HIGH-PROBABILITY GUARANTEES FOR LINEAR ACCESSIBILITY IN FEATURE SUPERPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural networks can leverage feature superposition to encode more concepts than dimensions, but cross-feature interference constrains the linear accessibility of simultaneously active features. By framing linear accessibility as a compressed sensing problem, we derive high-probability recovery bounds for fixed supports under subgaussian noise, proving that $d = O_\varepsilon(k \log m)$ dimensions suffice for the per-input regime in which networks typically operate in, in contrast to the quadratic scaling required by uniform worst-case guarantees. We characterize the asymmetry between active and inactive interference and the trade-off between interference and observation-noise budgets. We then validate these bounds across system parameters through numerical experiments based on Gaussian-tail approximations. We also introduce IHT-SAE, which uses learned iterative refinement to improve feature recovery beyond the limits of linear accessibility. These results quantify the geometric constraints of the linear representation hypothesis, providing a framework for evaluating sparse autoencoders, compositional generalization, and neural interpretability. Code to reproduce this work is available at: https://anonymous.4open.science/r/Linear-Availability-07CD

1 INTRODUCTION

Stochastic separation theorems show that in high dimensions, any point in a random set can be separated from the others by a hyperplane with high probability, even when the set grows exponentially with dimension (Sidorov & Zolotykh, 2020). Though traditionally applied to single features, these concentration principles extend to simultaneously active features. Compressed sensing formalizes how a vast number of features (m) can be encoded into a smaller d-dimensional space when a sparse subset of size k is active; mirroring feature superposition in neural networks (Elhage et al., 2022).

Even though LLM spaces are globally not uniform (Ethayarajh, 2019), concept subspaces are (Cai et al., 2021), and can be effectively approximated by Gaussian distributions (Zhao et al., 2025). It is this very local uniformity that gives rise to representational geometry (linear representation hypothesis) in vision encoders (Radford et al., 2021) and LLMs (Park et al., 2025; 2024), which encode features into composite structures (e.g., categorical, hierarchical, spatial, relational, etc).

Within this geometry, co-occurring features such as "Christmas" and "December" can occupy similar directions, producing constructive interference that aids decoding (Prieto et al., 2026), while destructive interference can suppress active features (Stevinson et al., 2025). Yet true compositional reasoning may require recovering unusual combinations such as "Christmas in July" without familiar co-occurrences, making control of cross-feature interference central to reliable linear decoding.

We apply the subgaussian Hoeffding inequality to derive high-probability bounds for linear decoding, showing that the sufficient dimension scales linearly with the number of active features, compared with quadratic scaling under uniform worst-case guarantees (Garg et al., 2026). We further use Gaussian-tail approximations to estimate failure rates and assess these predictions numerically.

Sparse autoencoders (SAEs) recover such features by reconstructing representations as sparse combinations of learned feature directions (Huben et al., 2024). Standard single-step encoders, however, remain constrained by linear accessibility. We therefore introduce IHT-SAE, which refines a learned sparse estimate through residual-based updates with learned step sizes and TopK thresholding. We further find that decoder performance depends on the coefficient distribution: strongly decaying language-model activations favor greedy residual subtraction, while flatter vision-encoder activations narrow the gap between strategies.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Our main contribution is a framework for understanding feature accessibility in superposition:

- We distinguish uniform worst-case from fixed-support high-probability linear accessibility guarantees, showing that in the latter, governed by cross-feature interference, a dimension $d = O_\varepsilon(k \log m)$ suffices for fixed error and failure tolerances when the noise satisfies $\sigma = O_\varepsilon(1/\sqrt{\log m})$.
- We place linear/nonlinear encoding/decoding within a shared representational regime, consistent with the linear representation hypothesis. Here, nonlinearities are used for setting/getting the feature coefficients to minimize or bypass the cross-feature interference.
- We connect these guarantees to SAEs, introduce IHT-SAE to recover features beyond the limits of linear accessibility, while decoupling iteration count from the sparsity budget, and demonstrate that the choice of nonlinear decoder should match the model's coefficient distribution.

2 THEORY: COMPRESSED SENSING FRAMEWORK

We model the problem in a compressed sensing framework, where a high-dimensional sparse vector is mapped to a lower-dimensional measurement space.

2.1 LINEAR VERSUS NONLINEAR ENCODING

Definition 1 (Linear encoding with independent noise)**.** *Let $x \in [-1, 1]^m$ be a k-sparse input signal, let $A = [a_1, \ldots, a_m] \in \mathbb{R}^{d \times m}$ be a dictionary of unit ℓ_2-norm feature embeddings ($\|a_i\|_2 = 1$), and let $z \in \mathbb{R}^d$ denote an independent observation noise vector. The linear encoding maps the input and noise to a compressed d-dimensional sketch y via:*

$$y = Ax + z = \sum_{j \in S} x_j a_j + z$$

where $S \subseteq [m]$ represents the active feature support of size $|S| \leq k$.

An alternative to this framework is encoding nonlinear representations. Linear representations possess a wider margin of error during decoding, operating as a probabilistic structure. Geometrically, decoding a linear sketch is akin to measuring the cosine similarity between nearly orthogonal feature directions; slight measurement errors yield proportionally small perturbations in the recovered state. In contrast, discontinuous nonlinear representations, such as RAM of storing the k largest coefficients alongside their indices, can achieve a high compression of $d = O(k)$ (Ba et al., 2010). However, slight errors during decoding of such structures (e.g., a corrupted index mapping) can completely break the structure, resulting in disparate representations rather than proportional deviations. Such a non-differentiable representation space renders gradient-based optimization intractable (but not impossible). We later show how nonlinearity can be employed for encoding linear representations while minimizing the margin of error when reading said features.

2.2 NONLINEAR DECODING

Classical compressed sensing demonstrates that a k-sparse vector in $\mathbb{R}^m$ can be embedded into approximately $d = O(k \log(m/k))$ measurements, assuming a known measurement matrix (Candes & Tao, 2005). Here, recovery can be achieved in polynomial time using nonlinear decoding algorithms (ℓ_1-minimization) (Candès et al., 2006), which extends to noisy environments, provided the noise z is bounded by $||z||_{\ell_2} \leq \varepsilon$ (Candès, 2008). Assuming the measurement matrix satisfies the restricted isometry property with a constant $\delta_{2k} < \sqrt{2} - 1$, this recovery is guaranteed to be stable in both the ℓ_1/ℓ_1 and ℓ_2/ℓ_1 regimes (Candès, 2008), with a matching lower bound in these regimes (Ba et al., 2010). Here, these proofs hold uniformly for all combinatorially many k-sparse supports.

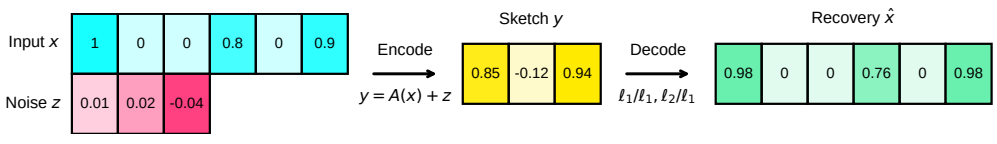


Figure 1: Compressed sensing encoding and decoding pipeline.

2.3 LINEAR DECODING

Linear decoding requires features to be linearly accessible, meaning they are recovered by a fixed linear decoder. As opposed to relatively clean recovery as in nonlinear decoding, linear decoding may produce cross-feature interference on both inactive and active coordinates (Stevinson et al., 2025). Depending on its sign, this interference can create spurious inactive estimates or distort active estimates. Decoding can therefore be framed as the existence of a threshold which separates active features from inactive ones.

Definition 2 (Linear decoding). *Given a compressed sketch $y \in \mathbb{R}^d$ and a dictionary $A \in \mathbb{R}^{d \times m}$, linear decoding is defined by the single-step matched-filter projection:*

$$\widehat{x} = A^\top y$$

such that the estimated coefficient for any feature $i \in [m]$ is recovered via the inner product $\widehat{x}_i = \langle a_i, y \rangle$.

The geometric limits of linear decoding are invariant to the magnitude of the compressed state y. By the bilinearity of the inner product, scaling the compressed state by a factor $\alpha > 0$ scales the entire decoding projection equally $\langle a_i, \alpha y \rangle = \alpha \langle a_i, y \rangle$. Because the compressed state is a linear combination of active features ($y = \sum x_j a_j$), this scaling magnifies both the target signal and the interference by the same factor. Figure 2 visualizes the recovery bounds in the empirical high-probability and worst-case scenarios, where a separating threshold could reside.

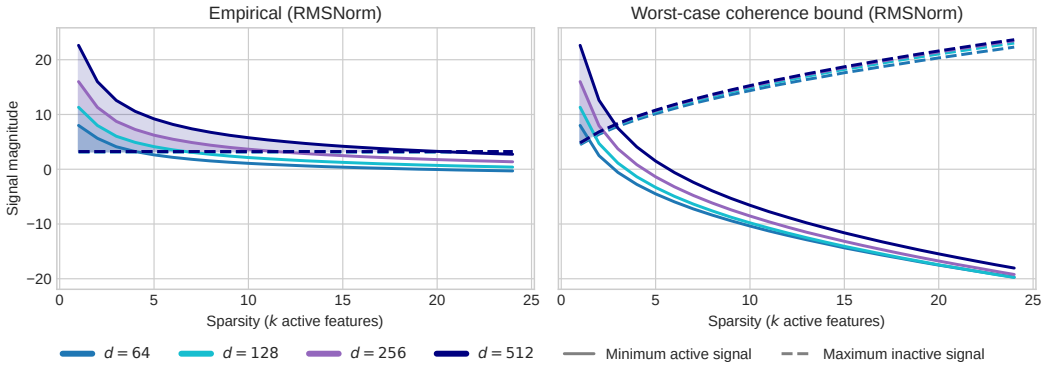


Figure 2: Signal magnitudes and recovery bounds in an RMSNorm vector space.

2.4 THE WORST-CASE SCENARIO OF LINEAR DECODING

Requiring the linear decoder to satisfy $\|A^\top A x - x\|_\infty < \varepsilon$ uniformly for every k-sparse vector $x \in [-1, 1]^m$ yields an upper bound $d = O_\varepsilon(k^2 \log m)$; a nearly matching lower bound of $d = \Omega_\varepsilon\left(\frac{k^2}{\log k} \log \frac{m}{k}\right)$ applies even when allowing any arbitrary linear decoder $B^\top$ (Garg et al., 2026). This k^2 scaling is a geometric limitation representing how effectively low-rank projections can suppress cross-feature interference uniformly over all combinatorially many k-sparse inputs.

2.5 THE EMPIRICAL HIGH-PROBABILITY LINEAR DECODING SCENARIO

By analyzing the joint probability space of the dictionary, signal, and noise, we can bound the total cross-feature interference and noise. We partition the error tolerance ε, allocating a budget of $\alpha \in (0, 1)$ to noise and $(1 - \alpha)$ to interference. A subgaussian tail bound controls both the interference I_i and the projected noise. Because subgaussian tails decay exponentially fast, applying a union bound across all features guarantees with high probability that both components remain bounded within allocated budget.

To bound the maximum interference, we partition the sum of individual tail probabilities into two:

$$\begin{aligned}
&\mathbb{P}_{A,x}\left(\max_{i\in\mathcal{I}}|I_i| \geq (1-\alpha)\varepsilon\right) \\
&\quad\leq \underbrace{2(m-|S|)\exp\left(-\frac{d(1-\alpha)^2\varepsilon^2}{2|S|}\right)}_{\text{inactive}} + \underbrace{\mathbb{I}_{|S|\geq 2}\cdot 2|S|\exp\left(-\frac{d(1-\alpha)^2\varepsilon^2}{2(|S|-1)}\right)}_{\text{active}} \qquad (1)\\
&\quad\leq 2m\exp\left(-\frac{d(1-\alpha)^2\varepsilon^2}{2|S|}\right) \leq \delta_{\text{int}}
\end{aligned}$$

The inactive term bounds the interference on inactive coordinates, where the interference is generated by all $|S|$ active features, whose erroneous inclusion would produce false positives. The active term bounds the interference on active coordinates, where the interference is generated by the remaining $|S|-1$ active features, whose erroneous exclusion would produce false negatives. The union bound includes both terms, so the condition controls the interference failures relevant to both error types.

Lemma 1 (Interference constraint)**.** *Let $d, m, k \in \mathbb{N}_{>0}$ with $1 \leq |S| \leq k \leq m$ for an arbitrarily chosen index set $S \subseteq [m]$. Suppose the dictionary columns $A = [a_1, \ldots, a_m]$ are mean-zero and sampled independently and uniformly from the unit sphere S^{d-1}, and the active signal coefficients $(x_j)_{j\in S}$ are drawn from a joint distribution $P_x \in \mathcal{P}_S$ satisfying $|x_j| \leq 1$ independently of A (with $x_j = 0$ for $j \notin S$). For any error threshold $\varepsilon > 0$, fraction $\alpha \in (0,1)$, and failure probability $\delta_{int} \in (0,1)$, if the dimension d satisfies:*

$$d \geq \frac{2|S|}{(1-\alpha)^2\varepsilon^2}\ln\left(\frac{2m}{\delta_{int}}\right)$$

then the maximum cross-feature interference is bounded by $(1-\alpha)\varepsilon$ with probability at least $1-\delta_{int}$.

For fixed α and δ_{int} independent of m, k, ε, substituting $|S| \leq k$ yields the dimension scaling $d_{\text{req}} = O_\varepsilon(k\log m)$, matching the asymptotic scaling in hyperdimensional computing theory (Thomas et al., 2021). However, unlike prior work that absorbs coefficients into feature vectors (implicitly assuming unit scale), we separate feature directions from their coefficient distributions, where our guarantees hold for arbitrary bounded coefficient distributions. This separation also yields a variance proxy for interference (Equation 14); combined with the noise variance, this proxy explains the budget borrowing dynamic. It also allows us to distinguish active from inactive interference.

Lemma 2 (Noise constraint)**.** *Let the dictionary A and signal x satisfy the conditions of Lemma 1. Suppose the observation noise $z = (z_1, \ldots, z_d)$ is drawn from a product distribution $P_z \in \mathcal{Z}_\sigma$ independent of (A, x), with independent, mean-zero coordinates satisfying the subgaussian norm bound $\|z_n\|_{\psi_2} \leq \sigma$ for $\sigma > 0$. For any fraction $\alpha \in (0,1)$, error tolerance $\varepsilon > 0$, and noise failure budget $\delta_{noise} \in (0,1)$, if the subgaussian scale satisfies:*

$$\sigma \leq \frac{\sqrt{c}\alpha\varepsilon}{\sqrt{\ln(2m/\delta_{noise})}}$$

then noise is bounded by $\alpha\varepsilon$ with probability at least $1-\delta_{noise}$, where $c > 0$ is an absolute constant.

For fixed α and δ_{noise}, this bounds the maximum certified noise scales as $\sigma_{\max} = O_\varepsilon(1/\sqrt{\log m})$.

Theorem 1 (Fixed-support decoding guarantee)**.** *Let $\varepsilon > 0$ and $\alpha \in (0,1)$, and let the failure budget be partitioned into $\delta_{int}, \delta_{noise} \in (0,1)$ such that $\delta_{int} + \delta_{noise} = \Delta < 1$. If the dimension d and noise scale σ satisfy Lemma 1 and Lemma 2, respectively, then:*

$$\forall\, S \subseteq [m] \text{ with } 1 \leq |S| \leq k, \qquad \sup_{P_x\in\mathcal{P}_S}\sup_{P_z\in\mathcal{Z}_\sigma}\mathbb{P}_{A,x,z}\left(\|\widehat{x}-x\|_\infty \geq \varepsilon\right) \leq \Delta$$

Meaning, for any arbitrary fixed support S across all valid distributions, the maximum estimation error exceeds ε with a joint failure probability of at most Δ.

Corollary 1 (High-probability support recovery)**.** *Given the ℓ_∞ error bound $\|\widehat{x}-x\|_\infty \leq \varepsilon$ from Theorem 1 and a known minimum active coefficient $x_{\min} \leq \min_{j\in S}|x_j|$, support recovery via the thresholding rule $\widehat{S} = \{i : |\widehat{x}_i| > \tau\}$ is guaranteed if there exists a threshold τ satisfying:*

$$\varepsilon_{inactive} < x_{\min} - \varepsilon_{active} \implies \underbrace{\varepsilon_{inactive}}_{\max_{i\notin S}|\widehat{x}_i|} < \tau < \underbrace{x_{\min} - \varepsilon_{active}}_{\min_{j\in S}|\widehat{x}_j|}$$

Full proof is in Appendix A.

3 NUMERICAL TESTS: EVALUATING THEORY

In our numerical tests (unless specified otherwise), all dictionary columns are drawn from Gaussian distribution and L2-normalized to unit length. For each sample, k active features are selected uniformly at random with coefficients set to 1, yielding noiseless observations.

As the Hoeffding inequality provides a conservative bound for any dictionary uniformly distributed on the unit sphere S^{d-1}, we can obtain a more precise estimate for the failure probability δ_{gauss} by modeling the interference terms independently (Equation 1) and computing their Gaussian approximation (Appendix A.3). Because this formulation bounds the tail-matching factor from below by a constant (0.5), the Gaussian approximation occasionally falls below the Hoeffding bound but never exceeds it. This approximation effectively captures the scaling dynamics of the system parameters. In configurations where the number of features is small relative to the dimension size, it yields a high probability of successful decoding; and instead when dimension size is relatively smaller, the probability decreases accordingly.

With minimum active coefficient $x_{\min} = 1$, the empirical maximum deviations are computed as $\max_{i \in S} |1.0 - \widehat{x}_i|$ and $\max_{i \notin S} |\widehat{x}_i|$ for active and inactive terms, respectively, from which we derive the empirical failure rate δ_{emp}. As δ approaches zero, the term $\ln(2m/\delta)$ grows indefinitely. To prevent finite-sample Monte Carlo noise from distorting the failure-rate ratio $\delta_{\text{gauss}}/\delta_{\text{emp}}$, we restrict our evaluation to regimes with at least 100 empirical failures. Because the empirical failure count N_{fail} for rare events approximates a Poisson distribution, its relative statistical error (coefficient of variation) is defined by the ratio of the standard deviation to the mean, $\sigma/\mu \approx \sqrt{N_{\text{fail}}}/N_{\text{fail}} = 1/\sqrt{N_{\text{fail}}}$ (Rubinstein & Kroese, 2016). By requiring $N_{\text{fail}} \geq 100$, the estimated coefficient of variation of the empirical denominator is at most $1/\sqrt{100} = 10\%$, reducing the influence of sampling noise on the failure-rate ratio. Meaning, $40,000$ trials enables evaluating high-probability decoding regimes down to a minimum failure rate of 0.25% (a 99.75% success rate).

As recovery success approaches 100%, the δ_{emp} converges to δ_{gauss} (Figure 3), demonstrating that the Gaussian approximation closely estimates the required dimension for high-probability decoding, providing empirical support that Hoeffding inequality governs the guarantee for high-probability (low failure-probability) linear accessibility. Similar results for Gaussian and Laplace[1] distributions are shown in Appendix D.

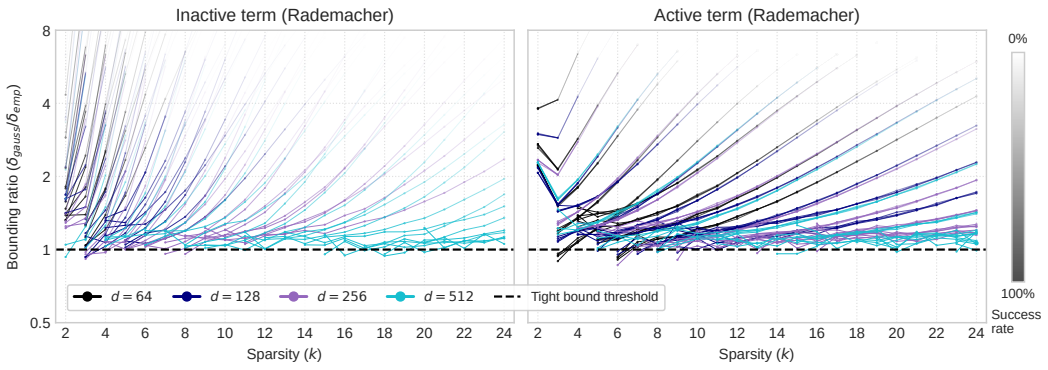


Figure 3: Tightness of Gaussian approximation for high-probability decoding on Rademacher distribution. $40,000$ trials were run for every combination of sparsity $k \in [2, 24]$, dictionary size $m \in \{512, 1024, 2048, 4096\}$ and dimension $d \in \{64, 128, 256, 512\}$, evaluated against various error thresholds $\varepsilon \in \{0.1, 0.2, \ldots, 0.9\}$. When $k = 2$ and active coefficients have equal magnitude, a projection $\langle a_i, a_j \rangle$ equally contributes to the failures of target i and j, so the union bound counts the same underlying event twice (negligible for larger k).

[1]Though Laplace distribution is subexponential, because we project the dictionary columns onto a unit sphere, it becomes bounded and subsequently subgaussian. These normalized columns are not uniform on the sphere, however. Nevertheless, under suitable central limit theorem conditions, centered, variance-normalized interference asymptotically approaches a normal distribution, motivating the Gaussian approximation.

3.1 SUPPORT RECOVERY

Figures 4 and 5 evaluate success by the existence of a valid separating threshold (Corollary 1).

In general vector space, where active features are simply summed together, the magnitude of the superposed state grows unbounded as the number of simultaneous features (k) increases. Subsequently, applying RMSNorm (Zhang & Sennrich, 2019) constrains the superposed state vector to a fixed magnitude of $\sqrt{d}$. Despite this difference in magnitudes, relative signal-to-interference ratios are identical as normalization scales the target signal and the background geometric interference by the same factor.

Our proofs partition the error tolerance ε into noise budget ($\alpha\varepsilon$) and interference budget ($(1-\alpha)\varepsilon$).

In practice, success simply requires that the sum of interference and noise remains below the tolerance ε. Because we model interference and additive noise as independent variables, Bienaymé's identity dictates that the variance of their sum equals the sum of their individual variances (we derive the variance proxy for interference in Equation 14). Meaning, the system exhibits budget borrowing, where high interference can consume the total tolerance ε, leaving no room for noise, and vice versa.

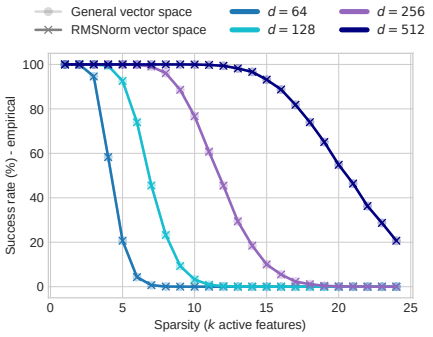


Figure 4: Recovery of general and RMSNorm vector spaces across dimensions and sparsities.

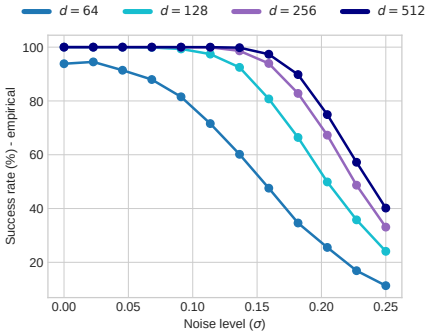


Figure 5: Recovery in noisy environment.

3.2 NONLINEAR DECODING ALGORITHMS

When dictionary is known, sparse inference can be performed using methods such as OMP (orthogonal matching pursuit), iterative hard thresholding, or LASSO (least absolute shrinkage and selection operator). OMP greedily orthogonalizes the residual of linearly most accessible feature (Pati et al., 1993). Iterative hard thresholding instead uses residual correlations to correct all feature activation estimates, then retains the k largest in magnitude (Blumensath & Davies, 2009). And LASSO estimates sparse coefficients by minimizing reconstruction error with an ℓ_1 penalty (Tibshirani, 1996). Meaning, it operates in a nonlinear decoding regime where the theoretical bounds of ℓ_1/ℓ_1 apply. This convex problem can be solved using ISTA (iterative shrinkage-thresholding algorithm) (Daubechies et al., 2004), which alternates residual-based coefficient updates with soft thresholding, or its accelerated variant, FISTA (Beck & Teboulle, 2009).

In unsupervised dictionary learning, the optimization must simultaneously discover feature directions and identify the unknown k-sparse activation support, where coefficients can be estimated using different sparse-inference methods, such as OMP, FISTA, or ISTA (Appendix B). Building on iterative soft thresholding, LISTA (learned iterative shrinkage-thresholding algorithm) unfolds ISTA into a fixed-depth network and learns its parameters through backpropagation (Gregor & LeCun, 2010).

3.3 FEATURE RECOVERY USING SAES

A TopK SAE (Gao et al., 2024) encoder resembles a learned hard-thresholding step, whereas a ReLU SAE encoder resembles LISTA without iterative refinement. Both amortize the cost of sparse inference via a learned encoder that predicts sparse codes in a single forward pass, jointly optimizing the encoder and decoder dictionary through gradient descent. As monotonic activation functions (e.g., ReLU) do not increase the capacity of linear accessibility (Garg et al., 2026), these standard SAEs operate in a linear decoding regime. This could explain why SAEs fail at compositional generalization within a compressed sensing framework when trained on LLM activations (Barin-Pacela et al., 2026). Because Transformers can provably implement LISTA-type iterative nonlinear decoding algorithms (Liu et al., 2025), they have the capacity to operate in a nonlinear decoding regime. Meaning, LLMs can represent features in denser superposition that remain inaccessible to the linear decoding regime of SAEs.

To elevate SAEs into a nonlinear decoding regime, MP-SAE incorporates matching pursuit into an SAE by learning a dictionary through an unrolled matching-pursuit encoder that performs sequential residual updates (Costa et al., 2025). To further improve robustness to unknown sparsity, building on iterative hard thresholding, we augment a TopK SAE with learned iterative refinement (as in LISTA), and tie the update matrices to the decoder dictionary in order to align refinement with the SAE's reconstruction objective.

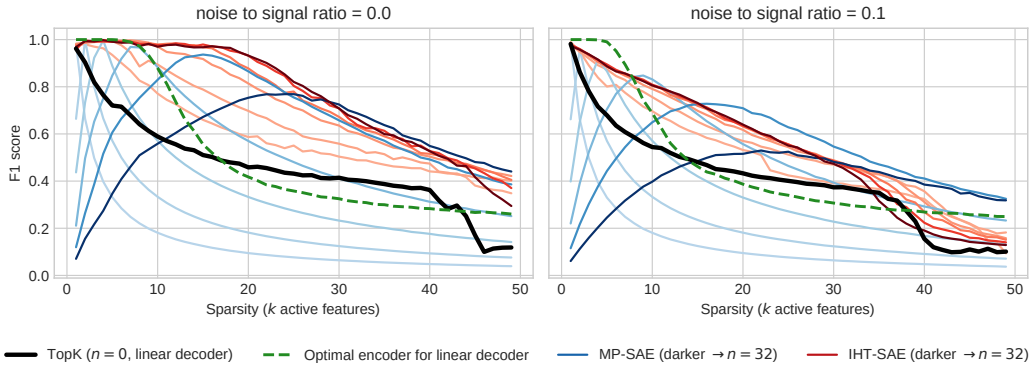


Figure 6: Support recovery using TopK, IHT-SAE and MP-SAE in RMSNorm vector space ($m = 1024$, $d = 128$). Curves show F1 scores versus active sparsity k at noise-to-signal power ratios of 0 and 0.1. To evaluate support recovery up to permutation and sign ambiguities, predicted feature activations are greedily aligned with the ground truth by mapping each true dictionary direction to the learned latent direction that maximizes their absolute cosine similarity. Reliable support recovery requires active coefficients to be sufficiently large to remain distinguishable from cross-feature interference and noise. We therefore impose a minimum active coefficient of 0.1 before normalization. Darker shades indicate more iterations ($n \in \{1, 2, 4, 8, 16, 32\}$); black denotes single-step TopK ($n = 0$). The dashed green line denotes the optimal encoder baseline. Similar results for general vector space in C.2. Detailed methodology in Section C.

Our IHT-SAE (algorithm in Appendix C.3) refines a learned sparse initialization through residual-based updates with learned step sizes. Each update adjusts coefficients jointly, then retains at most top-k values, followed by ReLU. MP-SAE instead starts from zero, greedily selecting a feature from the residual and subtracting its contribution. Thus, $n = k$ is an ideal iteration budget for MP-SAE if each iteration selects one new, correct feature. But choosing that n comes with the caveat of needing to know the active sparsity. IHT-SAE decouples iteration count from the sparsity budget by retaining at most k features after each update.

We also compute an optimal coefficient baseline that maximizes the separation margin between active and inactive features using gradient descent (methodology in Appendix C.1), outperforming a regular linear decoder. For context, every iteration in isolation within these SAEs is governed by linear accessibility, so future work could incorporate better encoding to preemptively account for interference and boost recovery performance even further. This also extends to foundational models, where one could maximize linear accessibility while minimizing false positives and negatives.

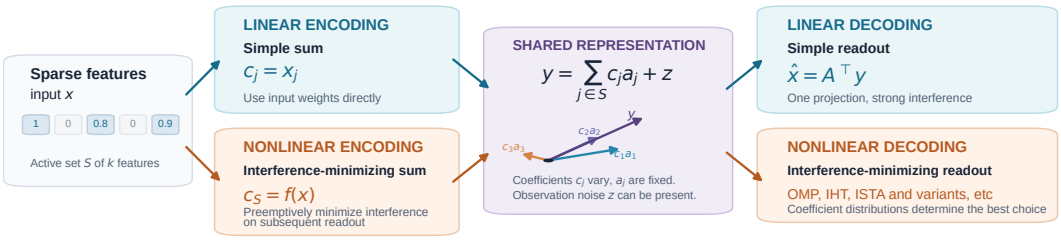


Figure 7: Overview of the compressed sensing framework of linear/nonlinear encoding/decoding.

3.3.1 EVALUATION ON FOUNDATION MODELS

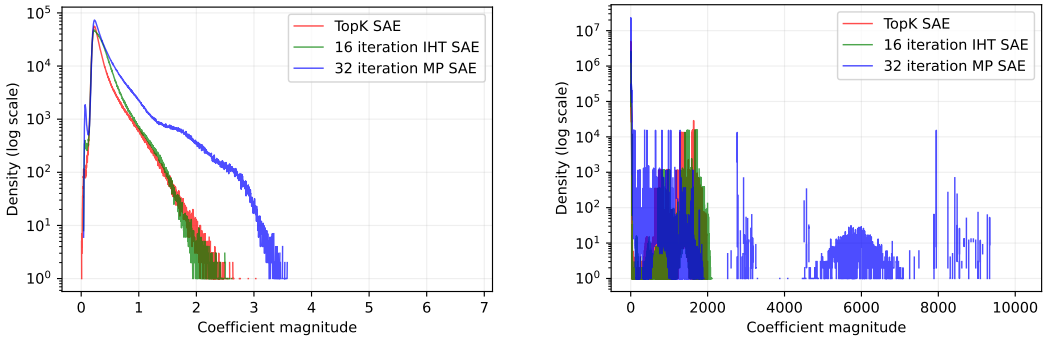
We evaluated both IHT-SAE and MP-SAE on activations extracted from foundational models. We probed layer index 18 of Qwen3-4B (Bai et al., 2023) using the TinyStories dataset (Eldan & Li, 2023) and layer index 5 of the CLIP ViT-Base (Radford et al., 2021) using the Oxford-IIIT Pet dataset (Parkhi et al., 2012). We report the reconstruction variance explained (R^2, Bricken et al. (2023)) and downstream performance preservation, measured by the degradation in cross-entropy (CE) loss when replacing original model activations with the SAE reconstructions. The training methodology mirrors the synthetic setup (Section C), where models were trained for $10,000$ steps with a batch size of 1024 using the Adam optimizer, incorporating ℓ_2-normalization of the decoder dictionary after every update. For these experiments, we fixed the sparsity budget to $k = 32$ and set the SAE dictionary size to four times the hidden dimension of the respective foundation model. We evaluated single-step TopK ($n = 0$), IHT-SAE ($n \in \{4, 8, 16\}$), and MP-SAE ($n = 32$). Here, 16 iteration IHT-SAE and 32 iteration MP-SAE take equal amount of FLOPs per forward pass, where each IHT iteration takes roughly the amount of FLOPs of baseline TopK forward pass.

Table 1: Downstream performance and reconstruction variance explained of SAE architectures on foundation models. The baseline row shows the original cross-entropy (CE) loss. SAE rows report mean $\pm$ sample standard deviation across three seeds for R^2 reconstruction and the increase in CE loss (Δ) when substituting original activations with SAE reconstructions.

	CLIP ViT		Qwen3 4B GPT	
Encoder	R^2	Δ CE loss	R^2	Δ CE loss
Baseline	–	1.6961	–	1.9833
TopK (0 iterations)	0.7000 ± 0.0003	$+0.0558 \pm 0.0036$	0.9989 ± 0.0000	$+0.0996 \pm 0.0022$
IHT (4 iterations)	0.7509 ± 0.0008	$+0.0286 \pm 0.0063$	0.9989 ± 0.0000	$+0.0976 \pm 0.0026$
IHT (8 iterations)	0.7576 ± 0.0010	$+0.0226 \pm 0.0010$	0.9989 ± 0.0000	$+0.1252 \pm 0.0075$
IHT (16 iterations)	0.7584 ± 0.0024	$+0.0255 \pm 0.0022$	0.9988 ± 0.0000	$+0.1318 \pm 0.0102$
MP (32 iterations)	$\mathbf{0.7623 \pm 0.0001}$	$\mathbf{+0.0201 \pm 0.0023}$	$\mathbf{0.9990 \pm 0.0000}$	$\mathbf{+0.0414 \pm 0.0008}$

Different nonlinear decoding algorithms handle coefficient distributions differently, meaning, the best algorithm depends on the coefficient distribution of the model being analyzed. OMP has been shown to perform well in α-strongly-decaying coefficient distributions (Wen et al., 2020), where the non-zero active coefficients, when sorted in descending order of magnitude such that $|x_{(1)}| \geq |x_{(2)}| \geq \cdots \geq |x_{(k)}| > 0$, satisfy the condition: $|x_{(i+1)}| \leq \alpha |x_{(i)}|$ for all $i \in \{1, \ldots, k-1\}$.

This allows it to greedily subtract massive outliers without them distorting the recovery of smaller features, which explains the better downstream reconstruction loss of MP-based SAEs over IHT in LLMs (Figure 8). As such, we provide the theoretical and empirical pieces for constructing SAEs whose inductive bias matches that of the model being analyzed.



(a) CLIP vision model

(b) GPT language model

Figure 8: Unfilled histogram of coefficient distributions on a logarithmic y-axis. MP SAE has an inductive bias towards an α-strongly-decaying coefficient distribution, highlighting its capacity to isolate extreme outlier activations in language models.

4 DISCUSSION

It is important to distinguish a concept manifold from features. While a conceptual manifold, all things that could be (e.g., days of the week) might not be one-dimensionally linear, its materialized instances (e.g., Monday) exists as a linearly separable point along that structure (Engels et al., 2025). While it has been shown to extend to higher dimensional manifold structures (numbers on a helix, Kantamneni & Tegmark (2025)), every instance (e.g, Friday the second) is still a composite feature of sparse activations.

4.1 UNKNOWN NONLINEAR TRANSFORM

A parallel line of work, known as World Models, has emerged which force nonlinear observations into a linear transform of the world's latent structure (Klindt et al., 2026). This is best understood through structural features, which are compositional properties that require nonlinearity to be recovered (Vompa & Tammet, 2026). Rather than assuming representations are inherently nonlinear, we hypothesize it is more akin to an unsolved problem where the base features are represented linearly, while higher-order properties are derived via iterative bottom-up perception (e.g., nonlinear encoding) or top-down reasoning (e.g., nonlinear decoding) (Vompa et al., 2026). Here, new information is added linearly to the residual stream, while successive layers can transform the representational basis and refine feature estimates. This enables iterative nonlinear encoding/decoding, providing a mechanism through which greater Transformer depth can improve compositional generalization (Vompa & Tammet, 2026).

4.2 LIMITATIONS

The connection between linear accessibility and LLM mechanics remains speculative, relying on external empirical evidence. Under the conditions of the Lindeberg central limit theorem (Lindeberg, 1922), centered, variance-normalized interference asymptotically converges to a normal distribution as $|S|$ grows; coupled with normalization (typically RMSNorm), their linear projections become bounded and subsequently subgaussian. This does not guarantee the uniform feature directions assumed by our theorem; nevertheless, we show that Gaussian tails are bounded by the Hoeffding bound theoretically and through numerical experiments. We describe the conditions where linear accessibility is possible, and subsequently the sufficient conditions for SAEs and more generally LRH, the extent LLMs utilize this remains unknown. This work does not claim a new state-of-the-art architecture, but rather clarifies the conditions governing a decoder choice.

5 CONCLUSION

We characterize the high-probability guarantees of linear accessibility in feature superposition, showing that a sufficient dimension for fixed-support decoding is $d = O(k \log m)$ dimensions. Within this framework, we demonstrate that linear/nonlinear encoding/decoding can share a linear representational regime (assumed by LRH) of a simple weighted sum over active features, where nonlinearities can be used for setting/getting the coefficients to bypass cross-feature interference. These results clarify the geometric limits of the linear representation hypothesis and suggest how learned superposition can support compositional reasoning.

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing proofs and implementing methods in code. We did not use generative AI tools to help develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, or interpret results. Furthermore, tasks such as generating synthetic datasets, cleaning and reformatting data, and supporting qualitative and thematic data analysis were not applicable to this work. We have reviewed all AI-assisted work. For example, all proofs were thoroughly checked manually and generated code was verified. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Khanh Do Ba, Piotr Indyk, Eric Price, and David P. Woodruff. Lower bounds for sparse recovery. In *SODA*, 2010. URL https://arxiv.org/abs/1106.0365.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report, 2023. URL https://arxiv.org/abs/2309.16609.

Vitória Barin-Pacela, Shruti Joshi, Isabela Camacho, Simon Lacoste-Julien, and David Klindt. Stop probing, start coding: Why linear probes and sparse autoencoders fail at compositional generalisation. In *2nd Workshop on Compositional Learning: Safety, Interpretability, and Agents*, 2026. URL https://openreview.net/forum?id=3NN3ehMKxW.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1), 2009. URL https://doi.org/10.1137/080716542.

Thomas Blumensath and Mike E. Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 2009. ISSN 1063-5203. URL https://doi.org/10.1016/j.acha.2009.04.002.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Xingyu Cai, Jiaji Huang, Yuchen Bian, and Kenneth Church. Isotropy in the contextual embedding space: Clusters and manifolds. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=xYGNO86OWDH.

Emmanuel Candes and Terence Tao. Decoding by linear programming. In *IEEE Transactions on Information Theory*, 2005. URL https://arxiv.org/abs/math/0502327.

Emmanuel J. Candès. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008. ISSN 1631-073X. URL https://www.sciencedirect.com/science/article/pii/S1631073X08000964.

Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics*, 59(8):1207–1223, 2006. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20124.

Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba E. Ba. From flat to hierarchical: Extracting sparse representations with matching pursuit. In *NeurIPS*, 2025. URL https://openreview.net/forum?id=Ll5miDx8KB.

I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 2004. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/cpa.20042.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL https://arxiv.org/abs/2305.07759.

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, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

K. Engan, S.O. Aase, and J. Hakon Husoy. Method of optimal directions for frame design. In *1999 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings. ICASSP99 (Cat. No.99CH36258)*, 1999. URL https://ieeexplore.ieee.org/document/760624.

Joshua Engels, Eric J Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. *ICLR*, 2025.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *EMNLP-IJCNLP*, 2019. URL https://aclanthology.org/D19-1006/.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Nikhil Garg, Jon Kleinberg, and Kenny Peng. How many features can a language model store under the linear representation hypothesis? In *COLT*, 2026. URL https://arxiv.org/abs/2602.11246.

Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, 2010. URL https://dl.acm.org/doi/10.5555/3104322.3104374.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *ICLR*, 2024. URL https://openreview.net/forum?id=F76bwRSLeK.

Subhash Kantamneni and Max Tegmark. Language models use trigonometry to do addition, 2025. URL https://arxiv.org/abs/2502.00873.

David Klindt, Yann LeCun, and Randall Balestriero. When does lejepa learn a world model?, 2026. URL https://arxiv.org/abs/2605.26379.

J. W. Lindeberg. Eine neue herleitung des exponentialgesetzes in der wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift*, 15:211–225, 1922. URL http://eudml.org/doc/167717.

Renpu Liu, Ruida Zhou, Cong Shen, and Jing Yang. On the learn-to-optimize capabilities of transformers in in-context sparse recovery. In *ICLR*, 2025. URL https://openreview.net/forum?id=NHhjczmJjo.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *ICML*, 2024. URL https://arxiv.org/abs/2311.03658.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *ICLR*, 2025. URL https://openreview.net/forum?id=bVTM2QKYuA.

Omkar M. Parkhi, Andrea Vedaldi, Andrew Zisserman, and C. V. Jawahar. Cats and dogs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, 1993. URL https://ieeexplore.ieee.org/document/342465.

Lucas Prieto, Edward Stevinson, Melih Barsbey, Tolga Birdal, and Pedro A. M. Mediano. From data statistics to feature geometry: How correlations shape superposition. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=7akSRQS5Xh.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *ICML*, 2021.

Reuven Y. Rubinstein and Dirk P. Kroese. *Simulation and the Monte Carlo Method*. 2016. URL https://app.dimensions.ai/details/publication/pub.1106805670.

Sergey Sidorov and Nikolai Zolotykh. Linear and fisher separability of random points in the d-dimensional spherical layer and inside the d-dimensional cube. *IJCNN*, 2020.

Edward Stevinson, Lucas Prieto, Melih Barsbey, and Tolga Birdal. Adversarial attacks leverage interference between features in superposition. In *Mechanistic Interpretability Workshop at NeurIPS 2025*, 2025. URL https://openreview.net/forum?id=LqI52GG2Ss.

Anthony Thomas, Sanjoy Dasgupta, and Tajana Rosing. A theoretical perspective on hyperdimensional computing. *Journal of Artificial Intelligence Research*, 72, October 2021. URL http://dx.doi.org/10.1613/jair.1.12664.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 01 1996. ISSN 0035-9246. URL https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, 2 edition, 2026.

C. Vignat and S. Bhatnagar. An extension of wick's theorem. *Statistics and Probability Letters*, 2008. URL https://www.sciencedirect.com/science/article/pii/S0167715208001405.

Enrico Vompa and Tanel Tammet. The scaling properties of implicit deductive reasoning in transformers, 2026. URL https://arxiv.org/abs/2605.04330.

Enrico Vompa, Tanel Tammet, and Mohit Vaishnav. Beyond the linear separability ceiling: Aligning representations in VLMs. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=3uX4p80bN0.

Jinming Wen, Rui Zhang, and Wei Yu. Signal-dependent performance analysis of orthogonal matching pursuit for exact sparse recovery. *IEEE Transactions on Signal Processing*, 2020. URL http://dx.doi.org/10.1109/TSP.2020.3016571.

Biao Zhang and Rico Sennrich. *Root mean square layer normalization*. 2019. URL https://arxiv.org/abs/1910.07467.

Haiyan Zhao, Heng Zhao, Bo Shen, Ali Payani, Fan Yang, and Mengnan Du. Beyond single concept vector: Modeling concept subspace in llms with gaussian distribution. *ICLR*, 2025. URL https://arxiv.org/abs/2410.00153.

A SUFFICIENT HIGH-PROBABILITY BOUND FOR LINEAR DECODING

Grounded in the linear representation hypothesis, the dictionary $A = [a_1, \ldots, a_m] \in \mathbb{R}^{d \times m}$ encodes semantic features as normalized latent directions ($\|a_i\|_2 = 1$). The observed state is a sparse linear superposition of these vectors $y = Ax + z$, where z is additive observation noise. We analyze the linear decoder:

$$\widehat{x} = A^\top y \tag{2}$$

Let $d, m, k \in \mathbb{N}_{>0}$ with $1 \le k \le m$, where d is the ambient dimension (number of observations), m the total feature count, and k the maximum sparsity. Fix an arbitrary deterministic index set $S \subseteq [m]$ with sparsity $1 \le |S| \le k$. We make the following assumptions throughout:

- Dictionary columns $a_1, \ldots, a_m$ are mean-zero, and sampled independently and uniformly from the unit sphere S^{d-1}.
- The active signal coefficients $(x_j)_{j \in S}$ are drawn from any joint distribution $P_x \in \mathcal{P}_S$, where $\mathcal{P}_S$ denotes the broad class of all valid bounded coefficient distributions ($|x_j| \le 1$) that are independent of A. We set $x_j = 0$ for $j \notin S$.
- Observation noise $z = (z_1, \ldots, z_d)$ is drawn from a distribution $P_z \in \mathcal{Z}_\sigma$, where $\mathcal{Z}_\sigma$ denotes the class of product noise distributions independent of (A, x) with independent, mean-zero coordinates satisfying the subgaussian norm bound $\|z_n\|_{\psi_2} \le \sigma$, where $\sigma > 0$.

Our analysis of this decoding draws largely upon the high-dimensional probability techniques detailed in Vershynin (2026), from which the theorems utilized below are sourced. We fixed a support S to avoid the uniform-support combinatorial penalty, with inactive coordinates set to zero ($x_j = 0$ for $j \notin S$). For any target coordinate $i \in [m]$, the number of interfering active features is denoted by $s_i := |S \setminus \{i\}|$. Because S is fixed, s_i is a fixed quantity bounded by $s_i = |S| - 1$ when $i \in S$, and $s_i = |S|$ when $i \notin S$. In either case, $s_i \le |S| \le k$.

A.1 ERROR DECOMPOSITION

Let $\varepsilon > 0$ denote the maximum acceptable global decoding error. To analyze this error for a target coordinate i, we expand the observation model $y = \sum_{j \in S} x_j a_j + z$ and subtract the true signal x_i from the decoder estimate $\widehat{x}_i = \langle a_i, y \rangle$:

$$\begin{aligned}
\widehat{x}_i - x_i &= \langle a_i, y \rangle - x_i \\
&= \left\langle a_i, \sum_{j \in S} x_j a_j + z \right\rangle - x_i \\
&= x_i \langle a_i, a_i \rangle + \sum_{j \in S \setminus \{i\}} x_j \langle a_i, a_j \rangle + \langle a_i, z \rangle - x_i \\
&= x_i (\langle a_i, a_i \rangle - 1) + \underbrace{\sum_{j \in S \setminus \{i\}} x_j \langle a_i, a_j \rangle}_{I_i} + \underbrace{\langle a_i, z \rangle}_{N_i}
\end{aligned} \tag{3}$$

Because the dictionary vectors are normalized ($\|a_i\|_2 = 1$), the self-term vanishes ($x_i(1-1) = 0$), isolating the decoding error to the cross-feature interference I_i and the projected noise N_i:

$$\widehat{x}_i - x_i = I_i + N_i \tag{4}$$

We define $\mathcal{I} := \{i : s_i \ge 1\}$ as the subset of coordinates that face cross-talk (worst-case having a cardinality of m); for $i \notin \mathcal{I}$, the interference is zero ($s_i = 0$ and $I_i = 0$). To separate the two contributions to the total error, we partition the global tolerance ε. We allocate a fraction $\alpha \in (0, 1)$ to the noise, and $(1 - \alpha)$ to the interference. By the triangle inequality, if the total error exceeds ε, at least one component must necessarily exceed its allocated budget. This bounds the total failure event within the union of individual failures:

$$\{|I_i + N_i| \ge \varepsilon\} \subseteq \{|I_i| \ge (1 - \alpha)\varepsilon\} \cup \{|N_i| \ge \alpha\varepsilon\} \tag{5}$$

Applying the union bound to these sets yields the decoupled probability bound:

$$\mathbb{P}(|I_i + N_i| \ge \varepsilon) \le \mathbb{P}(|I_i| \ge (1 - \alpha)\varepsilon) + \mathbb{P}(|N_i| \ge \alpha\varepsilon) \tag{6}$$

A.2 INTERFERENCE CONCENTRATION

We evaluate the interference concentration in the joint probability space of the dictionary and the signal. For any fixed target coordinate $i \in \mathcal{I}$ and any distribution $P_x \in \mathcal{P}_S$, with S fixed, we condition on the signal x and the target vector a_i realizations. Recall $I_i = \sum_{j \in S \setminus \{i\}} x_j U_j$, where $U_j = \langle a_i, a_j \rangle$. Since $a_j \sim \text{Unif}(S^{d-1})$ and a_i is a fixed unit vector, evaluating all moments allows us to bound the exponential moment.

To do so, we count the ways to partition $2r$ factors into disjoint pairs. Through the sphere version of Wick's theorem (Vignat & Bhatnagar, 2008), this count gives the even moments of U_j, which we then bound and substitute into the exponential series to bound its exponential moment. Let $I_{2r} = \{1, \ldots, 2r\}$ and let Π_{2r} be the set of partitions of I_{2r} into r disjoint pairs. For each integer $r \geq 1$, $U_j^{2r} = \prod_{k=1}^{2r} \langle a_i, a_j \rangle = \langle a_i, a_j \rangle \cdot \langle a_i, a_j \rangle \cdots \langle a_i, a_j \rangle$, with $2r$ identical factors. Because a_i is fixed and the expectation is over a_j, Wick's theorem pairs the a_i from one factor with the a_i from another, resolving every pair to $\langle a_i, a_i \rangle$. For a pairing $\sigma \in \Pi_{2r}$, let I_{2r}/σ contain one representative from each pair, giving r indices in total. Then:

$$\mathbb{E}_A[U_j^{2r} \mid a_i] = \frac{\Gamma(d/2)}{2^r \Gamma(r + d/2)} \underbrace{\sum_{\sigma \in \Pi_{2r}}}_{|\Pi_{2r}| \text{ terms}} \underbrace{\prod_{k \in I_{2r}/\sigma} \langle a_i, a_i \rangle}_{r \text{ factors, all } 1} = \frac{1 \cdot 3 \cdots (2r-1)}{d(d+2)\cdots(d+2r-2)} \leq \frac{(2r)!}{2^r r!\, d^r} \tag{7}$$

Here Γ is the Gamma function, satisfying $\Gamma(z+1) = z\Gamma(z)$, and $|\Pi_{2r}|$ denotes the number of pairings. Each pairing contributes 1 because $\langle a_i, a_i \rangle = \|a_i\|_2^2 = 1$. The final equality uses $|\Pi_{2r}| = 1 \cdot 3 \cdots (2r-1)$ and the Gamma recurrence to obtain the denominator $d(d+2)\cdots(d+2r-2)$. The inequality follows by bounding this denominator below by d^r and using $1 \cdot 3 \cdots (2r-1) = (2r)!/(2^r r!)$.

To control the probability of large interference via the exponential moment method, we first bound $\mathbb{E}_A[e^{\lambda U_j} \mid a_i]$. Since $|U_j| \leq 1$, we can expand the exponential and average its series term by term. The odd moments vanish by symmetry, and the even-moment bound above yields for every $\lambda \in \mathbb{R}$ (we will choose the exact value later):

$$\begin{aligned}
\mathbb{E}_A[e^{\lambda U_j} \mid a_i] &= 1 + \frac{\lambda^2}{2!}\mathbb{E}_A[U_j^2 \mid a_i] + \frac{\lambda^4}{4!}\mathbb{E}_A[U_j^4 \mid a_i] + \frac{\lambda^6}{6!}\mathbb{E}_A[U_j^6 \mid a_i] + \cdots && (8)\\
&= 1 + \sum_{r=1}^{\infty} \frac{\lambda^{2r}}{(2r)!}\mathbb{E}_A[U_j^{2r} \mid a_i] && (9)\\
&\leq 1 + \sum_{r=1}^{\infty} \frac{\lambda^{2r}}{(2r)!}\frac{(2r)!}{2^r r!\, d^r} && (10)\\
&= 1 + \sum_{r=1}^{\infty} \frac{1}{r!}\left(\frac{\lambda^2}{2d}\right)^r = \exp\left(\frac{\lambda^2}{2d}\right) && (11)
\end{aligned}$$

Conditional on x, a_i, the interference I_i is symmetric about zero, so its two tails have equal probability. For $t > 0$ and $\lambda > 0$, we apply Markov's inequality (Proposition 1.6.2) to $e^{\lambda I_i}$ and multiply by two. Conditional independence factors the expectation into a product, and the exponential-moment bound with parameter λx_j bounds each factor. Thus:

$$\begin{aligned}
\mathbb{P}_A(|I_i| \geq t \mid x, a_i) &= 2\mathbb{P}_A(I_i \geq t \mid x, a_i) = 2\mathbb{P}_A(e^{\lambda I_i} \geq e^{\lambda t} \mid x, a_i)\\
&\leq 2\frac{\mathbb{E}_A[e^{\lambda I_i} \mid x, a_i]}{e^{\lambda t}} = 2e^{-\lambda t} \prod_{j \in S \setminus \{i\}} \mathbb{E}_A[e^{\lambda x_j U_j} \mid x, a_i] \qquad (12)
\end{aligned}$$

Substituting Equation 11, with parameter λx_j, into Equation 12 yields:

$$\begin{aligned}\mathbb{P}_A(|I_i| \geq t \mid x, a_i) &\leq 2e^{-\lambda t} \prod_{j \in S \setminus \{i\}} \exp\left(\frac{(\lambda x_j)^2}{2d}\right) \\ &= 2e^{-\lambda t} \exp\left(\sum_{j \in S \setminus \{i\}} \frac{\lambda^2 x_j^2}{2d}\right) \\ &= 2\exp\left(-\lambda t + \frac{\lambda^2}{2d} \sum_{j \in S \setminus \{i\}} x_j^2\right).\end{aligned} \tag{13}$$

Using $|x_j| \leq 1$ and $\sum_{j \in S \setminus \{i\}} x_j^2 \leq s_i$, we obtain the variance proxy:

$$\frac{1}{d} \sum_{j \in S \setminus \{i\}} x_j^2 \leq \frac{s_i}{d} \tag{14}$$

For $s_i > 0$, choosing $\lambda = dt/s_i$ minimizes the quadratic exponent $-\lambda t + \lambda^2 s_i/(2d)$.

Thus, substituting Equation 14 into Equation 13 and choosing $\lambda = dt/s_i$, we obtain:

$$\mathbb{P}_A\left(|I_i| \geq t \mid x, a_i\right) \leq 2\exp\left(-\frac{dt^2}{s_i} + \frac{dt^2}{2s_i}\right) = 2\exp\left(-\frac{dt^2}{2s_i}\right) \tag{15}$$

We did this instead of applying the subgaussian Hoeffding inequality (Theorem 2.7.3) to avoid an unspecified absolute constant by proving it is 0.5. Furthermore, Theorem 2.2.1 assumes Rademacher variables, so we couldn't use it directly, but the steps in this proof follow the same exponential moment method.

Because the right hand side of this bound is deterministic and independent of the specific realizations of x and a_i, taking the expectation over $x \sim P_x$ and $a_i \sim \text{Unif}(S^{d-1})$ preserves the bound in the joint (A, x) probability space:

$$\mathbb{P}_{A,x}\left(|I_i| \geq t\right) \leq 2\exp\left(-\frac{dt^2}{2s_i}\right) \tag{16}$$

Intuitively, if the failure probability satisfies this bound for every individual realization of the signal and target vector, the expected failure probability across all random signals and vectors preserves this exact same bound. As such, to bound the worst-case cross-talk across all valid coordinates $i \in \mathcal{I}$, we seek the probability that the maximum interference exceeds our allocated threshold $t = (1-\alpha)\varepsilon$. Applying the union bound over $\mathcal{I}$ transforms the maximum into a sum of individual tail probabilities:

$$\mathbb{P}_{A,x}\left(\max_{i \in \mathcal{I}} |I_i| \geq (1-\alpha)\varepsilon\right) \leq \sum_{i \in \mathcal{I}} \mathbb{P}_{A,x}\left(|I_i| \geq (1-\alpha)\varepsilon\right) \tag{17}$$

Where $\max_{i \in \varnothing} |I_i| = 0$. We evaluate this sum by partitioning $\mathcal{I}$ into inactive coordinates ($i \notin S$, where $s_i = |S|$) and active coordinates ($i \in S$, where $s_i = |S| - 1$). This splits the sum into two components, yielding the joint interference constraint with failure probability at most δ_{int}:

$$\begin{aligned}&\mathbb{P}_{A,x}\left(\max_{i \in \mathcal{I}} |I_i| \geq (1-\alpha)\varepsilon\right) \\ &\leq \underbrace{2(m - |S|)\exp\left(-\frac{d(1-\alpha)^2\varepsilon^2}{2|S|}\right)}_{\text{inactive}} + \underbrace{\mathbb{I}_{|S|\geq 2} \cdot 2|S| \exp\left(-\frac{d(1-\alpha)^2\varepsilon^2}{2(|S|-1)}\right)}_{\text{active}} \\ &\leq 2m\exp\left(-\frac{d(1-\alpha)^2\varepsilon^2}{2|S|}\right) \leq \delta_{\text{int}}\end{aligned} \tag{18}$$

The *inactive* term controls the interference budget on inactive coordinates, where the interference is generated by all $s_i = |S|$ active features, whose erroneous inclusion would produce false positives. The *active* term controls the interference budget on active coordinates, where the interference is generated by the remaining $s_i = |S| - 1$ active features, whose erroneous exclusion would produce false negatives. Here, the term vanishes when $|S| = 1$. The union bound includes both terms, so the resulting condition controls the interference failures relevant to both error types. Inverting the rightmost inequality for d yields the Lemma 1.

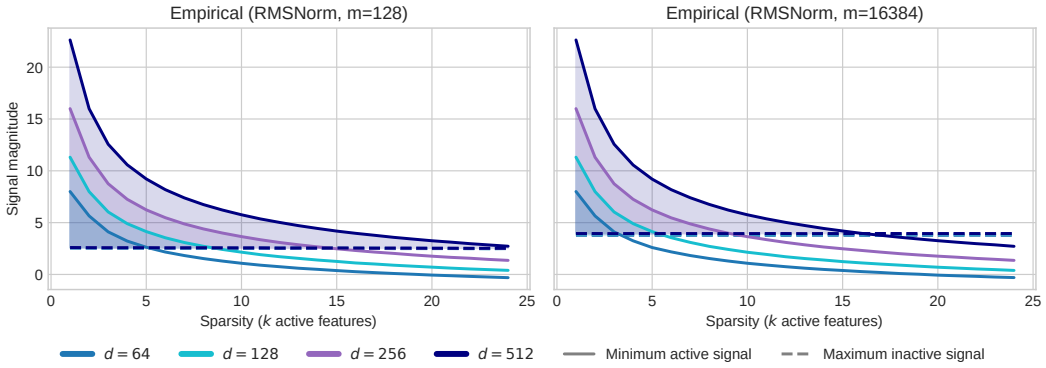


Figure A.1: Support recovery in an RMSNorm vector space for different dictionary sizes in a noiseless setting. Consistent with the theoretical high-probability bound $d = O_\epsilon(k \log m)$, the empirical recovery gap visualizes that linear accessibility supports exponentially many features. Success rates are evaluated across varying ambient dimensions d and sparsities k.

A.3 Gaussian approximation of δ_{INT}

Recall that interference $I_i = \sum_{j \in S \setminus \{i\}} x_j U_j$, where $U_j = \langle a_i, a_j \rangle$, is a linear combination of 1D projections. Since noise is absent, we simplify $(1 - \alpha)\varepsilon = \varepsilon$. By the projective central limit theorem (Theorem 3.3.9), 1D marginals of vectors uniformly distributed on the unit sphere converge to a normal distribution. Because $|x_j| \leq 1$, summing the independent projection variances of $1/d$ across the s_i active interfering features results in a worst-case total interference variance of $\sigma^2 = s_i/d$. Standardizing $I_i/\sigma \approx \mathcal{N}(0, 1)$ and defining the normalized threshold $t = \varepsilon/\sigma = \varepsilon\sqrt{d/s_i}$, we apply the Gaussian tail bound (Proposition 2.1.2) to approximate the two-sided failure probability:

$$\begin{aligned} \mathbb{P}(|I_i| \geq \varepsilon) = \mathbb{P}(|I_i/\sigma| \geq t) &\approx 2\left(\frac{1}{t\sqrt{2\pi}}\right) e^{-t^2/2} \\ &= 2\left(\frac{\sqrt{s_i}}{\varepsilon\sqrt{2\pi d}}\right) \exp\left(-\frac{d\varepsilon^2}{2s_i}\right) \end{aligned} \tag{19}$$

For comparison, we define a tail-matching factor by matching the Gaussian approximation to the form of the Hoeffding envelope (Equation 16). By allocating the entire error budget to the interference term, the threshold for the Hoeffding envelope becomes $t = \varepsilon$. Subsequently isolating a tail-matching factor c_{tail} yields:

$$\begin{aligned} 2\exp\left(-c_{\text{tail}}\frac{d\varepsilon^2}{s_i}\right) &= 2\left(\frac{\sqrt{s_i}}{\varepsilon\sqrt{2\pi d}}\right)\exp\left(-\frac{d\varepsilon^2}{2s_i}\right) \\ -c_{\text{tail}}\frac{d\varepsilon^2}{s_i} &= \ln\left(\frac{\sqrt{s_i}}{\varepsilon\sqrt{2\pi d}}\right) - \frac{d\varepsilon^2}{2s_i} \\ c_{\text{tail}} &= 0.5 - \frac{s_i}{d\varepsilon^2}\ln\left(\frac{\sqrt{s_i}}{\varepsilon\sqrt{2\pi d}}\right) \end{aligned} \tag{20}$$

To understand how an absolute constant can bound the tail-matching factor from below, let's substitute $\varepsilon\sqrt{d/s_i} = t$ back, yielding:

$$c_{\text{tail}} = 0.5 - \frac{1}{t^2}\ln\left(\frac{1}{t\sqrt{2\pi}}\right) \tag{21}$$

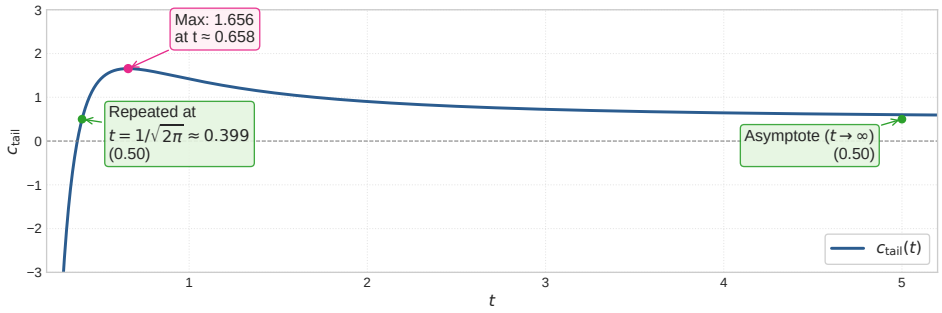


Figure A.2: $c_{\text{tail}}(t)$ is bounded from below by 0.5 for $t = \varepsilon\sqrt{d/s_i} \geq 1/\sqrt{2\pi}$.

To compute the Gaussian approximation for δ_{int}, we plug c_{tail} back into the Hoeffding envelope, and apply the union bound $M \in \{(m - |S|), |S|\}$ over the exponent:

$$\delta_{\text{gauss}} = 2M \exp\left(-\max\left(0.5, 0.5 - \frac{s_i}{d\varepsilon^2}\ln\left(\frac{\sqrt{s_i}}{\varepsilon\sqrt{2\pi d}}\right)\right)\frac{d\varepsilon^2}{s_i}\right) \tag{22}$$

This approximation captures the scaling dynamics of the system parameters. The floor of 0.5 within the exponent ensures that the numerical estimate does not exceed the baseline Hoeffding bound. The c_{tail} term is then utilized to occasionally get a better numerical estimate, yielding more accurate numerical results in regimes where the Gaussian tail decays faster than the worst-case Hoeffding envelope. In configurations where the number of features is small relative to the dimension size, the exponential decay yields a high probability of successful decoding; and instead when dimension size is relatively smaller, the probability decreases accordingly.

A.4 Noise Concentration

Conditional on A, because the projected noise $N_i = \langle a_i, z\rangle = \sum_{n=1}^{d} a_{i,n}z_n$ is a sum of independent, mean-zero subgaussian variables, we can apply the subgaussian Hoeffding inequality (Theorem 2.7.3), where $c > 0$ is a constant:

$$\mathbb{P}_z\left(|N_i| \geq t \mid A\right) \leq 2\exp\left(-\frac{ct^2}{\sum_{n=1}^{d}\|a_{i,n}z_n\|_{\psi_2}^2}\right) \tag{23}$$

By the definition of our noise class $\mathcal{Z}_\sigma$, the subgaussian norm of each coordinate is $\|z_n\|_{\psi_2} \leq \sigma$. When factoring out $a_{i,n}$ and using the fact that a_i is a unit vector ($\|a_i\|_2 = 1$), we get:

$$\sum_{n=1}^{d}\|a_{i,n}z_n\|_{\psi_2}^2 \leq \sigma^2\sum_{n=1}^{d}a_{i,n}^2 = \sigma^2 \tag{24}$$

Substituting Equation 24 into Equation 23 with our allocated error threshold $t = \alpha\varepsilon$, we obtain:

$$\mathbb{P}_z\left(|N_i| \geq \alpha\varepsilon \mid A\right) \leq 2\exp\left(-\frac{c\alpha^2\varepsilon^2}{\sigma^2}\right) \tag{25}$$

Applying a union bound over all m coordinates, analogous to Equation 17, yields the conditional noise constraint:

$$\mathbb{P}_z\left(\max_{i\in[m]}|N_i| \geq \alpha\varepsilon \,\middle|\, A\right) \leq \sum_{i=1}^{m}\mathbb{P}_z\left(|N_i| \geq \alpha\varepsilon \,|\, A\right) \leq 2m\exp\left(-\frac{c\alpha^2\varepsilon^2}{\sigma^2}\right) \leq \delta_{\text{noise}} \tag{26}$$

Since this conditional bound holds uniformly for every realization of A, it also holds under the joint distribution of (A, z), with failure probability at most δ_{noise}. Inverting the rightmost inequality for σ yields the Lemma 2.

B DICTIONARY RECOVERY

To measure dictionary recovery across varying dimensions ($d \in \{64, 128, 256\}$) and sparsities ($k \in [1, 50)$) for a dictionary of size $m = 1024$, we employ stochastic alternating minimization for up to 100 iterations using batches of $N = 50,000$ samples. The samples are generated via $Y = D_{\text{true}}X$ (general vector space), where k active coefficients per sample are drawn from a normal distribution $\mathcal{N}(0, 1)$ with uniformly random support.

Sparse codes are estimated using OMP (constrained to k non-zeros), FISTA (50 iterations with Nesterov acceleration and spectral norm-derived step sizes), and ISTA (unrolled for 10 iterations using weight matrices derived from the current dictionary), with both FISTA and ISTA optimizing the LASSO objective ($\lambda = 0.5$).

The dictionary is then updated from accumulated sufficient statistics XX^T and YX^T using ridge-regularized MOD with a penalty of 10^{-4}, followed by ℓ_2 column normalization (Engan et al., 1999). Degenerate atoms ($||d_i||_2 < 10^{-8}$) are dynamically reinitialized from random batch samples, and training halts early if the maximum column shift falls below 10^{-5}. Recovery is evaluated up to sign and permutation ambiguities by applying the Hungarian matching algorithm to the cost matrix $-|D_{\text{true}}^T D_{\text{pred}}|$.

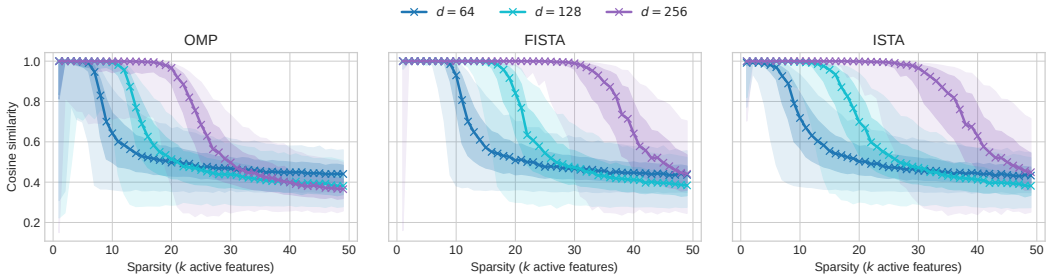


Figure A.3: Dictionary recovery using OMP, FISTA and ISTA, where the last two optimize a LASSO objective. Figures present the median, 5th, 25th, 75th, and 95th percentiles of the matched cosine similarities between corresponding ground-truth and learned dictionary columns in a noiseless setting. Feature directions and their coefficients are drawn from a normal distribution. Details in Section B.

C FEATURE RECOVERY USING SAEs

To measure SAE performance across varying sparsities ($k \in [1, 50)$) and noise levels ($\{0, 0.1\}$) for a dictionary of size $m = 1024$ embedded in $d = 128$ dimensions, we train models for $10,000$ steps using batches of 1024 samples. Samples are generated as $Y = D_{\text{true}}X + z$ (general vector space), optionally followed by RMSNorm, where the number of active features is drawn uniformly from $[1, k]$ during training and fixed to k during evaluation, and z is Gaussian noise scaled by the signal variance. Active coefficients are sampled from $|x| \sim |\mathcal{N}(0, 1)| + 0.1$.

We evaluate an iterative TopK SAE (unfolded with learned step sizes) and MP-SAE (utilizing sequential matching pursuit residual updates) across $n \in \{0, 1, 2, 4, 8, 16, 32\}$ iterations, where $n = 0$ is equivalent to a standard single-step TopK SAE. Models are optimized using Adam (learning rate 10^{-3}, cosine annealing) to minimize mean squared error alongside an auxiliary residual-based reconstruction loss for dead latent revival. The decoder dictionary columns are ℓ_2-normalized after every parameter update. Performance is evaluated across 2048 test samples using the F1-score for active support recovery.

C.1 OPTIMAL ENCODER BASELINE

We compute a coefficient vector c_S that maximizes the separation margin between active and inactive features for a single-step matched-filter readout. For a true active signal x_S with support S and active dictionary columns A_S, we optimize c_S such that the minimum active readout score exceeds the maximum inactive score by a target margin γ. Given the readout scores $s = A^\top A_S c_S$, the objective minimizes the margin violation:

$$\min_{c_S} \max(0, \gamma - (\min_{i \in S} s_i - \max_{j \notin S} s_j))$$

We solve this objective iteratively using the Adam optimizer with a learning rate of 0.01 for 200 iterations, setting the target margin $\gamma = 0.1$.

Simply put, when the true coefficient x_j is 1, but is measured $1 - 1 = 0$ because of interference, we add 1 to the transmitted signal so it measures $2 - 1 = 1$, which is the ground truth. However, increasing c_j might change the measurement for another active or inactive feature. This optimizes the coefficients for all active and inactive features simultaneously.

C.2 SAE IN GENERAL VECTOR SPACE

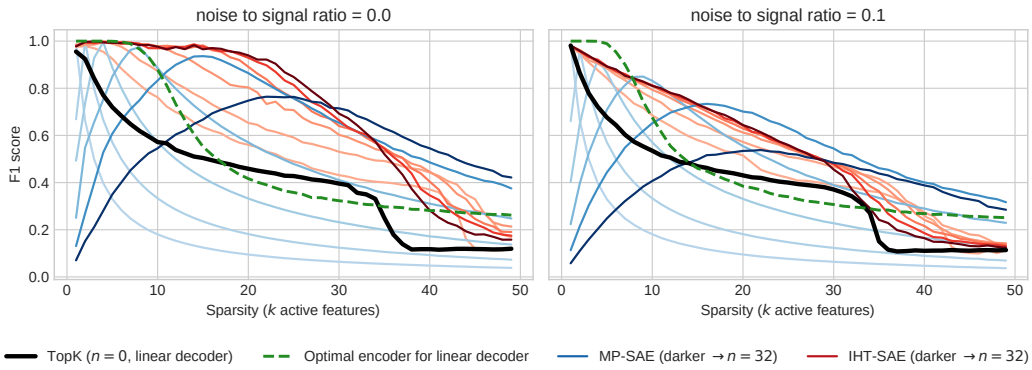


Figure A.4: Support recovery using TopK, IHT-SAE and MP-SAE in general vector space ($m = 1024$, $d = 128$). Curves show F1 scores versus active sparsity k at noise-to-signal power ratios of 0 and 0.1. Darker shades indicate more iterations ($n \in \{1, 2, 4, 8, 16, 32\}$); black denotes single-step TopK ($n = 0$).

C.3 IHT-SAE ALGORITHM

TopK-SAE and IHT-SAE

Input: x; encoder W_e; decoder D; biases $b_{\text{pre}}, b_{\text{lat}}$; sparsity budget k; IHT step sizes $\{\eta_t\}_{t=1}^{T}$.

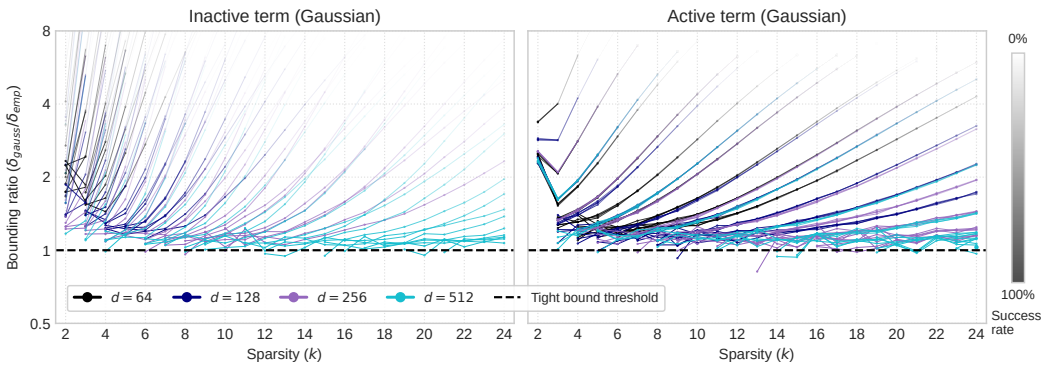
function $\text{TopKReLU}(v, k)$
  $S \leftarrow$ extract indices of the k largest entries in v
  $h \leftarrow$ initialize zero vector of the same shape as v
  $h_S \leftarrow \text{ReLU}(v_S)$
  return h
end function

TopK-SAE
$y \leftarrow x - b_{\text{pre}}$
$h_0 \leftarrow \text{TopKReLU}(W_e y + b_{\text{lat}}, k)$
$\hat{x}_{\text{TopK}} \leftarrow D h_0 + b_{\text{pre}}$

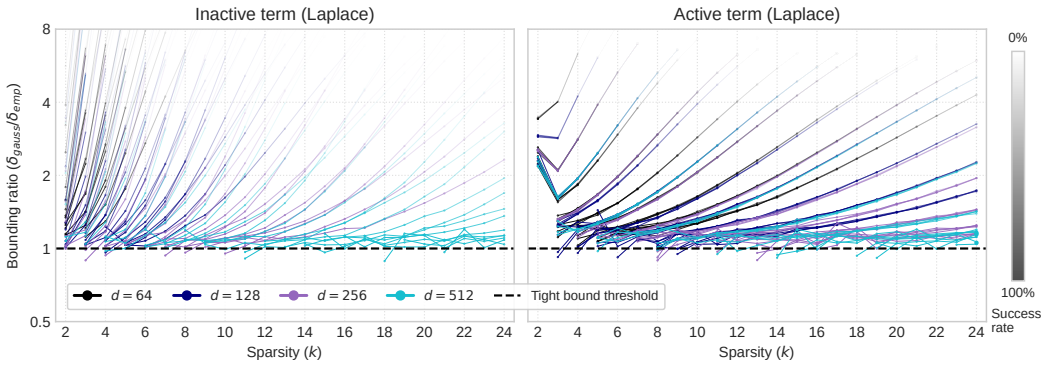
IHT-SAE
$h \leftarrow h_0$
for $n = 1, \ldots, N$ **do**
  $r \leftarrow y - Dh$
  $h \leftarrow \text{TopKReLU}(h + \eta_n D^\top r, k)$
end for
$\hat{x}_{\text{IHT}} \leftarrow Dh + b_{\text{pre}}$

Figure A.5: TopK-SAE and IHT-SAE forward passes. IHT-SAE refines a TopK initialization through residual-based updates; $N = 0$ recovers TopK-SAE. TopK selects the largest signed values before applying ReLU, retaining at most k active coefficients.

D GAUSSIAN APPROXIMATION ON OTHER DISTRIBUTIONS



(a) Gaussian distribution



(b) Laplace distribution

Figure A.6: Tightness of Gaussian approximation for high-probability decoding on other distributions. $40,000$ trials were run for every combination of sparsity $k \in [2, 24]$, dictionary size $m \in \{512, 1024, 2048, 4096\}$ and dimension $d \in \{64, 128, 256, 512\}$, evaluated against various error thresholds $\varepsilon \in \{0.1, 0.2, \ldots, 0.9\}$.