Progressive Feature Separation in Sparse Autoencoders via Cubic Scaling Orthogonality

Anonymous authors
Paper under double-blind review

Abstract

Understanding and controlling internal representations in large language models is crucial for ensuring their reliability and safety, yet existing methods struggle to isolate specific knowledge without degrading overall performance. We introduce a novel sparse autoencoder architecture that achieves targeted feature separation through progressive orthogonality constraints, addressing the fundamental challenge of balancing feature isolation with reconstruction quality. Our key innovation is a cubic scaling mechanism that smoothly adjusts separation pressure based on feature correlations, combined with activation-based weighting to focus on frequently co-activated features. Experiments on the Gemma-2-2b model demonstrate that our approach reduces reconstruction loss by 3.8% (from 200.23 to 192.60) while maintaining strong downstream performance (91.5% vs 95.1% on sparse probing tasks). We achieve this through careful engineering of the separation mechanism, including optimal batch sizing (4096) for stable ratio estimation and warmup scheduling aligned with learning rate progression. These results establish a new direction for controlled knowledge manipulation in language models, with implications for interpretability, bias mitigation, and targeted model editing.

1 Introduction

The rapid advancement of large language models has brought unprecedented capabilities OpenAI (2024), but also heightened concerns about their interpretability and controllability. While these models excel at diverse tasks, their distributed representations make it challenging to isolate and modify specific knowledge without unintended consequences. This limitation poses significant risks for deployment, particularly in safety-critical applications where precise control over model behavior is essential.

Sparse autoencoders (SAEs) have emerged as a promising tool for interpreting neural representations Goodfellow et al. (2016), but current implementations face fundamental challenges in achieving targeted feature separation. The core technical difficulty lies in balancing two competing objectives: maintaining high-quality reconstructions while enforcing meaningful separation between features. Our experiments reveal that naive approaches to feature separation can increase reconstruction loss by up to 15.9%, demonstrating the need for more sophisticated methods.

The challenge becomes particularly acute as model complexity increases Vaswani et al. (2017). Traditional approaches either sacrifice reconstruction quality for separation or achieve separation at the cost of model performance. This trade-off is evident in our baseline experiments, where aggressive feature separation led to a 214.41 reconstruction loss compared to the baseline 200.23.

This paper makes the following contributions:

- A novel cubic scaling mechanism for orthogonality penalties that provides smooth transitions in feature separation, reducing reconstruction loss from 200.23 to 192.60 while maintaining downstream performance
- An activation-based scaling approach that dynamically adjusts separation pressure based on feature co-activation patterns, preserving natural relationships between independent features
- A learning rate-aligned warmup schedule that allows initial feature development before introducing separation constraints

- Comprehensive empirical validation showing our method maintains strong performance on downstream tasks (91.5% vs 95.1% on sparse probing) while achieving better feature separation

Through systematic experimentation on the Gemma-2-2b model, we demonstrate consistent improvement from baseline (200.23) through quadratic scaling (193.48) to our final cubic scaling implementation (192.60). Our method achieves this while maintaining strong performance across diverse tasks, including sentiment analysis (93.55% vs 98.15

These results establish a foundation for more precise control over neural representations, with immediate applications in model interpretation, bias mitigation, and targeted knowledge editing. Looking ahead, our work opens new directions for research in dynamic penalty scaling, attention-specific separation mechanisms, and architectural innovations to overcome the current limitations in knowledge isolation.

2 Related Work

Several approaches have been proposed for achieving interpretable feature representations in neural networks. The β-VAE framework Higgins et al. (2017) and InfoGAN Chen et al. (2016) achieve disentanglement through training-time constraints, but these methods are not directly applicable to post-training analysis of language models. While they demonstrate the value of progressive constraints for feature separation, our method differs by operating on pre-trained representations and using activation-based scaling rather than information-theoretic objectives.

Recent work on sparse autoencoders for language model interpretation Lan et al. (2024) has shown that similar features emerge across different architectures, suggesting fundamental patterns in learned representations. However, their approach focuses on identifying universal features rather than controlled separation. Our cubic scaling mechanism builds on their insights while providing finer control over feature relationships, as demonstrated by our improved reconstruction loss (192.60 vs 200.23 baseline).

Theoretical foundations for sparse coding Arora et al. (2015) have established optimization guarantees, but these approaches typically assume independence between features. In contrast, our method explicitly models feature correlations through ratio-based ranking and progressive orthogonality constraints. This is particularly important for language models where feature interactions capture complex linguistic patterns Makelov et al. (2024).

The challenge of balancing feature isolation with performance preservation has been explored in biological systems Flesch et al. (2021), where neural populations project task representations onto orthogonal manifolds. While their findings inspired our approach, we extend beyond simple orthogonality by introducing activation-based scaling and cubic penalties. Recent work on feature disentanglement Zhou et al. (2024) uses similar progressive constraints but lacks our ratio-based targeting mechanism, leading to less precise feature separation.

Our approach differs from traditional regularization methods Suteu & Guo (2019) by dynamically adjusting separation pressure based on feature activation patterns. This builds on early work in self-organizing neural networks Sirosh (1995) while introducing modern optimization techniques and scale-appropriate batch sizes (4096 vs their 256) for stable ratio estimation. The effectiveness of this approach is demonstrated by our consistent improvement in reconstruction quality while maintaining downstream performance.

3 Background

The foundations of sparse coding trace back to seminal work in computational neuroscience Olshausen & Field (1996), which demonstrated that sparse, independent features naturally emerge when optimizing for efficient representations of sensory input. This principle was later formalized in deep learning through sparse autoencoders Bengio (2007), which combine dimensionality reduction with sparsity constraints to learn interpretable features.

Modern sparse autoencoders build upon several key advances:

- Efficient optimization through adaptive methods Kingma & Ba (2014)
- Stable training via normalization techniques Ba et al. (2016)
- Weight normalization for improved convergence Salimans & Kingma (2016)

These techniques enable SAEs to learn high-quality representations while maintaining sparsity, though achieving targeted feature separation remains challenging. Recent work has shown that careful initialization and progressive constraints are crucial for balancing reconstruction quality with feature independence Suteu & Guo (2019).

3.1 Problem Setting

Consider a pre-trained language model with hidden states $\mathbf{h} \in \mathbb{R}^d$. At a given layer l, we observe activation vectors $\mathbf{x} = f_l(\mathbf{h}) \in \mathbb{R}^d$. Our goal is to learn an encoder-decoder pair (E, D) that maps these activations to a sparse, interpretable space while preserving the model's computational capabilities.

Formally, we seek functions:

$$E : \mathbb{R}^d \to \mathbb{R}^k$$
$$D : \mathbb{R}^k \to \mathbb{R}^d$$

that minimize the objective:

$$\mathcal{L}(E, D) = \mathbb{E}_{\mathbf{x}}[\|\mathbf{x} - D(E(\mathbf{x}))\|_2^2 + \lambda\|E(\mathbf{x})\|_1 + \alpha\Omega(E, D)] \tag{1}$$

subject to the constraints:

$$\begin{aligned} \|E(\mathbf{x})\|_0 &\leq s \quad \text{(sparsity)} \\ \|\mathbf{w}_i^\top \mathbf{w}_j\| &\leq \epsilon_{ij} \quad \text{(feature separation)} \\ \mathcal{L}_{\text{task}}(M_{\text{SAE}}) &\leq (1 + \delta)\mathcal{L}_{\text{task}}(M) \quad \text{(performance)} \end{aligned}$$

where:

- λ, α control sparsity and orthogonality penalties
- $\mathbf{w}_i$ are decoder weight vectors
- ϵ_{ij} are feature-pair specific correlation thresholds
- δ bounds the allowable performance degradation
- M_{SAE} denotes the model with SAE intervention

This formulation extends classical sparse coding by introducing targeted feature separation through $\Omega(E, D)$ while maintaining model performance through explicit constraints. The challenge lies in optimizing these competing objectives without sacrificing reconstruction quality or downstream performance.

4 Method

Building on the problem formulation in Section 3, we introduce three key innovations to achieve targeted feature separation while maintaining reconstruction quality. Our approach extends the basic SAE framework by introducing dynamic constraints that smoothly adjust separation pressure based on feature activation patterns.

4.1 Progressive Orthogonality

The core challenge in optimizing the objective $\mathcal{L}(E, D)$ lies in balancing the reconstruction term with the feature separation constraint $\|\mathbf{w}_i^\top \mathbf{w}_j\| \leq \epsilon_{ij}$. We introduce a cubic scaling mechanism that provides finer control over feature interactions:

$$\Omega(E, D) = \sum_{i,j} |c_{ij}|^3 \cdot |\langle \mathbf{w}_i, \mathbf{w}_j \rangle| \tag{2}$$

where c_{ij} measures the correlation between features i and j in the encoded space. This cubic scaling ensures smooth transitions in separation pressure, with experimental validation showing improvement from baseline (loss 200.23) through quadratic (193.48) to cubic scaling (192.60).

4.2 Activation-Based Weighting

To satisfy the performance constraint $\mathcal{L}_{\text{task}}(M_{\text{SAE}}) \leq (1 + \delta)\mathcal{L}_{\text{task}}(M)$ while enforcing separation, we dynamically adjust ϵ_{ij} based on feature activation patterns:

$$\epsilon_{ij} = \epsilon_0 \cdot \exp(-\beta \log(1 + \bar{a}_i \bar{a}_j)) \tag{3}$$

where $\bar{a}_i$ is the mean activation of feature i and β controls separation strength. This allows frequently co-activated features to maintain necessary relationships while separating potentially interfering representations.

4.3 Training Procedure

The final optimization combines these mechanisms with the sparsity constraint $\|E(\mathbf{x})\|_0 \leq s$ through a progressive schedule:

$$\mathcal{L}_{\text{total}} = \|x - D(E(x))\|_2^2 + \lambda \|E(x)\|_1 + \gamma(t)\Omega(E, D) \tag{4}$$

where $\gamma(t)$ implements a warmup schedule aligned with learning rate progression. Using a batch size of 4096 for stable correlation estimates, we achieve an ℓ_0 sparsity of 85.21 while maintaining downstream performance (91.5

5 Experimental Setup

We evaluated our approach on the Gemma-2-2b model, focusing on layer 19 ($d = 2304$) based on preliminary analysis showing strong feature differentiation at this depth. Our implementation uses PyTorch with mixed-precision training (bfloat16) to handle the model's full context window.

Training Data We used 10 million tokens from the Pile dataset's uncopyrighted subset, maintaining the model's original tokenization and a context length of 128 tokens. Data was processed in batches of 4096 sequences using an activation buffer to efficiently capture intermediate model states.

Model Configuration The sparse autoencoder matches the hidden dimension ($d_{\text{in}} = d_{\text{sae}} = 2304$) with ReLU activation and normalized decoder weights. We initialize using Kaiming uniform initialization for both encoder and decoder weights, with small uniform noise (± 0.01) for biases.

Optimization Training uses Adam with the following hyperparameters:

- Learning rate: 3×10^{-4} with 1000-step warmup
- L1 sparsity penalty: $\lambda = 0.04$
- Orthogonality penalty: $\alpha = 0.001$
- Batch size: 4096 sequences
- Training steps: 4882 with checkpoints every 100 steps

Evaluation Protocol We assess model performance through:

- **Core Metrics**: Reconstruction MSE, activation sparsity (ℓ_0, ℓ_1), and explained variance
- **Feature Analysis**: Absorption scores and sparse probing accuracy across 8 classification tasks
- **Knowledge Separation**: SCR metrics on biographical and product review datasets
- **Model Integrity**: KL divergence and cross-entropy on held-out text

Each evaluation metric is computed on separate validation sets to ensure robust assessment of both reconstruction quality and knowledge separation. We run all experiments on a single GPU with 40GB memory, with evaluation batches sized to maximize GPU utilization while maintaining numerical stability.

6 Results

Our experiments evaluate the effectiveness of progressive feature separation through a series of controlled comparisons. The baseline model achieved a reconstruction loss of 200.23, with our final cubic scaling implementation reducing this to 192.60, a 3.8% improvement. Table 1 summarizes the key performance metrics.

Metric	Value	Baseline
Reconstruction Loss	192.60	200.23
Explained Variance	30.86%	-
ℓ_0 Sparsity	85.21	-
ℓ_1 Norm	458.0	-
KL Divergence	0.795	10.06
Cross-Entropy	0.789	12.44

Table 1: Core performance metrics comparing our method to baseline. Lower values are better for all metrics except explained variance.

Ablation Studies We conducted two key ablation experiments:

1. **Batch Size Impact**: Reducing batch size from 4096 to 2048 led to unstable ratio estimates and increased loss (214.41), demonstrating the importance of sufficient samples for reliable feature correlation estimation.

2. **Scaling Function**: We observed a clear progression in reconstruction quality:

- Initial ratio-based: 214.41 loss
- Quadratic scaling: 193.48 loss
- Cubic scaling: **192.60** loss
- Activation-based: 195.37 loss

Downstream Performance The model maintained strong performance across diverse tasks while achieving better feature separation:

- Sparse probing accuracy: 91.51% (baseline: 95.06%)
- Sentiment analysis: 93.55% (baseline: 98.15%)
- Code completion: 93.14% (baseline: 96.74%)
- Language identification: 95.72% (baseline: 99.94%)

Limitations Several important limitations emerged:

- The unlearning evaluation score of 0.0 indicates incomplete knowledge isolation
- Performance degradation of 3-4% on downstream tasks suggests room for improvement in preserving model capabilities
- The modest 3.8% improvement in reconstruction loss may indicate fundamental architectural constraints

These results demonstrate that while our method successfully improves feature separation without catastrophic performance degradation, achieving complete knowledge isolation remains challenging. The trade-off between separation strength and task performance suggests the need for more sophisticated architectures or training strategies.

7 Conclusions and Future Work

This work introduced a novel approach to feature separation in sparse autoencoders through progressive orthogonality constraints and cubic scaling. Our key contribution - the dynamic adjustment of separation pressure based on feature correlations and activation patterns - demonstrates that careful engineering of constraints can improve both reconstruction quality and feature interpretability. The success of our cubic scaling mechanism, particularly in maintaining downstream performance while achieving better feature separation, suggests a promising direction for controlled knowledge manipulation in large language models.

The progression of our experiments revealed fundamental insights about the relationship between feature separation and model performance. While we achieved meaningful improvements in reconstruction quality, the limitations encountered - particularly in complete knowledge isolation - point to deeper architectural challenges that warrant further investigation. The trade-off between separation strength and task performance appears to be a fundamental constraint rather than an implementation artifact.

Looking ahead, several promising research directions emerge:

- **Architectural Innovations:** Developing specialized structures for managing feature interactions while preserving task-critical relationships
- **Dynamic Constraints:** Exploring adaptive mechanisms that adjust separation pressure based on task requirements and feature importance
- **Scalable Solutions:** Investigating techniques to maintain separation effectiveness as model size and complexity increase

These findings lay the groundwork for more sophisticated approaches to model interpretability and controlled behavior modification. As language models continue to grow in capability and complexity, the ability to precisely manipulate their internal representations while preserving performance becomes increasingly crucial for their safe and reliable deployment.

References

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. pp. 113–149, 2015.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Yoshua Bengio. Learning deep architectures for ai. *Found. Trends Mach. Learn.*, 2:1–127, 2007.

Xi Chen, Yan Duan, Rein Houthooft, John Schulman, I. Sutskever, and P. Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. pp. 2172–2180, 2016.

Timo Flesch, Keno Juechems, T. Dumbalska, Andrew M. Saxe, and C. Summerfield. Rich and lazy learning of task representations in brains and neural networks. *bioRxiv*, 2021.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Irina Higgins, L. Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2017.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Michael Lan, Philip Torr, Austin Meek, Ashkan Khakzar, David Krueger, and Fazl Barez. Sparse autoencoders reveal universal feature spaces across large language models. *ArXiv*, abs/2410.06981, 2024.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards principled evaluations of sparse autoencoders for interpretability and control. *ArXiv*, abs/2405.08366, 2024.

B. Olshausen and D. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381:607–609, 1996.

OpenAI. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

Tim Salimans and Diederik P. Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. *ArXiv*, abs/1602.07868, 2016.

Joseph Sirosh. A self-organizing neural network model of the primary visual cortex. pp. 815–818, 1995.

Mihai Suteu and Yike Guo. Regularizing deep multi-task networks using orthogonal gradients. *ArXiv*, abs/1912.06844, 2019.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Nuoyan Zhou, Dawei Zhou, Decheng Liu, Xinbo Gao, and Nannan Wang. Mitigating feature gap for adversarial robustness by feature disentanglement. *ArXiv*, abs/2401.14707, 2024.