gap between $\mathcal{L}_{\text{WSAE}}$ and $\mathcal{L}_{\text{GT}}$). *When $W_m = W_p^\top$, we have*

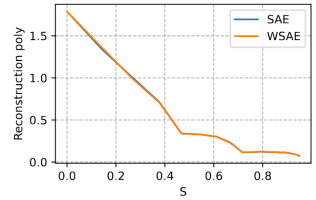
$$\mathcal{L}_{\text{WSAE}}(W_m; x_p) - \mathcal{L}_{\text{GT}}(W_m; x) = [x - \sigma(W_m W_p x)]^\top (W_p^\top \Gamma^\top \Gamma W_p - I_{n\times n})[x - \sigma(W_m W_p x)].$$

$$W_p^\top \Gamma^\top \Gamma W_p - I_{n\times n} = \begin{bmatrix} \gamma_1^2 - 1 & \cdots & \gamma_1^2 W_{p,[:,1]}^\top W_{p,[:,n]} \\ \cdots & & \cdots \\ \gamma_n^2 W_{p,[:,n]}^\top W_{p,[:,1]} & \cdots & \gamma_n^2 - 1 \end{bmatrix}$$

- Assign weights near 1 to monosemantic dimensions (small off-diagonal interference).
- Assign smaller weights to polysemantic dimensions to suppress negative interference.
- In experiments, variance/semantic consistency serves as proxy for monosemanticity.

Experimental Verification

(a) Reconstruction error on the non-sparse dimensions of the ground truth monosemantic features, showing a greater error gap between WSAE and SAE.



(b) The reconstruction error of the polysemantic features x_p, where the errors of the two methods are comparable.

Table 1: Auto-interpretability scores (%) of SAEs trained following different layers (0-11) of Pythia-160M with original SAE and weighted SAE loss. SAEs trained with weighted SAE loss obtain higher auto-interpretability scores (i.e., stronger monosemanticity) across different situations.

	0	1	2	3	4	5	6	7	8	9	10	11
Original SAE	74.7	74.1	76.7	77.8	78.5	79.5	79.3	77.8	74.6	75.6	71.6	72.5
Weighted SAE (α=0.5)	75.4	77.6	76.4	77.9	79.3	79.6	79.8	77.4	78.6	**79.3**	**76.1**	72.6
Gains	+0.7	+3.5	-0.3	+0.1	+0.8	+0.1	+0.5	-0.4	+4.0	**+3.7**	**+4.5**	+0.1
Weighted SAE (α=1)	**77.2**	**78.9**	**81.3**	**84.6**	**83.9**	**83.3**	**83.9**	**79.6**	**81.5**	77.6	72.4	**73.5**
Gains	**+2.5**	**+4.8**	**+4.6**	**+6.8**	**+5.4**	**+3.8**	**+4.6**	**+1.8**	**+6.9**	+2.0	+0.8	**+1.0**