Feature Recovery Requires Structured Event Regimes in Sparse Reconstruction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Sparse autoencoders are often used in mechanistic interpretability as if sparse reconstruction should recover the features represented by a model. Recent work shows that this recovery is fragile, but it remains unclear which failures come from the SAE architecture, the encoder, optimization, or finite data. We show that several failures can already be incentivized by the population-level sparse-reconstruction objective. We study by how much residual mass projects above the sparsity threshold in a positive linear latent-ray model; merging and splitting arise as static properties of this objective, while absorption and seed-dependent alternatives arise sequentially as earlier selections change the residual field. We also separate two notions often conflated in interpretability practice: recovering a ground-truth direction and recovering the activation pattern of that feature. Neither implies the other in general; they coincide under specific structural conditions, such as single-feature event dominance and regular simplex structure in learned symmetric geometries. Sparse reconstruction therefore recovers ground-truth features only in structured event regimes; outside them, the objective can favor non-canonical but reconstruction-useful directions, and a one-ReLU encoder introduces a further representability gap governed by whether the oracle gate is affine-ReLU approximable. Overall, our results refine the existing analysis of SAE behavior and provide a unified perspective on ground-truth feature recovery studies.

1 Introduction

Mechanistic interpretability often studies neural representations through the lens of features: directions or low-dimensional components that correspond to meaningful computational variables [23, 17]. Sparse autoencoders (SAEs) are widely used to extract such features from dense model activations by reconstructing activations using a sparse nonnegative code [3, 12]. This practice implicitly treats SAE latents as candidates for the features used by the model.

Recent work shows that this interpretation is fragile. SAEs can learn non-canonical features [20], learn different dictionaries from the same data under different random seeds [26], and exhibit pathologies such as feature absorption [8]. Toy-model recovery studies sharpen the point: recovery is possible only under restrictive sparsity, frequency, or hyperparameter alignment assumptions [9, 6, 11], and high reconstruction quality or decoder recovery need not identify the latent variables [28, 31, 19].

The question of this paper is therefore not whether SAEs can sometimes recover features. They can. The question is what the sparse-reconstruction objective incentivizes, even before finite data, encoder amortization, and the optimization landscape enters. The central object of our study is the population one-slot utility

$$\mathcal{U}_\lambda(v; R) = \frac{1}{2}\mathbb{E}[(\langle v, R\rangle - \lambda)_+^2],$$

Submitted to 40th Conference on Neural Information Processing Systems (NeurIPS 2026). Do not distribute.

which measures the improvement from adding a nonnegative ℓ_1-penalized code along direction v against residual R. Its support-event decomposition shows that the objective scores directions by thresholded residual mass on latent activation events. Sparse reconstruction is therefore not directly a feature oracle; it is an event-weighted thresholded residual objective.

The analysis separates several objects that are often conflated. A teacher ray may be visible in the data, but visibility is only raw evidence. The sparse-coding objective then decides whether that evidence is useful enough above threshold. The SAE encoder then has to realize the corresponding gate. Finally, even a recovered decoder direction does not always mean that feature's activation pattern would also be recovered. Failure modes appear when one object holds but the next one fails:

$$\text{visibility} \not\Rightarrow \text{preference} \not\Rightarrow \text{realizability} \not\Rightarrow \text{direction recovery} \not\Rightarrow \text{support recovery}.$$

Each missing implication has a concrete mechanism. Mixed above-threshold events create merging and hedging-like pressure; event families with varying residual directions create splitting; earlier surrogate features suppress later evidence, producing absorption-like false negatives; tied or near-tied event utilities create residual-path alternatives. The following sections establish these points and the appendix gives the corresponding propositions and proofs.

Our contributions are:

- We identify the one-slot utility and its support-event decomposition as the population objects controlling one-slot nonnegative sparse reconstruction.
- We show that one important recovery regime occurs when singleton rank-one events dominate off-singleton contamination, forcing maximizers to localize near teacher directions.
- We identify mechanisms of the population sparse reconstruction objective that can underlie several SAE failure modes: merging, splitting, absorption-like false negatives, and residual-path alternatives can arise before encoder amortization or optimization dynamics.
- We separate oracle preference from one-ReLU SAE realizability, and decoder direction recovery from support recovery.
- We provide a unified interpretation for recent toy-model recovery results, connecting their observations under a common thresholded utility framework.

Sparse reconstruction does not ask "what are the true features?" It asks which directions reduce thresholded residual loss. Feature recovery happens only when the data distribution makes true feature directions win that scoring game, and when the SAE can realize the winning gate.

The paper proceeds as follows. Section 2 defines the positive latent-ray residual model. Section 3 derives the one-slot utility and support-event decomposition. Section 4 explains recovery and objective-level failure modes in the oracle setting. Section 5 relates the oracle to one-ReLU SAE encoders. Section 6 separates direction recovery from support recovery and points to Section L for observer-only diagnostics when ground-truth supports are unavailable.

2 Exact positive latent-ray model setup

We write $[F] := \{1, \dots, F\}$ as a feature index set and $S^{d-1} := \{v \in \mathbb{R}^d : \|v\| = 1\}$ as a unit sphere. All probabilities and expectations are taken with respect to the population law of the random activation vector x and the latent coefficients that generate it. We assume finite second moments, for example $\mathbb{E}\|x\|^2 < \infty$, so that all squared losses and utilities below are finite.

We use the linear representation hypothesis, widely adopted in related literature [17, 7, 28, 6, 9, 11, 18], as background for our theoretical setup. Throughout, x follows the positive latent-ray model

$$x = \sum_{i=1}^{F} \alpha_i w_i, \qquad \alpha_i \geq 0, \tag{1}$$

where $x \in \mathbb{R}^d$ is the observed activation vector, α_i are nonnegative random scalars, $w_i \in \mathbb{R}^d \setminus \{0\}$ are fixed teacher directions, and the rays $\mathbb{R}_+ w_i := \{t w_i : t \geq 0\}$ are distinct. The latent support

is $S := \{i : \alpha_i > 0\} \subseteq [F]$. For a context index set $B \subseteq [F]$, chosen or given, and any remaining subset $T \subseteq [F] \setminus B$, write $E_T(B) := \{S \setminus B = T\}$. Let

$$V_B := \operatorname{span}\{w_i : i \in B\}, \qquad P_B : \mathbb{R}^d \to V_B, \qquad \pi_B := I - P_B : \mathbb{R}^d \to V_B^{\perp}$$

be the orthogonal projections onto the context subspace and its orthogonal complement. Throughout fixed-B statements are understood on non-vacuous residual charts, i.e. $\dim(V_B) < d$. We define the projected residual

$$R_B := \pi_B(x) = \sum_{i \notin B} \alpha_i \tilde{w}_i, \qquad \tilde{w}_i := \pi_B(w_i).$$

On $\{R_B \neq 0\}$, the normalized residual direction is $U_B := R_B / \|R_B\|$.

The context B is a way of asking a local question: after some directions have been accounted for, what residual evidence remains? If the residual on some event points along a single ray, then that event is clean evidence for a direction. If several support patterns produce the same or nearby residual direction, the observed activation law can support a direction without uniquely identifying the latent support semantics behind it.

With this residual chart fixed, visibility has a precise meaning. For $j \notin B$ with $\tilde{w}_j \neq 0$, write $u_j := \tilde{w}_j / \|\tilde{w}_j\|$. We say that the residualized teacher ray $\mathbb{R}_+ \tilde{w}_j$ is *visible at stage* B if the law of U_B has positive mass at u_j: $\mathbb{P}(U_B = u_j) > 0$. More generally, an observable direction $u \in S^{d-1}$ is visible at scale ε if $\mathbb{P}(\angle(U_B, u) \leq \varepsilon) > 0$.

The clean regime is a rank-one event: the residual points along one ray throughout the event.

Definition 2.1 (Rank-one event)**.** Let $u \in S^{d-1}$ and let $A \subseteq \{R_B \neq 0\}$ be measurable. A is a *rank-one event* for the ray $\mathbb{R}_+ u$ if $R_B = \xi u$ with probability one on A for some nonnegative random scalar ξ, and $\mathbb{P}(A \cap \{\xi > 0\}) > 0$.

Singleton support events $E_{\{j\}}(B)$ are the canonical rank-one events: on $E_{\{j\}}(B)$, the residual is $R_B = \alpha_j \tilde{w}_j$, so $U_B = u_j$ whenever $\alpha_j > 0$. Thus singleton events make teacher rays directly visible at the population level unless another support family produces residuals on the same ray; this is analogous to the separability condition in NMF-style dictionary learning [15].

We study a clean positive linear model so that failures cannot be blamed on nonlinearities, signs, finite data, or bad optimization. The key observable is the residual direction after removing a context subspace. Singleton residual events are the clean evidence for teacher rays.

3 What does ℓ_1 sparse coding prefer?

For a residual random vector $R \in \mathbb{R}^d$ with $\mathbb{E}\|R\|^2 < \infty$, a unit direction $v \in S^{d-1}$, and sparsity weight $\lambda \geq 0$, we define the one-slot loss

$$\mathcal{L}(v, a; R) := \mathbb{E}\left[\frac{1}{2}\|R - v\, a(R)\|_2^2 + \lambda\, a(R)\right]$$

over nonnegative measurable scalar codes $a(R) \geq 0$ with finite second moment. Optimizing this code gives the oracle utility.

This is the first place where the paper's interpretation changes. The one-slot analysis says that a direction is valuable when enough residual samples project above the sparsity threshold onto it.

Theorem 3.1 (Exact oracle utility)**.** *For fixed $v \in S^{d-1}$, the optimal nonnegative one-slot code is $a_v^\star(R) = (\langle v, R\rangle - \lambda)_+$, and its gain over using no feature is*

$$\mathcal{U}_\lambda(v; R) = \frac{1}{2}\mathbb{E}\big[(\langle v, R\rangle - \lambda)_+^2\big]. \tag{2}$$

The positive part in $(\langle v, R\rangle - \lambda)_+$ says that a sample contributes nothing unless its projection onto v beats the sparsity threshold. The square then makes large above-threshold residuals especially

valuable. Thus sparse coding does not first ask whether v is a teacher atom, it asks whether projection onto v exceeds the threshold often enough, and with enough magnitude, to pay for the ℓ_1 penalty. The support-event decomposition then asks where those above-threshold projections come from: singleton events, mixed events, or broader families of events.

Theorem 3.2 (Support-event decomposition)**.** *The utility decomposes as a sum over residual support events $E_T(B)$. The index $T = \varnothing$ denotes the empty event that contributes no utility:*

$$\mathcal{U}_\lambda(v; R_B) = \sum_{\varnothing \neq T \subseteq [F] \setminus B} \frac{1}{2} \mathbb{E}\left[\left(\sum_{i \in T} \alpha_i \langle v, \tilde{w}_i \rangle - \lambda \right)_+^2 \mathbf{1}_{E_T(B)} \right].$$

Each term in the sum is the thresholded utility contributed by a single latent support family T. If $T = \{j\}$, the term is clean singleton evidence for teacher j. If T contains several active atoms, the term is mixed evidence that may reward a non-teacher direction. This is the main organizing object of the paper: recovery and non-recovery are comparisons between eventwise utility contributions.

The sparse-coding objective has a simple population score. A direction is good when many residuals project above the threshold onto it. The decomposition says which latent support events pay for a direction.

4 When does sparse-coding preference recover teacher directions?

The support-event decomposition turns feature recovery into an eventwise dominance question. A teacher direction is preferred in the clean regime where rank-one singleton events dominate all competing residual events. When this dominance fails, the objective can prefer alternatives to atom recovery that are analogous to known SAE failure modes in gradient descent training regimes.

The theorem below gives the positive case: if clean singleton evidence has a margin over everything else, the objective localizes near the teacher direction. The following subsections explain what replaces recovery when that margin is absent. Mixed events create merge pressure, heterogeneous event families create split pressure, and sequential residual subtraction creates holes and path dependence.

4.1 The positive case: rank-one event dominance

Let $u_j = \tilde{w}_j / \|\tilde{w}_j\|$. For an angular scale $\varepsilon > 0$, define the singleton dominance gap

$$\mathfrak{M}^B_{j,\lambda}(\varepsilon) := \underbrace{\frac{1}{2}\mathbb{E}\Big[\big((\alpha_j \|\tilde{w}_j\| - \lambda)_+^2 - (\alpha_j \|\tilde{w}_j\| \cos\varepsilon - \lambda)_+^2\big)\mathbf{1}_{E_{\{j\}}(B)}\Big]}_{\Delta^B_{j,\lambda}(\varepsilon) \text{ singleton angular margin}} - \underbrace{\frac{1}{2}\mathbb{E}\big[\|R_B\|^2 \mathbf{1}_{E_{\{j\}}(B)^c}\big]}_{C_j^B \text{ off-singleton load}}.$$

$\Delta^B_{j,\lambda}(\varepsilon)$ is the singleton utility lost by rotating at least ε away from the teacher direction u_j. C_j^B is a deliberately crude budget for everything outside the singleton event that could compensate for that loss. If the singleton loss is larger than the off-singleton budget, then no far-away direction can win.

Theorem 4.1 (Singleton dominance implies directional localization)**.** *If $\mathfrak{M}^B_{j,\lambda}(\varepsilon) > 0$, then every global maximizer of $v \mapsto \mathcal{U}_\lambda(v; R_B)$ lies inside the spherical ε-ball around u_j.*

Thus ground-truth direction preference is a dominance regime: singleton rank-one evidence must exceed the total off-singleton competition budget. This gives an eventwise interpretation of extreme-sparsity recovery results such as [11]: when singleton events dominate, sparse reconstruction can recover teacher directions. Uniform-margin, low-coherence, support-fragility, random-geometry, and learned-teacher refinements are collected in Sections F and G.

4.2 Merging: one direction compresses several events

In the utility field, *merging* means that one direction receives high utility because it compresses several support events, or several atoms within a mixed event. The sparse-reconstruction objective can then

prefer a reconstruction-useful mixed direction over an individual teacher direction. The formal local certificate is a positive tangent derivative away from a teacher direction, stated in Proposition E.1. Singleton-only populations do not strictly favor merges (Proposition E.2), while independent co-occurrence can favor a merge once the joint projection crosses threshold (Proposition E.3).

The point is not that co-occurrence must be correlated or hierarchical. Correlation is one empirical source of above-threshold mixed events, but the oracle condition is eventwise thresholded transverse residual mass. Put differently, samples that already pay for one direction may also contain residual mass in another direction, so the objective can improve reconstruction by rotating toward a mixture. This is the population version of feature hedging: a learned latent carries a mixture because the mixture improves sparse reconstruction [5].

4.3 Splitting: several directions compress one broad family

Feature splitting can arise at the objective level when a broad event family does not have a fixed residual direction. No single direction compresses all subevents as well as specialized child directions. Proposition E.7 gives the general event-partition certificate, and Proposition E.8 gives a two-subevent angle-and-threshold specialization.

This is distinct from encoder-induced splitting. Objective-level splitting says different subevents prefer different directions even with unrestricted sample-wise gates. Encoder-induced splitting, treated in Section 5, says a single direction may be oracle-preferred while a one-ReLU encoder cannot realize its gate, so several SAE slots may partition the behavior.

The nontechnical picture is that a broad feature can be too geometrically heterogeneous for one direction to represent efficiently. The objective then rewards children not because the semantic feature is false, but because its residual evidence arrives through several different rays.

4.4 Residual suppression and absorption-like false negatives

Merging and splitting are properties of a fixed residual utility field. Absorption-like behavior is different: earlier selections change the residual field. We view absorption as a support-level false-negative pattern where an earlier learned surrogate fires on a subevent, carries part of a broader teacher direction, and suppresses later residual evidence for that direction [8]. The residual-suppression theorem is stated in Theorem E.4; the independent-support construction in Proposition E.5 shows that the same signature can arise without a parent–child generative hierarchy. For an event A, the localized utility used in the formal statement is $\mathcal{U}_{\lambda,A}(v;R) := \frac{1}{2}\mathbb{E}[(\langle v, R\rangle - \lambda)_+^2 \mathbf{1}_A]$. This is the first mechanism where order matters. In a fixed residual field, a teacher direction may have evidence. After a surrogate is selected and subtracted, the same teacher direction may no longer cross threshold on part of its support. The learned feature can then be direction-aligned but have systematic holes in recall.

4.5 Residual-path alternatives and seed sensitivity

Empirical seed instability refers to SAEs trained on the same data with different random seeds learning different feature sets [26]. We isolate a population residual-path non-identifiability that can serve as an objective-level substrate for this sensitivity. When the oracle utility has multiple tied or near-tied maximizers, residual suppression makes early selections persistent: the selected direction removes its event from the above-threshold residual field, so later utilities depend on the earlier choice. Proposition E.6 constructs an exact population model where tied pair directions give inequivalent non-teacher residual paths.

This is not a convergence theorem for SGD. A seed need not create a bad local optimum; in such regimes it can act as a tie-breaker, or near-tie breaker, among residual paths already present in the population objective. Related sparse dictionary learning non-identifiability and multiple-minima phenomena are discussed in [28].

Table 1 summarizes the section. Merging and splitting are fixed-field mechanisms; absorption-like holes and residual-path sensitivity require earlier choices because the residual field itself changes.

Additional oracle-level material is deferred to the appendix; see Proposition D.3, Theorem E.9, and section E.

Table 1: Mechanism map for the oracle sparse-coding framework. Oracle mechanisms are properties of a fixed residual utility field. Sequential mechanisms arise because earlier learned directions change the residual field.

Mechanism	When it appears	Result
Oracle utility: fixed residual field		
Atom recovery	Singleton rank-one events dominate off-singleton contamination at an angular margin, so oracle maximizers localize near a teacher direction.	Theorem 4.1
Merging	Mixed support events yield larger utility than singleton teacher events; locally, samples above threshold along one direction also contain transverse residual mass.	Proposition E.1 and Proposition E.3
Splitting	A broad event family decomposes into rank-one subevents with different residual directions, and specialized children beat any shared direction.	Proposition E.7 and Proposition E.8
Sequential residual path: earlier choices change later utility		
Residual suppression	A learned surrogate fires on an event supporting another direction and suppresses that event's above-threshold projection for the original direction.	Theorem E.4 and Proposition E.5
Residual-path sensitivity	Several directions are tied or nearly tied; selecting one removes its associated event from later above-threshold utility fields.	Proposition E.6

Together, these cases show the shape of the recovery regime. Teacher directions win when rank-one events dominate contamination in thresholded utility. Outside that regime, sparse reconstruction can prefer non-teacher directions that compress mixed events, allocate directions to subevents, or persist by suppressing residual evidence for the original atom. These mechanisms are population oracle and residual-greedy mechanisms, before encoder amortization or optimization dynamics.

Teacher features win only when their clean rank-one events dominate the other support events. If mixed or heterogeneous events carry more above-threshold residual mass, the same objective prefers merged, split, or path-dependent alternatives.

5 Does an SAE realize the oracle preference?

The oracle in Section 3 was allowed to choose a fresh nonnegative code for every sample. A SAE slot is more constrained: its activation must be produced by one encoder row. This adds a new layer to the implication chain. A direction can be preferred by sparse coding, yet fail to be realized by a one-ReLU SAE gate.

Let K be the number of SAE slots. Let $D = [d_1, \dots, d_K] \in \mathbb{R}^{d\times K}$ have unit-norm columns, let $E \in \mathbb{R}^{K\times d}$ be the encoder matrix, let $b_{\text{enc}} \in \mathbb{R}^K$ and $b_{\text{dec}} \in \mathbb{R}^d$ be encoder and decoder biases, and let $\sigma(t) = t_+$ act coordinatewise. The encoder is *untied*: E is not constrained to be equal to $D^\top$. The encoder input is $\nu = x - b_{\text{dec}}$, the code is $z = \sigma(E\nu + b_{\text{enc}}) \in \mathbb{R}^K_+$, and the reconstruction is $\hat{x} = Dz + b_{\text{dec}}$. For slot r, write $R_{-r} := \nu - \sum_{s\neq r} d_s z_s$ and $y_r := d_r^\top R_{-r} - \lambda$.

The blockwise algebra is simple but important. Once the other slots and the decoder direction are fixed, the encoder row is just trying to imitate the oracle threshold target $y_r = d_r^\top R_{-r} - \lambda$, with a nonnegative one-ReLU output.

Theorem 5.1 (Block decomposition)**.** *For a fixed unit decoder direction, fixed other slots, and a fixed sample value of the leave-one-out residual, the slot loss satisfies the pointwise identity*

$$\frac{1}{2}\|R_{-r} - d_r z_r\|^2 + \lambda z_r = \frac{1}{2}\|R_{-r}\|^2 - \frac{1}{2}y_r^2 + \frac{1}{2}(z_r - y_r)^2.$$

Under this conditioning, optimizing one encoder row is squared-error regression onto y_r with a nonnegative one-ReLU output; the unrestricted pointwise nonnegative optimum is $(y_r)_+$.

Thus the sample-wise sparse-coding gate for direction d and residual R is $g_d^\star = (d^\top R - \lambda)_+$. The SAE row can match this oracle only if the same random function can be written as a single affine-ReLU of the encoder input. To state this precisely, define

$$\mathcal{A}_\nu := \{\, (e^\top \nu + b)_+ : e \in \mathbb{R}^d,\ b \in \mathbb{R} \,\} \subset L^2_+(\Omega)$$

and, for square-integrable random variable gate $z \in L^2_+$,

$$\Gamma_\lambda(d, z; R) := \mathbb{E}\left[z(d^\top R - \lambda) - \frac{1}{2} z^2 \right], \qquad \mathcal{A}_\lambda(d; R, \nu) := \sup_{z \in \mathcal{A}_\nu} \Gamma_\lambda(d, z; R).$$

Here $\mathcal{A}_\lambda$ is the one-row SAE version of the oracle utility: same decoder direction, same residual, but the gate must be a single affine-ReLU of ν. The full envelope identity is deferred to Theorem H.2; the main consequence is the closure criterion below.

Theorem 5.2 (Exact one-ReLU realizability and closure criterion)**.** *Assume $R, \nu \in L^2$, and fix $d \in S^{d-1}$. Let $\overline{\mathcal{A}_\nu}^{L^2}$ denote the closure of $\mathcal{A}_\nu$ in $L^2(\Omega)$. Then*

$$\mathcal{A}_\lambda(d; R, \nu) = \mathcal{U}_\lambda(d; R) \quad \Longleftrightarrow \quad g_d^\star \in \overline{\mathcal{A}_\nu}^{L^2}.$$

If the supremum defining $\mathcal{A}_\lambda(d; R, \nu)$ is attained by some $z_{e,b} = (e^\top \nu + b)_+$, then equality holds if and only if $(e^\top \nu + b)_+ = g_d^\star$ a.s. Thus exact one-ReLU realizability is the attained case of the L^2-closure criterion.

This is the encoder-realizability gap. If $g_d^\star \notin \overline{\mathcal{A}_\nu}^{L^2}$, the one-row SAE value is strictly below the sample-wise oracle value even when its decoder direction is correct. If $g_d^\star$ lies in the closure but is not exactly affine-ReLU realizable, the oracle value can be approached but is not attained by a single row. Several slots may then partition a single oracle gate and reduce this gap, producing an architectural splitting mechanism distinct from the event-level splitting in Section 4.3.

Appendix M gives a small sanity check for the realizability layer. We first train SAE1 on the ground-truth generator, then train SAE2 either directly on SAE1's code or on SAE1's reconstructed hidden states. Direct code training is the easy endpoint: SAE2 receives a nonnegative coordinate representation and recovers SAE1 features with support F1 0.937, compared with 0.583 for GT$\to$SAE1 and 0.600 for GT$\to$SAE2. Training on reconstructed hidden states is harder, but the representation-relative pattern remains: SAE1$\to$SAE2 support F1 is 0.804 with learned decoders and 0.592 in the fixed-decoder output-student ablation, compared to the GT$\to$SAE1 values 0.583 and 0.571.

5.1 Transfer from latent residuals to SAE slots

There is one more transfer step. The latent sparse-coding score uses $R_B = \pi_B x$, while slot r sees R_{-r}. Define the perturbation scale $\eta(R, S) := \frac{1}{2}\|R - S\|_{L^2}(\|R\|_{L^2} + \|S\|_{L^2})$, with the full bound stated in Lemma I.2. Let $\eta_{B,-r} := \eta(R_B, R_{-r})$ and define the amortization gap

$$\mathrm{Gap}_\lambda(u; R_{-r}, \nu) := \mathcal{U}_\lambda(u; R_{-r}) - \mathcal{A}_\lambda(u; R_{-r}, \nu),$$

the utility lost at direction u because the gate must be a single affine-ReLU of ν. For a competitor set $\mathcal{C}$, the corresponding amortized SAE margin is

$$\mathrm{Marg}_{\mathcal{C}}^{\mathrm{SAE}}(u) := \mathcal{A}_\lambda(u; R_{-r}, \nu) - \sup_{v \in \mathcal{C}} \mathcal{A}_\lambda(v; R_{-r}, \nu).$$

The latent margin γ is the amount by which a direction beats its competitors in the clean residual field R_B. Two costs spend that margin before it reaches an SAE: residual mismatch and gate amortization.

Theorem 5.3 (Residual-field transfer)**.** *Let $\mathcal{C} \subseteq S^{d-1}$ be a set of competitor directions, and suppose $u \in S^{d-1}$ has latent margin $\gamma := \mathcal{U}_\lambda(u; R_B) - \sup_{v \in \mathcal{C}} \mathcal{U}_\lambda(v; R_B) > 0$. Then $\mathrm{Marg}_{\mathcal{C}}^{\mathrm{SAE}}(u) \geq \gamma - 2\eta_{B,-r} - \mathrm{Gap}_\lambda(u; R_{-r}, \nu)$. Hence the ranking transfers whenever this lower bound is positive.*

The factor $2\eta_{B,-r}$ is the cost of changing residual fields; the gap term is the cost of restricting the gate to one affine-ReLU. Exact residual compatibility is

$$R_{-r} = R_B \quad \Longleftrightarrow \quad b_{\mathrm{dec}} + \sum_{s \neq r} d_s z_s = P_B x \qquad \text{almost surely.}$$

A decoder bias inside V_B leaves the projected residual component unchanged, but a nonzero projected bias shifts thresholds and can create utility on empty residual-support events. Amortization can reverse an oracle ranking only through direction-dependent gaps; affine residual regimes eliminate those gaps, while pair-only worlds provide stress tests where both oracle and exact sparse coding prefer non-ground-truth directions (Sections H.7 and I).

Even if the oracle prefers a direction, an SAE slot still has to realize the right gate using a one-ReLU encoder row and the leave-one-out residual. Oracle preference and SAE realization are separate layers.

6 Do recovered directions imply support semantics?

The previous section separated oracle preference from encoder realizability. The next separation is semantic: even if a decoder direction is recovered, this need not mean that its gate fires on the intended latent examples. Direction recovery and support recovery are different objects.

The reason is visible in one moment equation. Let $a(x) \geq 0$ be an activation such that $\mathbb{E}[\|R_B\|a] < \infty$ and $\mathbb{E}[\alpha_i a] < \infty$ for every $i \notin B$. Define $\mu(a; B) := \mathbb{E}[R_B a] = \sum_{i \notin B} \mathbb{E}[\alpha_i a]\, \tilde{w}_i$.

When the decoder is fit by least squares to the gate, and in the corresponding stationary condition for the SAE decoder, the decoder direction aligns with this gated residual moment. The decoder therefore sees an average vector weighted by the gate. If the gate also fires when other atoms are active, their residual directions enter the average; if it fires on only a tiny but geometrically pure subset, the direction can look correct while support recall is poor.

Theorem 6.1 (Moment characterization of direction alignment)**.** *For $j \notin B$ with $\tilde{w}_j \neq 0$, let $a \geq 0$ satisfy $\mathbb{E}[\|R_B\|a] < \infty$ and $\mathbb{E}[\alpha_i a] < \infty$ for every $i \notin B$, and suppose $\mu(a; B) \neq 0$. Then $\mu(a; B) \in \mathrm{span}(\tilde{w}_j)$ iff the off-j gated moment lies in $\mathrm{span}(\tilde{w}_j)$. If the residualized atoms are linearly independent, this is equivalent to $\alpha_i a = 0$ a.s. for every $i \notin B$, $i \neq j$.*

Under linear independence, exact nonzero decoder alignment with a teacher ray is equivalent to zero gated co-activation with every other residualized atom. This is stronger than "the gate fires when the atom is present." It is also different from support recovery. The minimal counterexamples in Propositions J.1 and J.2 show both directions of failure: a gate can recover direction with arbitrarily poor support recall, and a support-faithful gate can yield a rotated decoder direction.

The claim that direction recovery implies support recovery is also disproven in [31]; here we isolate the population event-geometry behind this separation. Below, we show that the two objects coincide in the clean regime already used for oracle recovery: rank-one event structure.

Theorem 6.2 (Coincidence on rank-one events)**.** *If A is a rank-one event with $R_B = \xi u$ and the gate $a \geq 0$ is supported on A with $0 < \mathbb{E}[\xi a \mathbf{1}_A] < \infty$, then $\mu(a; B)$ is parallel to u.*

This is only a sufficient coincidence regime: event support plus positive gated amplitude on a rank-one residual event yields direction alignment. It does not say that arbitrary gates aligned with u recover A. The robust moment bound and support-fragility refinements are in Section J; learned-teacher simplex symmetry gives another special coincidence regime by cancelling exchangeable off-target moments (Section G). These are structured exceptions, not generic consequences of sparse reconstruction.

6.1 Sparsity matching is not support-event matching

Recent toy-model work emphasizes sparsity-level and firing-frequency alignment as recovery-relevant quantities [6, 9]. In our framework these quantities are real necessary shadows of support recovery,

but they are not the support-event law itself. Let $Z_i := \mathbf{1}_{\{\alpha_i>0\}}$, $S := \{i : Z_i = 1\}$, $N := |S|$, $p_i := \mathbb{P}(Z_i = 1)$, and $s_\star := \mathbb{E}N$. Exact support recovery implies samplewise sparsity recovery, which implies expected sparsity matching: $\widehat{S} = S$ a.s. $\Longrightarrow |\widehat{S}| = |S|$ a.s. $\Longrightarrow \mathbb{E}|\widehat{S}| = s_\star$. The converses fail because expected sparsity fixes only $\sum_i p_i$, and marginal-frequency matching fixes only the one-feature marginals. The sparse-reconstruction objective decomposes over full support events $E_T(B)$, so it depends on joint probabilities $\mathbb{P}(S = T)$, eventwise amplitudes, and residual geometry. The missing information already appears at second order:

$$\mathrm{Var}(N) = \sum_i p_i(1-p_i) + 2\sum_{i<j}\mathrm{Cov}(Z_i, Z_j), \qquad \mathbb{E}\binom{N}{2} = \sum_{i<j}\mathbb{P}(Z_i = Z_j = 1).$$

Pairwise co-activation events are exactly the lowest-order mixed support events in Theorem 3.2. They can be invisible to a global sparsity target while still dominating the thresholded utility. Thus sparsity and frequency matching can remove one mismatch between teacher and learned codes, but they do not replace support-event dominance. Section K gives the exact two-feature marginal-frequency separation, amplitude-threshold mismatch under the correct orthogonal dictionary, power-law asymptotics, load transitions, and pairwise-only support constructions.

Appendix M also illustrates the direction/support distinction. In fixed-decoder ablations, decoder rows are correct by construction. Nevertheless, GT support recovery remains poor: direction-matched GT→SAE2 support F1 is 0.569 when SAE2 is trained on SAE1's code, and 0.391 when SAE2 is trained on SAE1 reconstructed hidden states with the GT decoder fixed. In the latter case, L_0 is also closely matched to the teacher target (1.153 versus 1.175) and explained variance is high (0.968), so the failure is not just a gross sparsity or reconstruction mismatch. The closest support-matched features are geometrically nearby, about 0.94; the issue is that their support events do not match. Also, while GT supports are independend, learned supports have substantial off-diagonal correlations.

The observer-level version is even weaker. In real activations the latent support S is not observed, so hidden states alone cannot identify latent support semantics. What can be certified is residual geometry: observable residual-direction atoms or caps, directional preference of the observed sparse-coding utility field, and whether learned features match those caps by direction, precision, recall, intrinsic ambiguity, and gate consistency. Section L gives the formal observer-only diagnostics.

> A decoder direction and a feature's firing pattern are different semantic objects. Direction alignment is a gated-moment condition; support recovery is an event condition. Sparsity and frequency matching preserve useful low-order statistics, but the objective depends on the full support-event law.

7 Conclusion

This paper isolates what nonnegative sparse reconstruction prefers before an SAE encoder, finite data, or gradient dynamics enter. The one-slot utility $\mathcal{U}_\lambda(v; R) = \frac{1}{2}\mathbb{E}[(\langle v, R\rangle - \lambda)_+^2]$ is a population thresholded gain, and its support-event decomposition shows which latent events make a direction valuable. Teacher directions win only in structured dominance regimes, where rank-one or single-feature events dominate off-target contamination above threshold. When this dominance fails, the same objective can prefer non-canonical but reconstruction-useful directions: mixed events create merging, heterogeneous event families create splitting, earlier surrogates create absorption-like false negatives, and near-tied event utilities create residual-path alternatives.

The claims fit together through the implication chain

$$\text{visibility} \not\Rightarrow \text{preference} \not\Rightarrow \text{realizability} \not\Rightarrow \text{direction recovery} \not\Rightarrow \text{support recovery}.$$

Each missing implication is witnessed by an explicit construction in Section C. Thus sparse reconstruction can recover ground-truth features, but only when the data distribution, threshold, geometry, and encoder class align; outside those regimes, such failures should not be attributed only to bad optimization or architecture, because population objectives can already incentivize them.

The limitations are structural. The nonnegative cone in Equation (1) is essential for the residual-ray and support-semantics interpretation of Theorem 3.2. Signed activations, nonlinear mixing, SAE optimization dynamics and prediction of SAE preferences, finite-sample observer diagnostics, and empirical validation in pretrained-model SAEs remain open. The fixed-B analysis applies only on non-vacuous residual charts with $\dim(V_B) < d$, and target-feature claims require $\pi_B w_j \neq 0$.

References

[1] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing*, 54(11):4311–4322, 2006.

[2] Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, efficient, and neural algorithms for sparse coding. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 113–149, Paris, France, 03–06 Jul 2015. PMLR.

[3] 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.

[4] E.J. Candes and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

[5] David Chanin, Tomáš Dulka, and Adrià Garriga-Alonso. Feature hedging: Correlated features break narrow sparse autoencoders, 2025.

[6] David Chanin and Adrià Garriga-Alonso. Sparse but wrong: Incorrect l0 leads to incorrect features in sparse autoencoders, 2025.

[7] David Chanin and Adrià Garriga-Alonso. Synthsaebench: Evaluating sparse autoencoders on scalable realistic synthetic data, 2026.

[8] David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Isaac Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2026.

[9] Siyu Chen, Heejune Sheen, Xuyuan Xiong, Tianhao Wang, and Zhuoran Yang. Taming polysemanticity in LLMs: Theory-grounded feature recovery via sparse autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2026.

[10] Valérie Costa, Thomas Fel, Ekdeep Singh Lubana, Bahareh Tolooshams, and Demba Ba. Evaluating sparse autoencoders: From shallow design to matching pursuit, 2025.

[11] Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the limits of sparse autoencoders: A theoretical framework and reweighted remedy. In *The Fourteenth International Conference on Learning Representations*, 2026.

[12] Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023.

[13] Jean-François Determe, Jérôme Louveaux, Laurent Jacques, and François Horlin. On the exact recovery condition of simultaneous orthogonal matching pursuit. *IEEE Signal Processing Letters*, 23(1):164–168, 2016.

[14] David L. Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

[15] David L. Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? In *NIPS*, pages 1141–1148, 2003.

[16] D.L. Donoho, M. Elad, and V.N. Temlyakov. Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, 52(1):6–18, 2006.

[17] 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.

[18] Nikhil Garg, Jon Kleinberg, and Kenny Peng. How many features can a language model store under the linear representation hypothesis?, 2026.

[19] Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do saes beat random baselines?, 2026.

[20] Patrick Leask, Bart Bussmann, Michael Pearce, Joseph Bloom, Curt Tigges, Noura Al Moubayed, Lee Sharkey, and Neel Nanda. Sparse autoencoders do not find canonical units of analysis, 2025.

[21] Julien Mairal, Francis R. Bach, Jean Ponce, and Guillermo Sapiro. Online dictionary learning for sparse coding. In *ICML*, pages 689–696, 2009.

[22] S.G. Mallat and Zhifeng Zhang. Matching pursuits with time-frequency dictionaries. *IEEE Transactions on Signal Processing*, 41(12):3397–3415, 1993.

[23] Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. Zoom in: An introduction to circuits. *Distill*, 2020. https://distill.pub/2020/circuits/zoom-in.

[24] Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by v1? *Vision Research*, 37(23):3311–3325, 1997.

[25] Charles O'Neill, Alim Gumran, and David Klindt. Compute optimal inference and provable amortisation gap in sparse autoencoders. In *Forty-second International Conference on Machine Learning*, 2025.

[26] Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features, 2025.

[27] Daniel A. Spielman, Huan Wang, and John Wright. Exact recovery of sparsely-used dictionaries. In Shie Mannor, Nathan Srebro, and Robert C. Williamson, editors, *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 37.1–37.18, Edinburgh, Scotland, 25–27 Jun 2012. PMLR.

[28] Yiming Tang, Harshvardhan Saini, Zhaoqian Yao, Zheng Lin, Yizhen Liao, Jingyi Cui, Yisen Wang, Mengnan Du, and Dianbo Liu. A unified theory of sparse dictionary learning in mechanistic interpretability: Piecewise biconvexity and spurious minima, 2026.

[29] Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, Hoagy Cunningham, Nicholas L Turner, Callum McDougall, Monte MacDiarmid, C. Daniel Freeman, Theodore R. Sumers, Edward Rees, Joshua Batson, Adam Jermyn, Shan Carter, Chris Olah, and Tom Henighan. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024.

[30] J.A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

[31] Zhenyu Zhu, Marco Fumero, Francesco Locatello, and Volkan Cevher. Diagnosing and fixing latent recovery in sparse autoencoders. In *ICLR 2026 Workshop on Unifying Concept Representation Learning*, 2026.

A Appendix roadmap

The appendix collects the formal material behind the main claims. It preserves the deferred theorem and proposition statements, gives complete derivations, includes refinements that clarify positive regimes, and gives stress tests for the boundary between oracle preference and amortized realization.

Section C gives the counterexamples behind the implication chain. Sections D to F contain the oracle derivations, mechanisms, and positive recovery refinements. Section G studies structured learned-teacher geometries. Sections H and I give the SAE bridge and residual-transfer details. Sections J and K contain the direction-support and sparsity-frequency branches. Section L contains observer-only audit refinements. Section M gives synthetic two-stage SAE illustrations for the realizability layer and the direction/support separation.

B Related work

The related literature gives the recovery side of the story and the pathology side of the story. The point of comparison here is the population sparse-reconstruction objective: when the assumptions behind recovery guarantees weaken, which residual events does that objective reward?

> Prior recovery work asks when sparse methods recover features. The complementary question here is what the population sparse-reconstruction objective rewards when recovery conditions do not dominate.

Sparse recovery and dictionary learning. Classical sparse recovery studies when sparse codes or dictionaries are identifiable from data, typically under incoherence, restricted-isometry, separability, or sparse-use assumptions [14, 4, 27, 2]. Classical dictionary-learning algorithms such as sparse coding, K-SVD, and online dictionary learning optimize reconstruction with sparsity constraints and alternate between code inference and dictionary updates [24, 1, 21]. This literature primarily gives sufficient conditions for recovery. Our question is complementary: when recovery conditions do not dominate the population objective, which residual support events does sparse reconstruction reward, and what non-teacher directions can become optimal?

Matching pursuit and greedy sparse approximation. The one-slot oracle in Section 3 is the population analogue of a nonnegative, ℓ_1-thresholded matching-pursuit step. Classical matching pursuit selects atoms by greedy residual reconstruction gain [22]; OMP and related analyses characterize when greedy residual correlations recover the correct support under dictionary assumptions [30, 16]. Simultaneous matching pursuit extends this logic to multiple signals and common-support recovery [13]. Our setting differs in two ways. First, the direction ranges over the sphere rather than over a fixed dictionary. Second, latent supports vary across samples, so the relevant object is not a single support to recover but the support-event decomposition of the population gain. This decomposition identifies when greedy sparse reconstruction favors teacher directions, merged directions, split subevents, absorption-like false negatives, or residual-path alternatives.

Sparse autoencoders and feature recovery. SAEs are widely used to extract interpretable features from neural activations [12, 3, 29]. Recent theory studies when such features recover known ground truth in toy latent models. Recovery can hold under restrictive sparsity, frequency, or hyperparameter alignment assumptions [9, 6, 11, 7]. These results align with our positive regimes: extreme sparsity, matched firing frequencies, and correct sparsity levels are ways of making thresholded support events identify teacher directions. Our contribution is not another recovery guarantee for a specific SAE algorithm; it is an event-level explanation of why the same sparse-reconstruction objective can either support recovery or prefer non-canonical directions.

Non-canonical features and SAE pathologies. Empirical and theoretical work has documented non-canonical SAE features, feature absorption, feature hedging, seed-dependent feature sets, and other failures [20, 8, 5, 26]. We do not claim priority for observing these phenomena. We isolate objective-level mechanisms for a central subset of them. Mixed above-threshold events create merge and hedging-like pressure; variable rank-one subevents create event-partition splitting; earlier surrogates create absorption-like false negatives by suppressing residual evidence; and tied or near-

tied event utilities create residual-path non-identifiability. These mechanisms occur before finite samples, encoder amortization, gradient dynamics, or local minima.

Direction recovery versus code or support recovery. Several recent works emphasize that reconstruction quality or dictionary recovery does not imply recovery of latent variables [28, 31]. Zhu et al. [31] are especially close to our direction/support separation: they show that decoder recovery does not guarantee latent-code recovery. The work [25] also argues that standard SAE encoders cannot optimally represent the underlying code, and that expressivity of encoder class is important for recovery. We refine this distinction for interpretability semantics. Decoder direction recovery is a gated-moment condition, while support recovery is an event-level condition; neither implies the other in general. They coincide only under additional structure, such as rank-one event support or symmetric cancellation in learned geometries.

Matching-pursuit SAEs and broader SDL theory. Recent MP-SAE work uses residual-guided matching-pursuit inference inside SAE architectures to better handle correlated, hierarchical, or conditionally orthogonal features [**?** 10]. This is architecturally adjacent to our oracle, but our use is diagnostic rather than constructive: residual-guided sparse reconstruction can itself prefer non-canonical directions at the population level. This also complements unified SDL theory, which studies optimization landscapes, non-identifiability, and spurious minima for sparse dictionary learning methods [28]. Our analysis is orthogonal: we study the population thresholded sparse-coding gain first, then separately track encoder representability and direction/support semantics.

C Separation witnesses

The implication chain in the introduction needs exact witnesses. Each construction below shows that one semantic object can hold without the next one in the chain, so none of the missing implications is silently assumed.

Each witness breaks one tempting implication: visibility, preference, realizability, direction recovery, and support recovery are different objects.

Proposition C.1 (Separation witnesses)**.** *The implication chain separated in the introduction cannot be collapsed: visibility need not imply oracle preference; oracle preference need not imply one-ReLU realizability; one-ReLU realizability need not imply teacher-direction recovery; teacher-direction recovery need not imply support recovery; and support recovery need not imply teacher-direction recovery.*

Proof. Visibility need not imply preference: in Proposition E.3, with $p \in (0,1)$ and $A \le \lambda < \sqrt{2}A$, the singleton event for u has positive probability, so u is visible, but $\mathcal{U}_\lambda(u;x) = 0 < \mathcal{U}_\lambda((u+c)/\sqrt{2};x)$.

Preference need not imply one-ReLU realizability with respect to the encoder input: take the encoder input to be a scalar random variable ν with $\mathbb{E}[\nu^4] < \infty$ and support containing an interval, set $R(\nu) = \nu^2 u$, and take $\lambda = 0$.

One-ReLU realizability need not imply teacher-direction recovery: in the same independent two-atom affine model used in Proposition E.3, the merged direction $m = (u+c)/\sqrt{2}$ has affine-ReLU gate $(m^\top x - \lambda)_+$ and is oracle-preferred in the threshold regime, but m is not a teacher ray.

Teacher-direction recovery without support recovery is Proposition J.1. Support recovery without teacher-direction recovery is Proposition J.2. □

D Core oracle derivations

These derivations support Sections 2–3. They establish what the oracle utility measures, why residual support events decompose its utility, and how marginal insertion relates the one-slot calculation to sparse-coding improvement.

The central algebra is: singleton events create residual-ray atoms, one-slot sparse coding gives the thresholded utility, and support events split that utility into interpretable pieces.

Proposition D.1 (Residual-ray observability from singleton witnesses)**.** *Fix $j \notin B$. Assume*

$$\tilde{w}_j \neq 0, \qquad \mathbb{P}(S \setminus B = \{j\}) > 0.$$

Then the law of $U_B := R_B/\|R_B\|$ on $\{R_B \neq 0\}$ has an atom at

$$\tilde{u}_j := \frac{\tilde{w}_j}{\|\tilde{w}_j\|}.$$

Moreover, suppose that for every family $T \subseteq [F] \setminus B$ with $\mathbb{P}(E_T(B)) > 0$, the following hold:

$$j \notin T,\ T \neq \emptyset \quad \Longrightarrow \quad \mathbb{R}_+\tilde{w}_j \cap \mathrm{cone}_+\{\tilde{w}_i : i \in T\} = \{0\},$$

and

$$j \in T,\ T \neq \{j\} \quad \Longrightarrow \quad \mathbb{R}\tilde{w}_j \cap \mathrm{cone}_+\{\tilde{w}_i : i \in T \setminus \{j\}\} = \{0\}.$$

Then no other positive residual support family generates the same ray, so the atom is semantically unambiguous relative to the assumed latent factorization.

D.1 Proof of Proposition D.1

On the singleton event $E_{\{j\}}(B)$, one has

$$R_B = \alpha_j \tilde{w}_j,$$

so on $E_{\{j\}}(B) \cap \{\alpha_j > 0\}$,

$$U_B = \frac{\tilde{w}_j}{\|\tilde{w}_j\|} = \tilde{u}_j.$$

Thus the law of U_B assigns at least the positive mass $\mathbb{P}(E_{\{j\}}(B) \cap \{\alpha_j > 0\})$ to the singleton direction $\tilde{u}_j$. This is residual-ray observability, not latent-support identifiability from x alone. The cone-separation condition

$$\mathbb{R}_+\tilde{w}_j \cap \underset{+}{\mathrm{cone}}\{\tilde{w}_i : i \in T\} = \{0\}$$

for every other support family T rules out semantic collisions with other positive support families under the assumed latent factorization.

For $T \not\ni j$, the first cone condition rules out $R_B \in \mathbb{R}_+\tilde{w}_j$ on $E_T(B)$. For $T \ni j$, $T \neq \{j\}$, write

$$R_B = \alpha_j \tilde{w}_j + \sum_{i \in T \setminus \{j\}} \alpha_i \tilde{w}_i.$$

If $R_B \in \mathbb{R}_+\tilde{w}_j$, then

$$\sum_{i \in T \setminus \{j\}} \alpha_i \tilde{w}_i \in \mathbb{R}\tilde{w}_j,$$

contradicting the second cone condition.

D.2 Proof of Theorem 3.1

For fixed $R = r$,

$$\frac{1}{2}\|r - va\|_2^2 + \lambda a = \frac{1}{2}\|r\|_2^2 - a\langle v, r\rangle + \frac{1}{2}a^2 + \lambda a.$$

This is a strictly convex quadratic in $a \geq 0$. The unconstrained minimizer is $\langle v, r\rangle - \lambda$, so the constrained minimizer is $(\langle v, r\rangle - \lambda)_+$. Substituting back yields the gain

$$\frac{1}{2}(\langle v, r\rangle - \lambda)_+^2.$$

Taking expectations gives the stated utility.

D.3 Proof of Theorem 3.2

The support events $\{E_T(B)\}_{T\subseteq[F]\setminus B}$ form a measurable partition. On $E_T(B)$,

$$R_B = \sum_{i\in T} \alpha_i \tilde{w}_i.$$

Substitute this into Theorem 3.1 and sum over the partition. Thus

$$\mathcal{U}^B_{\lambda,T}(v) := \frac{1}{2}\,\mathbb{E}\left[\Big(\sum_{i\in T}\alpha_i\langle v,\tilde{w}_i\rangle - \lambda\Big)_+^2 \mathbf{1}_{E_T(B)}\right],$$

and the empty event contributes zero for the projected residual R_B.

Proposition D.2 (utility gradient)**.** *Let $R \in L^2(\mathbb{R}^d)$. Then*

$$\nabla_v \mathcal{U}_\lambda(v;R) = \mathbb{E}[(\langle v,R\rangle - \lambda)_+ R],$$

and the Riemannian gradient on the sphere is

$$\operatorname*{grad}_{S^{d-1}} \mathcal{U}_\lambda(v;R) = P_{v^\perp}\mathbb{E}[(\langle v,R\rangle - \lambda)_+ R].$$

D.4 Proof of Proposition D.2

Let $\phi(t) = \frac{1}{2}(t-\lambda)_+^2$, so $\phi'(t) = (t-\lambda)_+$. Differentiate under the expectation:

$$D\mathcal{U}_\lambda(v)[\delta] = \mathbb{E}[\phi'(\langle v,R\rangle)\,\langle \delta,R\rangle] = \langle \mathbb{E}[(\langle v,R\rangle - \lambda)_+ R],\,\delta\rangle\,.$$

This gives the Euclidean gradient. Project onto $v^\perp$ to obtain the sphere gradient.

Proposition D.3 (marginal insertion bound)**.** *Let*

$$\Phi_\lambda(D) := \mathbb{E}\left[\min_{z\geq 0} \frac{1}{2}\,\|x - Dz\|_2^2 + \lambda \mathbf{1}^\top z\right].$$

Fix a $(K-1)$-slot dictionary D_{-r} and let $R^\star_{-r}$ be the corresponding leave-one-out exact residual. Then for every unit vector v,

$$\Phi_\lambda(D_{-r}) - \Phi_\lambda(D_{-r}\cup\{v\}) \geq \mathcal{U}_\lambda(v;R^\star_{-r}).$$

D.5 Proof of Proposition D.3

For each sample x, start from the exact $(K-1)$-slot optimum and optimize only the new coefficient $a \geq 0$ along v. By Theorem 3.1, the gain from this restricted reoptimization is exactly

$$\frac{1}{2}(\langle v, R^\star_{-r}(x)\rangle - \lambda)_+^2.$$

The full K-slot optimum may further reoptimize the old coefficients, so the true improvement is at least this large. Take expectations.

E Ideal sparse-coding mechanism details

The mechanism claims from Section 4 are made precise below: first-order merge pressure, event-partition splitting, active-set maxima, and residual suppression. These are the exact formal statements and proofs behind the mechanism narrative.

The same utility field creates merge pressure, split pressure, absorption-like residual suppression, and path-dependent alternatives.

Deferred merge-pressure statement. For $u \in S^{d-1}$ and $h \in u^\perp \cap S^{d-1}$, define the great-circle perturbation $v_\theta := \cos\theta\, u + \sin\theta\, h$.

Proposition E.1 (first-order merge pressure). *Let $u \in S^{d-1}$ and $h \in u^\perp \cap S^{d-1}$. Then*

$$\frac{d}{d\theta}\mathcal{U}_\lambda(v_\theta; R)\Big|_{\theta=0} = \mathbb{E}\big[(\langle u, R\rangle - \lambda)_+ \langle h, R\rangle\big].$$

Consequently, if $\mathbb{E}\big[(\langle u, R\rangle - \lambda)_+ \langle h, R\rangle\big] > 0$, then u is not a local maximizer of $\mathcal{U}_\lambda(v; R)$ on S^{d-1}.

E.1 Proof of Proposition E.1

By Proposition D.2,

$$\nabla_v \mathcal{U}_\lambda(v; R) = \mathbb{E}[(\langle v, R\rangle - \lambda)_+ R].$$

Since v_θ is a great-circle perturbation through u with tangent h,

$$\frac{d}{d\theta}\mathcal{U}_\lambda(v_\theta; R)\Big|_{\theta=0} = \langle \nabla_v \mathcal{U}_\lambda(u; R), h\rangle = \mathbb{E}[(\langle u, R\rangle - \lambda)_+ \langle h, R\rangle].$$

If the derivative is positive, utility increases along that tangent direction, so u is not a local maximizer.

Deferred singleton-only statement. The proposition below shows that merging is a mixed-event phenomenon, when higher-order support events are present.

Proposition E.2 (singleton-only populations do not strictly favor merges). *Let $x = AZ_1 u + AZ_2 c$, where $u \perp c$, $A > 0$, $Z_1, Z_2 \in \{0, 1\}$, and $Z_1 Z_2 = 0$ a.s. Write $\pi_i := \mathbb{P}(Z_i = 1)$. If $0 < \lambda < A$, then*

$$\sup_{v \in S^{d-1}} \mathcal{U}_\lambda(v; x) = \frac{1}{2}\max\{\pi_1, \pi_2\}(A - \lambda)^2.$$

A teacher direction with maximal singleton probability is globally optimal. If $\pi_1 = \pi_2$, both u and c are global maximizers, and no mixed direction strictly improves on them. If $\lambda = 0$, then

$$\sup_{v \in S^{d-1}} \mathcal{U}_0(v; x) = \frac{1}{2}A^2 \max\{\pi_1, \pi_2\}.$$

When $\pi_1 = \pi_2$, every unit direction $v = au + bc$ with $a, b \geq 0$ and $a^2 + b^2 = 1$ is globally optimal, so mixed directions in the positive quadrant can tie but cannot strictly improve over teacher directions. Directions with a negative coordinate in the (u, c) basis are not optimal. If $\lambda \geq A$, then $\mathcal{U}_\lambda(v; x) = 0$ for every v.

E.2 Proof of Proposition E.2

For $v \in S^{d-1}$, write

$$a := \langle v, u\rangle, \qquad b := \langle v, c\rangle, \qquad \tau := \lambda/A.$$

Since $u \perp c$, we have

$$a^2 + b^2 \leq 1.$$

Because $Z_1 Z_2 = 0$ almost surely,

$$\mathcal{U}_\lambda(v; x) = \frac{1}{2}A^2\Big[\pi_1(a - \tau)_+^2 + \pi_2(b - \tau)_+^2\Big].$$

Assume first that $0 < \tau < 1$. Set

$$y_1 := (a - \tau)_+, \qquad y_2 := (b - \tau)_+.$$

If $y_1 = y_2 = 0$, the desired upper bound is immediate. Otherwise, on the active coordinates $I := \{i : y_i > 0\}$, we have $a_i = y_i + \tau$, where $a_1 = a$ and $a_2 = b$. Hence

$$\sum_{i \in I} (y_i + \tau)^2 \le a^2 + b^2 \le 1.$$

Therefore

$$\|y\|_2^2 + 2\tau \|y\|_1 + |I|\tau^2 \le 1.$$

Since $|I| \ge 1$ and $\|y\|_1 \ge \|y\|_2$,

$$\|y\|_2^2 + 2\tau \|y\|_2 + \tau^2 \le 1,$$

so

$$\|y\|_2 \le 1 - \tau.$$

Thus

$$(a - \tau)_+^2 + (b - \tau)_+^2 = y_1^2 + y_2^2 \le (1 - \tau)^2.$$

It follows that

$$\mathcal{U}_\lambda(v; x) \le \frac{1}{2} A^2 \max\{\pi_1, \pi_2\} \big((a - \tau)_+^2 + (b - \tau)_+^2 \big) \le \frac{1}{2} A^2 \max\{\pi_1, \pi_2\} (1 - \tau)^2.$$

Since $A^2(1 - \tau)^2 = (A - \lambda)^2$, this gives

$$\mathcal{U}_\lambda(v; x) \le \frac{1}{2} \max\{\pi_1, \pi_2\} (A - \lambda)^2.$$

The bound is attained by $v = u$ if $\pi_1 = \max\{\pi_1, \pi_2\}$, and by $v = c$ if $\pi_2 = \max\{\pi_1, \pi_2\}$. This proves the claim for $0 < \lambda < A$.

If $\lambda = 0$, then

$$\mathcal{U}_0(v; x) = \frac{1}{2} A^2 \left[\pi_1 a_+^2 + \pi_2 b_+^2 \right] \le \frac{1}{2} A^2 \max\{\pi_1, \pi_2\} (a^2 + b^2) \le \frac{1}{2} A^2 \max\{\pi_1, \pi_2\}.$$

The upper bound is attained by a highest-probability teacher direction. If $\pi_1 = \pi_2$, equality requires $a, b \ge 0$ and $a^2 + b^2 = 1$, which gives exactly the positive-quadrant unit directions in $\operatorname{span}\{u, c\}$.

If $\lambda \ge A$, then for every unit v,

$$A \langle v, u \rangle \le A \le \lambda, \qquad A \langle v, c \rangle \le A \le \lambda,$$

so both positive parts vanish and $\mathcal{U}_\lambda(v; x) = 0$.

Angle dependence of the merge mechanism. The orthogonal construction isolates the threshold effect. If the same independent co-occurrence model is used with nonorthogonal unit directions u, c and

$$\rho := \langle u, c \rangle = \cos\theta \in (-1, 1),$$

then the normalized bisector is

$$m := \frac{u + c}{\|u + c\|} = \frac{u + c}{\sqrt{2(1 + \rho)}}.$$

On singleton events,

$$\langle m, Au \rangle = \langle m, Ac \rangle = A \sqrt{\frac{1 + \rho}{2}},$$

whereas on the joint event,

$$\langle m, A(u + c) \rangle = A \sqrt{2(1 + \rho)}.$$

Thus the threshold-only merge regime in which singleton teacher projections of size A are suppressed while the joint bisector projection remains active is

$$A \le \lambda < A \sqrt{2(1 + \rho)}.$$

This interval is nonempty exactly when

$$\rho > -\frac{1}{2} \qquad \Longleftrightarrow \qquad \theta < \frac{2\pi}{3}.$$

Hence, for fixed amplitude and threshold, more aligned features create stronger merge pressure, while sufficiently opposed features do not produce this threshold-only merge regime.

The local derivative also depends explicitly on the angle. Let

$$h := \frac{c - \rho u}{\sqrt{1-\rho^2}},$$

the unit tangent at u pointing toward c, and define

$$v_\varphi := \cos\varphi\, u + \sin\varphi\, h.$$

For

$$x = As_1 u + As_2 c, \qquad s_1, s_2 \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p),$$

the tangent derivative is

$$\frac{d}{d\varphi}\mathcal{U}_\lambda(v_\varphi; x)\Big|_{\varphi=0} = A\sqrt{1-\rho^2}\Big[p(1-p)(A\rho-\lambda)_+ + p^2(A(1+\rho)-\lambda)_+\Big].$$

For $\rho = 0$, this reduces to

$$Ap^2(A-\lambda)_+,$$

the orthogonal formula in Proposition E.3.

Deferred independent co-occurrence statement. The proposition below shows that strict correlation is not necessary for merging to happen, since higher-order support events drive it.

Proposition E.3 (independent co-occurrence can favor a merge)**.** *Let $x = As_1u + As_2c$, where $u \perp c$, $A > 0$, and $s_1, s_2 \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$. Let $m = (u+c)/\sqrt{2}$. If $A \le \lambda < \sqrt{2}A$, then*

$$\mathcal{U}_\lambda(u; x) = 0 \qquad \textit{while} \qquad \mathcal{U}_\lambda(m; x) > 0$$

for every $p > 0$. If $\lambda < A$, then

$$\frac{d}{d\theta}\mathcal{U}_\lambda(\cos\theta\, u + \sin\theta\, c; x)\Big|_{\theta=0} = Ap^2(A-\lambda) > 0.$$

E.3 Proof of Proposition E.3

The support events are $\{1\}$, $\{2\}$, and $\{1,2\}$ with probabilities $p(1-p)$, $p(1-p)$, and p^2. For u, only the events containing atom 1 matter, so

$$\mathcal{U}_\lambda(u; x) = \frac{1}{2}p(A-\lambda)_+^2.$$

For $m = (u+c)/\sqrt{2}$, singleton events contribute overlap $A/\sqrt{2}$ and the pair event contributes overlap $\sqrt{2}A$, yielding the stated formula by direct substitution into Theorem 3.1. The derivative toward c follows from Proposition E.1: only the joint event has nonzero tangent projection, giving

$$\frac{d}{d\theta}\mathcal{U}_\lambda(\cos\theta\, u + \sin\theta\, c; x)\Big|_{\theta=0} = Ap^2(A-\lambda)_+.$$

Deferred absorption statements. For an event A, write $\mathcal{U}_{\lambda,A}(v; R) := \frac{1}{2}\mathbb{E}[(\langle v, R\rangle - \lambda)_+^2 \mathbf{1}_A]$.

Theorem E.4 (post-learning residual suppression on a rank-one event)**.** *Let A be an event on which $R = \xi u$, with $\|u\| = 1$ and $\xi \ge 0$. Let $v \in S^{d-1}$ be a previously selected feature with overlap $\rho := \langle u, v\rangle \in [0,1]$ and define $\mathcal{R}_v(R) := R - v(\langle v, R\rangle - \lambda)_+$. If $\rho\xi > \lambda$ almost surely on A, then*

$$\mathcal{U}_{\lambda,A}(u; \mathcal{R}_v(R)) = \frac{1}{2}(1-\rho)^2\mathbb{E}\big[((1+\rho)\xi - \lambda)^2 \mathbf{1}_A\big].$$

Proposition E.5 (absorption-like false negatives from independent support intersections)**.** *Let $u, w \in S^{d-1}$ satisfy $u \perp w$, and let $x = A(s_1 u + s_2 w)$, $s_1, s_2 \overset{\text{i.i.d.}}{\sim} \text{Bernoulli}(p)$, with $A > 0$. Let $m = (u+w)/\sqrt{2}$, and assume $\frac{1}{\sqrt{2}} < \frac{\lambda}{A} < 1$. Then there exists $p_\star(\lambda/A) < 1$ such that, for every $p > p_\star(\lambda/A)$, m is the unique global maximizer of $v \mapsto \mathcal{U}_\lambda(v; x)$. After subtracting the oracle code along m, $R^+ := x - m(\langle m, x\rangle - \lambda)_+$, we have*

$$\mathcal{U}_\lambda(u; x) = \frac{1}{2}p(A-\lambda)^2, \qquad \mathcal{U}_\lambda(u; R^+) = \frac{1}{2}p(1-p)(A-\lambda)^2.$$

Eventwise, the later oracle gate for u fires on $\{s_1 = 1, s_2 = 0\}$ and not on $\{s_1 = s_2 = 1\}$. Hence, conditional on the teacher event $\{s_1 = 1\}$, its recall is $1 - p$.

E.4 Proofs for Section 4.4

Proof of Theorem E.4. On A, we have $R = \xi u$, and therefore

$$\langle v, R\rangle = \xi\langle v, u\rangle = \rho\xi.$$

By assumption, $\rho\xi > \lambda$ almost surely on A, so

$$(\langle v, R\rangle - \lambda)_+ = \rho\xi - \lambda.$$

Hence, on A,

$$\mathcal{R}_v(R) = \xi u - v(\rho\xi - \lambda).$$

Taking the inner product with u gives

$$\begin{aligned}\langle u, \mathcal{R}_v(R)\rangle &= \xi - \rho(\rho\xi - \lambda)\\ &= (1-\rho^2)\xi + \rho\lambda.\end{aligned}$$

Thus

$$\begin{aligned}\langle u, \mathcal{R}_v(R)\rangle - \lambda &= (1-\rho^2)\xi - (1-\rho)\lambda\\ &= (1-\rho)((1+\rho)\xi - \lambda).\end{aligned}$$

Since $\rho\xi > \lambda$ on A, we have $(1+\rho)\xi - \lambda > 0$ on A. Taking the positive part and squaring yields

$$(\langle u, \mathcal{R}_v(R)\rangle - \lambda)_+^2 = (1-\rho)^2((1+\rho)\xi - \lambda)^2 \qquad \text{on } A.$$

Multiplying by $\mathbf{1}_A$, taking expectation, and dividing by 2 proves the claim. □

Proof of Proposition E.5. Write

$$t := \frac{\lambda}{A}, \qquad a := \langle v, u\rangle, \qquad b := \langle v, w\rangle.$$

Only the projection of v onto $\text{span}\{u, w\}$ affects the utility, so (a, b) ranges over the closed unit disk

$$D := \{(a, b) : a^2 + b^2 \le 1\}.$$

For $v \in S^{d-1}$, the utility equals

$$\begin{aligned}\mathcal{U}_\lambda(v; x) = \frac{1}{2}p(1-p)(Aa-\lambda)_+^2 + \frac{1}{2}p(1-p)(Ab-\lambda)_+^2\\ + \frac{1}{2}p^2(A(a+b)-\lambda)_+^2.\end{aligned}$$

Equivalently, up to the positive factor $A^2/2$, define

$$F_p(a, b) := p(1-p)(a-t)_+^2 + p(1-p)(b-t)_+^2 + p^2(a+b-t)_+^2.$$

At $p = 1$,

$$F_1(a, b) = (a+b-t)_+^2.$$

Over D, the linear functional $a + b$ is uniquely maximized at

$$(a, b) = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right),$$

corresponding to $v = m = (u + w)/\sqrt{2}$. Hence m is the unique maximizer of F_1.

Because D is compact and $F_p \to F_1$ uniformly as $p \to 1$, strict uniqueness of the maximizer of F_1 implies that, for every sufficiently small neighborhood N of $(1/\sqrt{2}, 1/\sqrt{2})$, all maximizers of F_p lie in N for p sufficiently close to 1. Since $t > 1/\sqrt{2}$, choose N small enough that

$$a < t, \qquad b < t \qquad \text{for all } (a, b) \in N.$$

On this neighborhood the singleton terms vanish, and

$$F_p(a, b) = p^2(a + b - t)_+^2.$$

This function is uniquely maximized over D at

$$(a, b) = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right).$$

Therefore there exists $p_\star(t) < 1$ such that, for every $p \in (p_\star(t), 1]$, m is the unique global maximizer of $v \mapsto \mathcal{U}_\lambda(v; x)$.

It remains to compute the effect of subtracting the oracle code along m. Before subtraction,

$$\langle u, x\rangle = As_1,$$

so, since $A > \lambda$,

$$\mathcal{U}_\lambda(u; x) = \frac{1}{2}\mathbb{P}(s_1 = 1)(A - \lambda)^2 = \frac{1}{2}p(A - \lambda)^2.$$

Now define

$$R^+ := x - m(\langle m, x\rangle - \lambda)_+.$$

Because $t > 1/\sqrt{2}$,

$$\langle m, Au\rangle = \langle m, Aw\rangle = \frac{A}{\sqrt{2}} < \lambda,$$

so the m-code is inactive on the singleton events. Because $t < 1$,

$$\langle m, A(u + w)\rangle = A\sqrt{2} > \lambda,$$

so the m-code is active on the coactivation event $C = \{s_1 = s_2 = 1\}$.

Thus on $\{s_1 = 1, s_2 = 0\}$, the residual remains

$$R^+ = Au,$$

and the later u-gate fires with value $A - \lambda$. On C,

$$\begin{aligned} R^+ &= A(u + w) - m(A\sqrt{2} - \lambda) \\ &= A(u + w) - \frac{u + w}{\sqrt{2}}(A\sqrt{2} - \lambda) \\ &= \frac{\lambda}{\sqrt{2}}(u + w). \end{aligned}$$

Therefore

$$\langle u, R^+\rangle = \frac{\lambda}{\sqrt{2}} < \lambda \qquad \text{on } C,$$

so the later u-gate is zero on the coactivation event. On all remaining events, either $s_1 = 0$ or the u-gate is also zero. Hence

$$\mathcal{U}_\lambda(u; R^+) = \frac{1}{2}\mathbb{P}(s_1 = 1, s_2 = 0)(A - \lambda)^2 = \frac{1}{2}p(1 - p)(A - \lambda)^2.$$

Finally, conditional on the teacher event $\{s_1 = 1\}$, the later u-gate fires exactly when $s_2 = 0$. Since s_1 and s_2 are independent,

$$\mathbb{P}(s_2 = 0 \mid s_1 = 1) = 1 - p.$$

This proves the recall statement. □

Deferred residual-path statement. The proposition below shows a mechanism analogous to seed instability in our framework.

Proposition E.6 (population residual-path non-identifiability)**.** *Let u_1, u_2, u_3 be orthonormal and define $q_{ij} := (u_i + u_j)/\sqrt{2}$. Suppose the residual takes values $\sqrt{2}Aq_{12}, \sqrt{2}Aq_{13}, \sqrt{2}Aq_{23}$ with probabilities $\pi_{12}, \pi_{13}, \pi_{23} > 0$. If $\beta := \lambda/\sqrt{2}A \in \left(\sqrt{3}/2, 1\right)$, then the global maximizers of $v \mapsto \mathcal{U}_\lambda(v; R)$ are exactly the pair directions q_{ij} whose probabilities π_{ij} are maximal. In particular, tied probabilities give multiple inequivalent non-teacher population maximizers.*

E.5 Proof of Proposition E.6

Let

$$\mathcal{P} := \{12, 13, 23\}.$$

For $a \in \mathcal{P}$, write q_a for the corresponding pair direction and π_a for its probability. For any unit vector v, set

$$t_a := \langle v, q_a \rangle.$$

The pair directions satisfy

$$\langle q_a, q_b \rangle = \frac{1}{2} \qquad (a \neq b).$$

We first show that at most one t_a can exceed β. If $t_a > \beta$ and $t_b > \beta$ for $a \neq b$, then

$$2\beta < t_a + t_b = \langle v, q_a + q_b \rangle \leq \|q_a + q_b\| = \sqrt{2 + 2\langle q_a, q_b \rangle} = \sqrt{3},$$

contradicting $\beta > \sqrt{3}/2$.

Therefore at most one event contributes positive utility for any v. The utility is

$$\mathcal{U}_\lambda(v; R) = \frac{1}{2} \sum_{a \in \mathcal{P}} \pi_a (\sqrt{2}A\, t_a - \lambda)_+^2 = A^2 \sum_{a \in \mathcal{P}} \pi_a (t_a - \beta)_+^2.$$

Since at most one term can be active and $t_a \leq 1$,

$$\mathcal{U}_\lambda(v; R) \leq A^2 (1 - \beta)^2 \max_{a \in \mathcal{P}} \pi_a.$$

This bound is attained by $v = q_a$ for every a with

$$\pi_a = \max_{b \in \mathcal{P}} \pi_b.$$

Indeed, for $v = q_a$,

$$t_a = 1, \qquad t_b = \frac{1}{2} < \beta \quad (b \neq a),$$

so only the a-event contributes. Hence the global maximizers are exactly the most probable pair directions.

The utility gap between two pair directions is explicit:

$$\mathcal{U}_\lambda(q_a; R) - \mathcal{U}_\lambda(q_b; R) = A^2 (1 - \beta)^2 (\pi_a - \pi_b).$$

Thus exact ties give exact population non-identifiability, and near ties give proportionally small population gaps.

Teacher directions have zero utility. For example,

$$\langle u_1, \sqrt{2}Aq_{12} \rangle = A, \qquad \langle u_1, \sqrt{2}Aq_{13} \rangle = A, \qquad \langle u_1, \sqrt{2}Aq_{23} \rangle = 0.$$

Since

$$\lambda > \frac{\sqrt{3}}{2} \sqrt{2}A = \sqrt{\frac{3}{2}} A > A,$$

all projections are below threshold. Thus

$$\mathcal{U}_\lambda(u_1; R) = 0,$$

and similarly for u_2, u_3.

Now suppose q_a is selected first and the oracle code

$$z_a(R) := (\langle q_a, R\rangle - \lambda)_+$$

is subtracted. On the a-event,

$$R = \sqrt{2}Aq_a, \qquad z_a = \sqrt{2}A - \lambda,$$

so the new residual is

$$R^+ = \lambda q_a.$$

For every unit direction v,

$$\langle v, R^+\rangle - \lambda = \lambda\langle v, q_a\rangle - \lambda \le 0.$$

Thus the selected event contributes no future above-threshold utility to any direction. On every other pair event $b \neq a$, the code is inactive because

$$\langle q_a, \sqrt{2}Aq_b\rangle = \frac{\sqrt{2}A}{2} < \lambda,$$

where the last inequality follows from $\beta > \sqrt{3}/2 > 1/2$. Hence all unselected pair events are unchanged. The later population utility field is therefore the original pair-event utility with the selected event removed.

At $\beta = \sqrt{3}/2$, the strict exclusion argument becomes non-strict: two pair directions can reach the threshold simultaneously, so the winner-takes-one proof no longer applies. Below this threshold, simultaneous active pair events are possible.

Deferred splitting statement. The proposition below shows a role of event partition in splitting.

Proposition E.7 (event-partition splitting certificate)**.** *Suppose an event family A decomposes into m disjoint rank-one subevents $A_1, \dots, A_m$, with $R_B = \xi_k u_k$ on A_k, where each $u_k \in S^{d-1}$ and each ξ_k is a nonnegative random scalar. If the best one-direction utility on A is strictly smaller than the utility achieved by the specialized directions $u_1, \dots, u_m$, then the oracle objective strictly favors allocating separate directions to the subevents.*

E.6 Proof of Proposition E.7

The exact sufficient condition used by Proposition E.7 is

$$\sup_{v\in S^{d-1}} \sum_{k=1}^{m} \mathbb{E}\left[(\xi_k \langle v, u_k\rangle - \lambda)_+^2 \mathbf{1}_{A_k}\right] < \sum_{k=1}^{m} \mathbb{E}\left[(\xi_k - \lambda)_+^2 \mathbf{1}_{A_k}\right],$$

where $A = \bigsqcup_{k=1}^{m} A_k$, $R_B = \xi_k u_k$ a.s. on A_k, $\xi_k \ge 0$, and $\|u_k\| = 1$. This is the formal version of event-partition splitting: one broad direction is worse than specialized children when no single v can match the eventwise rank-one gains.

On each subevent A_k, the residual is $\xi_k u_k$. A single direction v therefore achieves eventwise gain

$$\frac{1}{2}\sum_{k=1}^{m} \mathbb{E}[(\xi_k \langle v, u_k\rangle - \lambda)_+^2 \mathbf{1}_{A_k}].$$

By contrast, the child family $\{u_k\}$ achieves gain at least

$$\frac{1}{2}\sum_{k=1}^{m} \mathbb{E}[(\xi_k - \lambda)_+^2 \mathbf{1}_{A_k}]$$

by the feasible code $z_k = (\xi_k - \lambda)_+ \mathbf{1}_{A_k}$. If the former is strictly smaller than the latter uniformly over v, then splitting is oracle-preferred.

Deferred two-subevent splitting statement. Proposition below nuances the viewpoint above.

Proposition E.8 (two-subevent splitting regime)**.** *Let $A = A_1 \sqcup A_2$, with $\mathbb{P}(A_1) = \mathbb{P}(A_2) = q/2$. Suppose $R_B = A_0 u_k$ on A_k, where $A_0 > 0$, $u_1, u_2 \in S^{d-1}$, and $\angle(u_1, u_2) = \theta > 0$. If $0 \le \lambda < A_0$, then the two specialized directions u_1, u_2 achieve strictly larger eventwise oracle utility on A than any single shared direction. With $\tau = \lambda/A_0$, their guaranteed eventwise advantage is at least*

$$\frac{q}{2}(A_0 - \lambda)^2 - \sup_{v \in S^{d-1}} \frac{q}{4} \sum_{k=1}^{2} (A_0 \langle v, u_k \rangle - \lambda)_+^2 \ge \frac{q A_0^2}{4} \left[(1-\tau)^2 - (\cos(\theta/2) - \tau)_+^2 \right] > 0.$$

E.7 Proof of Proposition E.8

The specialized directions u_1, u_2 achieve at least the eventwise utility

$$\frac{1}{2} \sum_{k=1}^{2} \mathbb{E}[(A_0 - \lambda)^2 \mathbf{1}_{A_k}] = \frac{q}{2}(A_0 - \lambda)^2.$$

For a single shared direction v, write $t_k = \langle v, u_k \rangle$ and $\tau = \lambda/A_0$. The eventwise utility is

$$\frac{q A_0^2}{4} \left[(t_1 - \tau)_+^2 + (t_2 - \tau)_+^2 \right].$$

For every unit v, at least one of t_1, t_2 is at most $\cos(\theta/2)$. Otherwise

$$\langle v, u_1 + u_2 \rangle > 2\cos(\theta/2) = \|u_1 + u_2\|,$$

which contradicts Cauchy–Schwarz. Since also $t_k \le 1$, we have

$$(t_1 - \tau)_+^2 + (t_2 - \tau)_+^2 \le (1-\tau)^2 + (\cos(\theta/2) - \tau)_+^2.$$

Subtracting this upper bound from the specialized utility gives the displayed margin. The margin is strictly positive because $0 \le \tau < 1$ and $\cos(\theta/2) < 1$.

For the optional active-set characterization below, specialize to an atomic residual law

$$R = A u_i \quad \text{with probability } p_i, \qquad i = 1, \dots, M,$$

where $A > 0$, $u_i \in S^{d-1}$, $p_i > 0$, and $\sum_i p_i = 1$. Write $\beta := \lambda/A$. For $S \subseteq [M]$, define the active cell

$$\mathcal{C}_S := \{ v \in S^{d-1} : \ \langle v, u_i \rangle > \beta \text{ if and only if } i \in S \},$$

and the weighted active-set moments

$$G_S := \sum_{i \in S} p_i u_i u_i^\top, \qquad b_S := \sum_{i \in S} p_i u_i.$$

Theorem E.9 (exact active-set characterization of nonboundary strict local maxima)**.** *Assume $d \geq 2$ and the atomic residual law*

$$\mathbb{P}(R = Au_i) = p_i, \qquad i = 1, \dots, M,$$

where $A > 0$, $u_i \in S^{d-1}$, $p_i > 0$, and $\sum_{i=1}^{M} p_i = 1$. Set

$$\beta := \frac{\lambda}{A}.$$

For $S \subseteq [M]$, define the strict active cell

$$\mathcal{C}_S^\circ := \left\{ v \in S^{d-1} : \langle v, u_i \rangle > \beta \; \forall i \in S, \quad \langle v, u_i \rangle < \beta \; \forall i \notin S \right\}.$$

For $S \neq \varnothing$, define

$$G_S := \sum_{i \in S} p_i u_i u_i^\top, \qquad b_S := \sum_{i \in S} p_i u_i.$$

Let $S \neq \varnothing$ and $v \in \mathcal{C}_S^\circ$. Then, on a neighborhood of v in S^{d-1}, the active set is fixed and

$$\mathcal{U}_\lambda(v; R) = \frac{A^2}{2} \left(v^\top G_S v - 2\beta b_S^\top v + \beta^2 \sum_{i \in S} p_i \right).$$

Moreover, v is a spherical critical point if and only if there exists $\eta \in \mathbb{R}$ such that

$$G_S v - \beta b_S = \eta v.$$

At such a critical point, v is a strict local maximizer of $\mathcal{U}_\lambda$ on S^{d-1} if and only if

$$\xi^\top G_S \xi < \eta \, \|\xi\|^2 \qquad \forall \xi \in v^\perp \setminus \{0\}.$$

Equivalently,

$$\eta > \lambda_{\max} \left(P_{v^\perp} G_S P_{v^\perp} \big|_{v^\perp} \right).$$

Thus every nonboundary strict local maximizer is a threshold-consistent stable solution of the shifted active-set eigenproblem

$$G_S v - \beta b_S = \eta v.$$

E.8 Proof of Theorem E.9

Since $v \in \mathcal{C}_S^\circ$, the active inequalities are strict. Hence, on a neighborhood of v in S^{d-1}, the active set is fixed and

$$\mathcal{U}_\lambda(v; R) = \frac{A^2}{2} \sum_{i \in S} p_i (\langle v, u_i \rangle - \beta)^2.$$

Expanding gives

$$\mathcal{U}_\lambda(v; R) = \frac{A^2}{2} \left(v^\top G_S v - 2\beta b_S^\top v + \beta^2 \sum_{i \in S} p_i \right).$$

Thus, inside the active cell,

$$\nabla_v \mathcal{U}_\lambda(v; R) = A^2 (G_S v - \beta b_S).$$

The spherical critical-point condition is

$$P_{v^\perp} \nabla_v \mathcal{U}_\lambda(v; R) = 0,$$

which is equivalent to the existence of $\eta \in \mathbb{R}$ such that

$$G_S v - \beta b_S = \eta v.$$

At such a critical point, the Euclidean Hessian inside the active cell is $A^2 G_S$. Therefore the Riemannian Hessian on S^{d-1}, restricted to $v^\perp$, is

$$\mathrm{Hess}_{S^{d-1}} \mathcal{U}_\lambda(v) = A^2 \left(P_{v^\perp} G_S P_{v^\perp} - \eta I_{v^\perp} \right).$$

Hence the Riemannian Hessian is negative definite if and only if

$$\xi^\top G_S \xi < \eta \|\xi\|^2 \qquad \forall \xi \in v^\perp \setminus \{0\}.$$

Negative definiteness is sufficient for v to be a strict local maximizer.

It remains to show necessity. Suppose v is a strict local maximizer. Then the Riemannian Hessian is negative semidefinite, so

$$\xi^\top G_S \xi \leq \eta \|\xi\|^2 \qquad \forall \xi \in v^\perp.$$

Assume, for contradiction, that equality holds for some $\xi \in v^\perp$ with $\|\xi\| = 1$. Consider the spherical geodesic

$$v(t) = v \cos t + \xi \sin t.$$

For sufficiently small $|t|$, $v(t) \in \mathcal{C}_S^\circ$. Define the rescaled active-cell objective, up to an irrelevant additive constant, by

$$f(w) := w^\top G_S w - 2\beta b_S^\top w.$$

Using $G_S v - \beta b_S = \eta v$, $\xi \perp v$, and $\xi^\top G_S \xi = \eta$, a direct expansion gives

$$f(v(t)) - f(v) = \beta(b_S^\top v)(1 - \cos t)^2 - 2\beta(b_S^\top \xi) \sin t (1 - \cos t).$$

If $\beta b_S^\top \xi \neq 0$, the second term has leading order

$$-\beta(b_S^\top \xi) t^3,$$

which changes sign with t. Hence v cannot be a local maximizer.

If $\beta b_S^\top \xi = 0$, then either $\beta = 0$ or $b_S^\top \xi = 0$. If $\beta = 0$, the expression above is identically zero along this geodesic, so v cannot be a strict local maximizer. If $\beta > 0$, then $b_S^\top \xi = 0$, and

$$f(v(t)) - f(v) = \beta(b_S^\top v)(1 - \cos t)^2.$$

Since $v \in \mathcal{C}_S^\circ$,

$$b_S^\top v = \sum_{i \in S} p_i \langle v, u_i \rangle > \beta \sum_{i \in S} p_i > 0.$$

Therefore $f(v(t)) - f(v) > 0$ for all sufficiently small $t \neq 0$, again contradicting local maximality.

Thus equality cannot occur in any nonzero tangent direction. The Hessian is negative definite, proving the strict inequality and completing the characterization.

E.9 Singleton-cell and pairwise-exclusion lemmas

If $\max_{j \neq i} \langle u_i, u_j \rangle < \beta$, then at u_i one has $\langle u_i, u_i \rangle = 1 > \beta$ and $\langle u_i, u_j \rangle < \beta$ for every $j \neq i$. These inequalities persist in a neighborhood, so u_i lies in the interior of the singleton cell.

For pairwise exclusion, suppose two atoms u_i, u_j were simultaneously active at some v, so that $\langle v, u_i \rangle > \beta$ and $\langle v, u_j \rangle > \beta$. Then

$$\langle v, u_i + u_j \rangle > 2\beta.$$

But

$$\langle v, u_i + u_j \rangle \leq \|u_i + u_j\| = \sqrt{2 + 2\langle u_i, u_j \rangle} \leq 2\nu,$$

contradicting $\beta > \nu$, where

$$\nu := \max_{i \neq j} \sqrt{\frac{1 + \langle u_i, u_j \rangle}{2}}.$$

E.10 Split rank as an architectural refinement

The oracle mechanisms above interact with an architectural mechanism specific to amortized SAEs: the oracle gate associated with a decoder direction may fail to lie in the single affine-ReLU class. This motivates a simple complexity notion.

Definition E.10 (split rank)**.** For a nonnegative measurable target $g : \mathbb{R}^d \to \mathbb{R}_+$, define

$$\operatorname{srank}(g) := \min\Big\{ m \in \mathbb{N} : g(\nu) = \sum_{k=1}^{m} (e_k^\top \nu + b_k)_+ \text{ a.s.} \Big\},$$

when such an m exists, and set $\operatorname{srank}(g) = \infty$ otherwise.

If the oracle gate of a decoder direction has split rank greater than one, then a standard one-ReLU feature must either incur amortization error or factor the target across several slots. This mechanism is distinct from oracle event-partition splitting, which concerns which support partition is reward-optimal even with unrestricted activations.

The split-rank definition is therefore an architectural refinement, not another primitive oracle mechanism. It explains why the same oracle direction can appear as several learned slots when a single affine-ReLU gate cannot express the event rule.

F Positive and robust recovery refinements

The counterexamples are complemented by sufficient conditions: singleton mass can dominate contamination, directional moments can remain stable under small support errors, and low coherence can control off-target signal. These results delimit the regimes in which the usual feature-recovery intuition is scientifically justified.

Singleton dominance can survive margins, perturbations, low coherence, and random or learned geometry assumptions.

F.1 Proof of Theorem 4.1

The quantities used in the theorem are

$$\Delta^B_{j,\lambda}(\varepsilon) := \frac{1}{2}\mathbb{E}\Big[\big((\alpha_j \|\tilde{w}_j\| - \lambda)_+^2 - (\alpha_j \|\tilde{w}_j\| \cos\varepsilon - \lambda)_+^2\big)\mathbf{1}_{E_{\{j\}}(B)}\Big],$$

and

$$C_j^B := \frac{1}{2}\mathbb{E}\big[\|R_B\|^2 \mathbf{1}_{E_{\{j\}}(B)^c}\big].$$

On $E_{\{j\}}(B)$,

$$R_B = \alpha_j \tilde{w}_j,$$

so for every v with $\angle(v, u_j) \geq \varepsilon$,

$$\langle u_j, R_B\rangle = \alpha_j \|\tilde{w}_j\|, \qquad \langle v, R_B\rangle \leq \alpha_j \|\tilde{w}_j\| \cos\varepsilon.$$

Therefore the utility difference restricted to $E_{\{j\}}(B)$ is at least $\Delta^B_{j,\lambda}(\varepsilon)$. On the complement $E_{\{j\}}(B)^c$, competitor utility is bounded by

$$\frac{1}{2}\mathbb{E}[\|R_B\|^2 \mathbf{1}_{E_{\{j\}}(B)^c}] = C_j^B,$$

while the utility of u_j is nonnegative. Subtracting the two bounds yields the claim.

Sharper threshold-aware competitor load. The proof above uses the threshold-blind off-singleton load C_j^B. A sharper but less explicit quantity is

$$C^B_{j,\lambda}(\varepsilon) := \sup_{\angle(v,u_j)\geq\varepsilon} \frac{1}{2}\mathbb{E}\Big[(\langle v, R_B\rangle - \lambda)_+^2 \mathbf{1}_{E_{\{j\}}(B)^c}\Big].$$

The same argument proves localization under the weaker condition

$$\Delta^B_{j,\lambda}(\varepsilon) > C^B_{j,\lambda}(\varepsilon).$$

Indeed, for every v with $\angle(v, u_j) \geq \varepsilon$, the singleton-event contribution to

$$\mathcal{U}_\lambda(u_j; R_B) - \mathcal{U}_\lambda(v; R_B)$$

is at least $\Delta^B_{j,\lambda}(\varepsilon)$, while the off-singleton contribution of v is at most $C^B_{j,\lambda}(\varepsilon)$, and the off-singleton contribution of u_j is nonnegative. Hence

$$\mathcal{U}_\lambda(u_j; R_B) - \mathcal{U}_\lambda(v; R_B) \geq \Delta^B_{j,\lambda}(\varepsilon) - C^B_{j,\lambda}(\varepsilon) > 0.$$

Thus no maximizer can lie outside the ε-ball around u_j.

The simpler load C^B_j used in the main theorem is a conservative upper bound on $C^B_{j,\lambda}(\varepsilon)$. For every unit v,

$$(\langle v, R_B\rangle - \lambda)^2_+ \leq (\langle v, R_B\rangle)^2_+ \leq \langle v, R_B\rangle^2 \leq \|R_B\|^2,$$

and therefore

$$C^B_{j,\lambda}(\varepsilon) \leq \frac{1}{2}\mathbb{E}\big[\|R_B\|^2 \mathbf{1}_{E_{\{j\}}(B)^c}\big] = C^B_j.$$

Thus the main-text condition

$$\Delta^B_{j,\lambda}(\varepsilon) > C^B_j$$

is a simple sufficient condition for the sharper threshold-aware condition.

Corollary F.1 (uniform singleton-margin criterion)**.** *Suppose that on $E_{\{j\}}(B)$,*

$$\alpha_j \|\tilde{w}_j\| \geq a_j > \lambda \qquad \textit{a.s.},$$

and write

$$q^B_j := \mathbb{P}(E_{\{j\}}(B)).$$

Then

$$\Delta^B_{j,\lambda}(\varepsilon) \geq \frac{q^B_j}{2}\Big[(a_j - \lambda)^2 - (a_j \cos\varepsilon - \lambda)^2_+\Big].$$

Hence the sufficient condition

$$q^B_j\Big[(a_j - \lambda)^2 - (a_j \cos\varepsilon - \lambda)^2_+\Big] > \mathbb{E}\big[\|R_B\|^2 \mathbf{1}_{E_{\{j\}}(B)^c}\big]$$

localizes all global maximizers near u_j.

F.2 Proof of Corollary F.1

On $E_{\{j\}}(B)$, the quantity $\alpha_j \|\tilde{w}_j\|$ is bounded below by $a_j > \lambda$ almost surely. Therefore the integrand defining $\Delta^B_{j,\lambda}(\varepsilon)$ is bounded below by

$$(a_j - \lambda)^2 - (a_j \cos\varepsilon - \lambda)^2_+.$$

Multiplying by the event probability q^B_j and the prefactor $1/2$ gives the stated lower bound. The localization condition then follows by combining this lower bound with Theorem 4.1.

Proposition F.2 (robust witness-concentration moment bound)**.** *Let A be a measurable event and suppose that for some unit u and activation $a \geq 0$ with $\mathbb{E}[\|R_B\| a] < \infty$,*

$$\mu(a; B) = \mathbb{E}[R_B a] = \beta_A(a)u + \eta_A(a), \qquad \eta_A(a) \in u^\perp, \qquad \beta_A(a) > 0.$$

Then

$$\tan\angle(\mu(a; B), u) = \frac{\|\eta_A(a)\|}{\beta_A(a)}.$$

If $\beta_A(a) > 2\|\eta_A(a)\|$, then

$$\left\| \frac{\mu(a; B)}{\|\mu(a; B)\|} - u \right\| \leq \frac{2\|\eta_A(a)\|}{\beta_A(a)}.$$

F.3 Proof of Proposition F.2

Write

$$\mu(a; B) = \beta_A(a)u + \eta_A(a).$$

Since $\eta_A(a) \in u^\perp$,

$$\|P_{u^\perp}\mu(a; B)\| = \|\eta_A(a)\|, \qquad \langle u, \mu(a; B)\rangle = \beta_A(a).$$

Taking the ratio of the transverse norm and the longitudinal component gives the tangent identity. The normalized-vector bound follows from the assumption $\beta_A(a) > 2\|\eta_A(a)\|$ and a standard perturbation estimate.

F.4 Support is non-robust while direction is robust

Proposition F.3 (support is non-robust while direction is robust)**.** *Let A be a rank-one event for the ray $\mathbb{R}_+u$, and let $a \geq 0$, with $\mathbb{E}[\|R_B\|\, a] < \infty$, satisfy*

$$a = 0 \quad \text{a.s. on } A^c, \qquad \mu(a; B) \neq 0.$$

Let $h \geq 0$ be any measurable function supported on A^c and nonzero on a positive-probability subset of A^c, with $\mathbb{E}[\|R_B\|\, h] < \infty$. Define

$$a_\varepsilon := a + \varepsilon h, \qquad \varepsilon > 0.$$

Then for every $\varepsilon > 0$, the activation a_ε no longer fires exactly on A. However,

$$\mu(a_\varepsilon; B) = \mu(a; B) + \varepsilon\mathbb{E}[R_B h],$$

and if

$$\varepsilon\mathbb{E}[\|R_B\|\, h] \leq \frac{1}{2}\,\|\mu(a; B)\|\,,$$

then

$$\left\| \frac{\mu(a_\varepsilon; B)}{\|\mu(a_\varepsilon; B)\|} - \frac{\mu(a; B)}{\|\mu(a; B)\|} \right\| \leq \frac{4\varepsilon\mathbb{E}[\|R_B\|\, h]}{\|\mu(a; B)\|}.$$

Proof. Exact support recovery fails immediately because h is supported on A^c and is nonzero there. The moment identity is linear in the activation. The perturbation satisfies

$$\left\|\varepsilon\mathbb{E}[R_B h]\right\| \leq \varepsilon\mathbb{E}[\|R_B\|\, h].$$

Under the assumed bound, this perturbation has norm at most $\|\mu(a; B)\|/2$, so the standard normalized-vector perturbation inequality yields the stated estimate. □

F.5 Low coherence controls off-target contamination

Proposition F.4 (low coherence controls off-target contamination)**.** *Let $w_1^\star, \dots, w_F^\star \in S^{d-1}$ be teacher directions with coherence*

$$\gamma := \max_{i \neq j} \left| \langle w_i^\star, w_j^\star \rangle \right|.$$

Assume

$$x = \sum_{i=1}^{F} f_i w_i^\star, \qquad f_i \geq 0, \qquad |\{i : f_i > 0\}| \leq s, \qquad 0 \leq f_i \leq a_{\max}.$$

Then for every j,

$$\left| \sum_{i \neq j} \langle w_j^\star, w_i^\star \rangle f_i \right| \leq \gamma(s-1)a_{\max},$$

and therefore

$$\langle w_j^\star, x \rangle \geq f_j - \gamma(s-1)a_{\max}.$$

Proof. For every j,

$$\left| \sum_{i \neq j} \langle w_j^\star, w_i^\star \rangle f_i \right| \leq \sum_{i \neq j} \left| \langle w_j^\star, w_i^\star \rangle \right| f_i \leq \gamma \sum_{i \neq j} f_i \leq \gamma(s-1)a_{\max}.$$

Since

$$\langle w_j^\star, x \rangle = f_j + \sum_{i \neq j} \langle w_j^\star, w_i^\star \rangle f_i,$$

the lower bound follows. □

Proposition F.5 (random learned geometry is low-coherence with high probability)**.** *Let $w_1^\star, \dots, w_F^\star$ be i.i.d. uniform on S^{d-1}. Then there exists a universal constant $c > 0$ such that, with probability at least $1 - \delta$,*

$$\max_{i \neq j} \left| \langle w_i^\star, w_j^\star \rangle \right| \leq c\sqrt{\frac{\log(F/\delta)}{d}}.$$

Proof. For a fixed pair (i, j), rotational invariance and spherical concentration imply

$$\mathbb{P}\left(\left| \langle w_i^\star, w_j^\star \rangle \right| > t \right) \leq 2e^{-c_0 d t^2}$$

for a universal constant $c_0 > 0$ and all $0 \leq t \leq 1$. Choose

$$t = C\sqrt{\frac{\log(F/\delta)}{d}}$$

with C large enough that $2\binom{F}{2}e^{-c_0 d t^2} \leq \delta$. A union bound over the $\binom{F}{2}$ pairs yields

$$\max_{i \neq j} \left| \langle w_i^\star, w_j^\star \rangle \right| \leq C\sqrt{\frac{\log(F/\delta)}{d}}$$

with probability at least $1 - \delta$. Absorbing C into the universal constant c gives the claim. □

This is a standard concentration fact included only to make the low-coherence positive regime explicit.

G Learned teacher geometry

If the teacher dictionary is itself learned by an upstream sparse-superposition objective, the oracle picture gains one useful nuance: upstream geometry can make support-to-direction alignment easier in special symmetric regimes, but exact reconstruction of rich support families forces strong Gram-structure constraints.

Symmetry and simplex-like learned geometry can cancel off-target moments, so support-faithful gates can recover directions even outside literal singleton-only settings.

Let

$$f \in \mathbb{R}_+^F$$

be drawn from a sparse generator, and consider the tied feature autoencoder

$$\hat{f} = \sigma(Gf + b), \qquad \sigma(t) = t_+, \qquad G = WW^\top \succeq 0, \qquad \operatorname{rank}(G) \le d,$$

trained with

$$\mathcal{J}(G, b) = \mathbb{E}\big[\|f - \sigma(Gf + b)\|_2^2\big].$$

The learned teacher directions are the rows of W, denoted

$$w_1^\star, \dots, w_F^\star \in \mathbb{R}^d,$$

so that

$$G_{ij} = \langle w_i^\star, w_j^\star \rangle.$$

Downstream, the hidden state is

$$x = \sum_{i=1}^{F} f_i w_i^\star.$$

For a support family $T \subseteq [F]$, write

$$E_T := \{\operatorname{supp}(f) = T\},$$

and let $\Omega_T \subseteq (0, \infty)^T$ be the topological support of f_T conditional on E_T. We say that E_T is full-dimensional if Ω_T contains a nonempty open subset of $\mathbb{R}^T$.

Theorem G.1 (active-block rigidity of the learned teacher)**.** *Assume E_T is full-dimensional and the feature autoencoder reconstructs it exactly:*

$$\sigma(Gf + b) = f \qquad \text{a.s. on } E_T.$$

Then

$$G_{TT} = I, \qquad b_T = 0.$$

Moreover, for every $j \notin T$,

$$G_{jT} s + b_j \le 0 \qquad \forall s \in \Omega_T.$$

Consequently, if every pair support $E_{\{i,j\}}$ is full-dimensional and reconstructed exactly, then

$$G_{ij} = 0 \qquad \forall i \ne j.$$

If this holds for all pairs and $\operatorname{rank}(G) \le d < F$, exact reconstruction is impossible.

Proof. On the support event E_T, one has $f_{T^c} = 0$. Exact reconstruction gives

$$\sigma(G_{TT} f_T + b_T) = f_T, \qquad \sigma(G_{T^cT} f_T + b_{T^c}) = 0$$

almost surely conditional on E_T.

Because $f_T \in (0, \infty)^T$ on E_T, the first identity implies

$$G_{TT} f_T + b_T = f_T$$

almost surely on E_T. The map

$$s \mapsto (G_{TT} - I)s + b_T$$

is continuous. Since the conditional topological support Ω_T contains a nonempty open set, and the continuous map vanishes almost surely, it vanishes on Ω_T, hence on a nonempty open set. An affine map that vanishes on a nonempty open set is identically zero. Therefore

$$G_{TT} = I, \qquad b_T = 0.$$

For $j \notin T$, exact inactivity gives

$$G_{jT} f_T + b_j \le 0$$

almost surely on E_T. By the same continuity and support argument,

$$G_{jT} s + b_j \le 0 \qquad \forall s \in \Omega_T.$$

If every pair event $E_{\{i,j\}}$ is full-dimensional and reconstructed exactly, applying the first part with $T = \{i, j\}$ gives

$$G_{\{i,j\},\{i,j\}} = I_2,$$

so

$$G_{ij} = G_{ji} = 0.$$

If this holds for all pairs, then $G = I_F$. But $\mathrm{rank}(I_F) = F$, contradicting $\mathrm{rank}(G) \le d < F$. Hence exact reconstruction is impossible in the undercomplete regime. □

Lemma G.2 (singleton sign constraint)**.** *Assume each singleton event $E_{\{i\}}$ has at least two positive amplitudes in its conditional support and is reconstructed exactly. Then*

$$G_{ii} = 1, \qquad b_i = 0, \qquad G_{ji} \le 0 \quad (j \ne i).$$

Proof of Lemma G.2. Assume each singleton event $E_{\{i\}}$ has at least two positive amplitudes in its conditional support and is reconstructed exactly. On $E_{\{i\}}$, exact reconstruction of the active coordinate gives

$$(G_{ii} f_i + b_i)_+ = f_i.$$

Since $f_i > 0$, this implies

$$G_{ii} f_i + b_i = f_i$$

on at least two distinct positive values of f_i. Hence

$$G_{ii} = 1, \qquad b_i = 0.$$

Now fix $j \ne i$. Exact inactivity of coordinate j on $E_{\{i\}}$ gives

$$G_{ji} f_i + b_j \le 0.$$

Since the previous argument applied to singleton j gives $b_j = 0$, and since $f_i > 0$, we obtain

$$G_{ji} \le 0.$$

□

Theorem G.3 (permutation symmetry induces equiangular Gram blocks)**.** *Let $T \subseteq [F]$, $|T| = k$. Assume the law of f is invariant under all permutations of the coordinates in T, and let (G, b) be a teacher solution with the same symmetry:*

$$PGP^\top = G, \qquad Pb = b \qquad \forall P \in \mathfrak{S}_T.$$

Then

$$G_{TT} = \alpha I + \beta \mathbf{1}\mathbf{1}^\top, \qquad b_T = \tau \mathbf{1}$$

for some scalars α, β, τ.
If, in addition, all singleton events $E_{\{i\}}$, $i \in T$, are nondegenerate and reconstructed exactly, then

$$b_T = 0, \qquad G_{TT} = (1-\rho)I + \rho \mathbf{1}\mathbf{1}^\top$$

for some $\rho \leq 0$. If also $\operatorname{rank}(G_{TT}) < k$, then

$$\rho = -\frac{1}{k-1},$$

so G_{TT} is the Gram matrix of a regular simplex.

Proof. A matrix commuting with every permutation matrix on T lies in the commutant of the permutation representation of $\mathfrak{S}_T$. This commutant consists exactly of matrices of the form

$$\alpha I + \beta \mathbf{1}\mathbf{1}^\top.$$

Similarly, the only vectors fixed by all permutations of T are constant vectors, so

$$b_T = \tau \mathbf{1}.$$

If all singleton events $E_{\{i\}}$, $i \in T$, are nondegenerate and reconstructed exactly, Lemma G.2 gives

$$G_{ii} = 1, \qquad b_i = 0, \qquad G_{ij} \leq 0 \quad (i \neq j).$$

By symmetry, all off-diagonal entries in G_{TT} are equal. Denote their common value by $\rho \leq 0$. Then

$$G_{TT} = (1-\rho)I + \rho \mathbf{1}\mathbf{1}^\top, \qquad b_T = 0.$$

The eigenvalues of this matrix are

$$1 - \rho \quad \text{with multiplicity } k-1,$$

and

$$1 + (k-1)\rho \quad \text{with multiplicity } 1.$$

Since $G_{TT} \succeq 0$, these eigenvalues are nonnegative. Because $\rho \leq 0$, the eigenvalue $1-\rho$ is strictly positive. Therefore $\operatorname{rank}(G_{TT}) < k$ can only occur when

$$1 + (k-1)\rho = 0,$$

i.e.

$$\rho = -\frac{1}{k-1}.$$

This is the regular-simplex Gram matrix. □

Proposition G.4 (effective threshold in an equiangular learned block)**.** *Let $T \subseteq [F]$ be a cluster of size k satisfying*

$$\|w_i^\star\| = 1, \qquad \langle w_i^\star, w_j^\star \rangle = \rho \quad (i \neq j,\ i, j \in T),$$

with

$$-\frac{1}{k-1} < \rho < 1.$$

Suppose the oracle nonnegative Lasso with fixed decoder $W^\star$ has active support exactly T, and there is no active off-cluster contribution into T, meaning $(W_T^\star)^\top x = G_{TT}^\star f_T$ on the event under consideration. Then, for every $i \in T$,

$$z_i^\star = f_i - \frac{\lambda}{1 + (k-1)\rho}.$$

Thus the effective threshold inside the block is

$$\lambda_{\text{eff}}(k, \rho) = \frac{\lambda}{1 + (k-1)\rho}.$$

Proof. On active support T, the KKT equations for the nonnegative Lasso are

$$G_{TT}^\star z_T^\star = (W_T^\star)^\top x - \lambda \mathbf{1}.$$

Under the assumption that there is no active off-cluster contribution into T,

$$(W_T^\star)^\top x = G_{TT}^\star f_T.$$

Therefore

$$z_T^\star = f_T - \lambda (G_{TT}^\star)^{-1} \mathbf{1}.$$

For an equiangular block,

$$G_{TT}^\star = (1-\rho) I + \rho \mathbf{1}\mathbf{1}^\top.$$

The vector $\mathbf{1}$ is an eigenvector with eigenvalue

$$1 + (k-1)\rho.$$

Thus

$$(G_{TT}^\star)^{-1} \mathbf{1} = \frac{1}{1 + (k-1)\rho} \mathbf{1},$$

and hence

$$z_i^\star = f_i - \frac{\lambda}{1 + (k-1)\rho} \qquad \forall i \in T.$$

□

Proposition G.5 (simplex cancellation restores direction recovery)**.** *Assume*

$$\{w_1^\star, \dots, w_m^\star\}$$

forms a regular simplex, so

$$\sum_{i=1}^m w_i^\star = 0.$$

Fix $j \in [m]$, and let $a(x) \geq 0$ be a learned activation satisfying

$$\mathbf{1}_{\{a>0\}} = \mathbf{1}_{\{f_j>0\}} \qquad \textit{a.s.}$$

Assume the off-target mixed moments are exchangeable:

$$\mathbb{E}[f_i a] = \beta \qquad \forall i \neq j, \qquad \mathbb{E}[f_j a] = \gamma.$$

Then

$$\mu(a) := \mathbb{E}[xa] = (\gamma - \beta) w_j^\star.$$

In particular, if $\gamma \neq \beta$, then the decoder moment is parallel to $w_j^\star$; if $\gamma > \beta$, it lies on the positive ray $\mathbb{R}_+ w_j^\star$.

Proof. By definition,

$$\mu(a) = \mathbb{E}[xa] = \sum_{i=1}^{m} \mathbb{E}[f_i a] w_i^\star.$$

Using the exchangeability assumption,

$$\mu(a) = \gamma w_j^\star + \beta \sum_{i \neq j} w_i^\star.$$

Since the simplex relation gives

$$\sum_{i \neq j} w_i^\star = -w_j^\star,$$

we obtain

$$\mu(a) = (\gamma - \beta) w_j^\star.$$

□

Remark G.6 (symmetry and multiplicity)**.** If a finite orthogonal group $\Gamma \subset O(d)$ preserves both the learned teacher dictionary and the distribution of f, then the induced residual law is Γ-invariant. Therefore

$$\mathcal{U}_\lambda(gv; R) = \mathcal{U}_\lambda(v; R) \qquad \forall g \in \Gamma.$$

Every local or global maximizer of the oracle utility field then comes with its symmetry orbit. Symmetric learned geometries can restore support-to-direction alignment while simultaneously increasing multiplicity.

H Amortized bridge proofs and stress tests

The bridge between SAE gradient dynamics, oracle one-slot utility, and one-ReLU realizability has two sides. First, decoder updates follow a gated residual moment and one encoder row tries to realize the oracle gate. Second, amortization gaps and pair-only worlds can change the ranking of directions.

The bridge from oracle sparse coding to SAE slots turns on whether the encoder can realize the oracle gate; the stress tests show when amortization or pair-only structure changes the recovery picture.

Deferred decoder-gradient statement. Proposition below is a characterization of moment-level condition of the decoder gate.

Proposition H.1 (decoder projected-gradient moment)**.** *In the normalized untied SAE model, with activations fixed as functions of the encoder input and with $\|d_r\| = 1$, the sphere-projected decoder gradient flow for slot r is proportional to*

$$P_{d_r^\perp} \mathbb{E}[R_{-r} z_r].$$

H.1 Proof of Theorem 5.1 and Proposition H.1

Write the sample loss as

$$\ell(\nu; D, E, b_{\text{enc}}) := \frac{1}{2} \|\nu - Dz\|_2^2 + \lambda \mathbf{1}^\top z, \qquad z = \sigma(E\nu + b_{\text{enc}}),$$

with residual $r = \nu - Dz$. Since $\partial r / \partial d_r = -z_r$ and z_r does not depend on d_r in the untied model,

$$\nabla_{d_r} \ell = -r z_r.$$

Taking expectations gives

$$\nabla_{d_r} \mathcal{L}_\lambda = -\mathbb{E}[r z_r].$$

Next differentiate through the gate $z_r = (e_r^\top \nu + b_r)_+$. Under the no-boundary-mass assumption $\mathbb{P}(e_r^\top \nu + b_r = 0) = 0$, we may differentiate almost surely with

$$\frac{\partial z_r}{\partial e_r} = m_r \nu, \qquad \frac{\partial z_r}{\partial b_r} = m_r, \qquad m_r = \mathbf{1}_{\{e_r^\top \nu + b_r > 0\}}.$$

Because only z_r depends on (e_r, b_r),

$$\frac{\partial \ell}{\partial z_r} = -d_r^\top r + \lambda.$$

Therefore

$$\nabla_{e_r} \ell = (\lambda - d_r^\top r) m_r \, \nu, \qquad \partial_{b_r} \ell = (\lambda - d_r^\top r) m_r.$$

Taking expectations gives the displayed formulas.

Since

$$R_{-r} = \nu - \sum_{s \neq r} d_s z_s = r + d_r z_r,$$

and $\|d_r\| = 1$,

$$d_r^\top r = d_r^\top R_{-r} - z_r.$$

Hence if

$$y_r(\nu) := d_r^\top R_{-r}(\nu) - \lambda,$$

then

$$\lambda - d_r^\top r = z_r - y_r,$$

which yields the equivalent encoder-gradient formulas.

For the block decomposition, expand

$$\frac{1}{2} \|R_{-r} - d_r z_r\|_2^2 + \lambda z_r = \frac{1}{2} \|R_{-r}\|_2^2 - z_r d_r^\top R_{-r} + \frac{1}{2} z_r^2 + \lambda z_r.$$

Using $y_r = d_r^\top R_{-r} - \lambda$, this becomes

$$\frac{1}{2} \|R_{-r}\|_2^2 - z_r y_r + \frac{1}{2} z_r^2 = \frac{1}{2} \|R_{-r}\|_2^2 - \frac{1}{2} y_r^2 + \frac{1}{2} (z_r - y_r)^2.$$

Taking expectations shows that, for fixed R_{-r}, blockwise optimization over (e_r, b_r) is exactly

$$\min_{e_r, b_r} \ \frac{1}{2} \mathbb{E}[(z_r - y_r)^2].$$

Finally, the decoder is constrained to the sphere $\|d_r\| = 1$. The ambient negative gradient direction is $\mathbb{E}[r z_r]$, so the projected gradient flow is

$$\dot{d}_r = P_{d_r^\perp} \mathbb{E}[r z_r].$$

Since $P_{d_r^\perp}(d_r z_r^2) = 0$ pointwise and $R_{-r} = r + d_r z_r$, this is equivalently

$$\dot{d}_r = P_{d_r^\perp} \mathbb{E}[R_{-r} z_r].$$

Theorem H.2 (Oracle utility as the sample-wise activation envelope)**.** *Let $\mathcal{G} := L_+^2(\Omega)$ denote arbitrary nonnegative square-integrable gates. Assume $R \in L^2$, and let $d \in S^{d-1}$. Then $\mathcal{U}_\lambda(d; R) = \sup_{z \in \mathcal{G}} \Gamma_\lambda(d, z; R)$, and the unique maximizer in $\mathcal{G}$ is $g_d^\star = (d^\top R - \lambda)_+$. Moreover, for every $z \in \mathcal{G}$,*

$$\mathcal{U}_\lambda(d; R) - \Gamma_\lambda(d, z; R) = \mathbb{E}\left[\frac{1}{2} (z - g_d^\star)^2 + z(\lambda - d^\top R)_+ \right].$$

Consequently, $\mathcal{A}_\lambda(d; R, \nu) \leq \mathcal{U}_\lambda(d; R)$.

H.2 Proof of Theorem H.2

For fixed d and sample value of (R, ν), optimize pointwise over nonnegative scalar z:

$$q := d^\top R - \lambda, \qquad \sup_{z \geq 0} \left\{ zq - \frac{1}{2} z^2 \right\} = \frac{1}{2} q_+^2, \qquad z^\star = q_+.$$

Thus the pointwise maximizer is

$$g_d^\star = (d^\top R - \lambda)_+.$$

Moreover, for any $z \geq 0$,

$$\frac{1}{2} q_+^2 - \left(zq - \frac{1}{2} z^2 \right) = \frac{1}{2} (z - q_+)^2 + z(-q)_+.$$

Substituting $q = d^\top R - \lambda$ and taking expectations gives the envelope formula and the exact gap identity.

H.3 One-ReLU realizability criterion and closure remark

If $g : \mathbb{R}^d \to \mathbb{R}_+$ is continuous and piecewise affine, exact one-ReLU realizability means

$$g(\nu) = (e^\top \nu + b)_+.$$

Then $\{g > 0\}$ must be a halfspace and g must agree there with one affine map. Conversely, any nonnegative function with those two properties is a single affine-ReLU.

Strictly speaking, equality $\mathcal{A}_\lambda(d; R, \nu) = \mathcal{U}_\lambda(d; R)$ means that the oracle gate lies in the L^2-closure of the one-ReLU class. Under optimizer attainment, this reduces to exact one-ReLU realization as stated in Theorem 5.2.

H.4 Proof of Theorem 5.2

Proof. The identity in the theorem implies that any sequence $z_n \in \mathcal{A}_\nu$ with $\Gamma_\lambda(d, z_n; R) \to \mathcal{U}_\lambda(d; R)$ must satisfy $\|z_n - g_d^\star\|_{L^2} \to 0$. Conversely, if $z_n \in \mathcal{A}_\nu$ and $z_n \to g_d^\star$ in L^2, then the same identity and Cauchy–Schwarz imply $\Gamma_\lambda(d, z_n; R) \to \mathcal{U}_\lambda(d; R)$, since $g_d^\star = 0$ on $\{d^\top R - \lambda < 0\}$ and $(\lambda - d^\top R)_+ \in L^2$. The attained statement follows because equality in the nonnegative gap identity forces $z = g_d^\star$ almost surely. □

Remark H.3 (Different envelope for encoder-measurable unrestricted gates)**.** If instead one optimizes over all nonnegative $\sigma(\nu)$-measurable gates, the envelope is generally different. Writing $Y = d^\top R - \lambda$, one obtains

$$\sup_{z \in L_+^2(\sigma(\nu))} \mathbb{E}\left[zY - \frac{1}{2} z^2 \right] = \frac{1}{2} \mathbb{E}\left[\big(\mathbb{E}[Y \mid \nu] \big)_+^2 \right],$$

with optimizer $z = (\mathbb{E}[Y \mid \nu])_+$. This is not the oracle utility $\mathcal{U}_\lambda(d; R)$ used in this paper. The gap studied here is therefore the gap between the sample-wise sparse-coding oracle and the one-ReLU amortized encoder class.

H.5 Subgradient formula for amortized utility

For fixed (e, b), the map $d \mapsto \Gamma_\lambda(d, e, b; R, \nu)$ is affine with gradient

$$\nabla_d \Gamma_\lambda(d, e, b; R, \nu) = \mathbb{E}[R z_{e,b}(\nu)].$$

Therefore $\mathcal{A}_\lambda$ is a supremum of affine functionals. Under optimizer attainment and boundedness of the active-gradient set, the standard Danskin–Clarke theorem gives

$$\partial^\circ \mathcal{A}_\lambda(d; R, \nu) = \operatorname{conv}\{ \mathbb{E}[R z_{e,b}(\nu)] : (e, b) \in \mathcal{M}(d) \}.$$

Projecting onto $d^\perp$ yields valid sphere ascent directions.

Proposition H.4 (common-field approximation)**.** *Let*

$$g_r^\star = (d_r^\top R_{-r} - \lambda)_+, \qquad h_r^\star = (d_r^\top r - \lambda)_+.$$

Then

$$\left\| \dot{d}_r - \operatorname*{grad}_{S^{d-1}} \mathcal{U}_\lambda(d_r; R_{-r}) \right\| \le \mathbb{E}[\|R_{-r}\| |z_r - g_r^\star|].$$

Moreover,

$$\left\| \operatorname*{grad}_{S^{d-1}} \mathcal{U}_\lambda(d_r; R_{-r}) - \operatorname*{grad}_{S^{d-1}} \mathcal{U}_\lambda(d_r; r) \right\| \le \mathbb{E}[\|r\| z_r].$$

Consequently,

$$\|\dot{d}_r - \operatorname*{grad}_{S^{d-1}} \mathcal{U}_\lambda(d_r; r)\| \le \mathbb{E}[\|R_{-r}\| |z_r - g_r^\star|] + \mathbb{E}[\|r\| z_r].$$

H.6 Proof of Proposition H.4

Let $g_r^\star = (d_r^\top R_{-r} - \lambda)_+$. By Theorem H.2, the oracle one-slot field for the leave-one-out residual satisfies

$$\operatorname*{grad}_{S^{d-1}} \mathcal{U}_\lambda(d_r; R_{-r}) = P_{d_r^\perp} \mathbb{E}[R_{-r} g_r^\star].$$

By Proposition H.1, the actual decoder flow is

$$\dot{d}_r = P_{d_r^\perp} \mathbb{E}[R_{-r} z_r].$$

Subtracting yields

$$\dot{d}_r - \operatorname*{grad}_{S^{d-1}} \mathcal{U}_\lambda(d_r; R_{-r}) = P_{d_r^\perp} \mathbb{E}[R_{-r}(z_r - g_r^\star)],$$

and therefore

$$\left\| \dot{d}_r - \operatorname*{grad}_{S^{d-1}} \mathcal{U}_\lambda(d_r; R_{-r}) \right\| \le \mathbb{E}[\|R_{-r}\| \, |z_r - g_r^\star|],$$

which is the amortization term.

For the common-field comparison, define

$$h_r^\star := (d_r^\top r - \lambda)_+.$$

Then

$$\operatorname*{grad}_{S^{d-1}} \mathcal{U}_\lambda(d_r; r) = P_{d_r^\perp} \mathbb{E}[r h_r^\star].$$

Hence

$$\operatorname*{grad}_{S^{d-1}} \mathcal{U}_\lambda(d_r; R_{-r}) - \operatorname*{grad}_{S^{d-1}} \mathcal{U}_\lambda(d_r; r) = P_{d_r^\perp} \mathbb{E}[R_{-r} g_r^\star - r h_r^\star].$$

Using $R_{-r} = r + d_r z_r$ and $P_{d_r^\perp}(d_r z_r g_r^\star) = 0$, this becomes

$$P_{d_r^\perp} \mathbb{E}[r(g_r^\star - h_r^\star)].$$

Since $t \mapsto (t - \lambda)_+$ is 1-Lipschitz and

$$d_r^\top R_{-r} - d_r^\top r = z_r,$$

we have

$$|g_r^\star - h_r^\star| \le z_r.$$

Therefore

$$\left\| \operatorname*{grad}_{S^{d-1}} \mathcal{U}_\lambda(d_r; R_{-r}) - \operatorname*{grad}_{S^{d-1}} \mathcal{U}_\lambda(d_r; r) \right\| \le \mathbb{E}[\|r\| \, z_r].$$

Combining this bound with the amortization term proves the result.

H.7 Amortization-induced ranking reversal and pair-only worlds

This stress-test subsection proves the amortization-rescue claims used in Section 5. The deferred encoder-measurable envelope remark above clarifies what is and is not meant by oracle utility. The main points are: an amortized SAE can reverse an oracle ranking only through a direction-dependent amortization gap; this reversal is impossible when the residual map is affine in the encoder input; and pair-only worlds provide a simple case where both one-slot and exact multi-slot sparse coding favor non-ground-truth directions.

Proposition H.5 (Amortization-induced ranking reversal)**.** *For a unit direction d, define*

$$\mathrm{Gap}_\lambda(d; R, \nu) := \mathcal{U}_\lambda(d; R) - \mathcal{A}_\lambda(d; R, \nu) \geq 0.$$

Then for any unit directions u, v,

$$\mathcal{A}_\lambda(u; R, \nu) - \mathcal{A}_\lambda(v; R, \nu) = \mathcal{U}_\lambda(u; R) - \mathcal{U}_\lambda(v; R) + \mathrm{Gap}_\lambda(v; R, \nu) - \mathrm{Gap}_\lambda(u; R, \nu).$$

Consequently, if

$$\mathcal{U}_\lambda(v; R) > \mathcal{U}_\lambda(u; R),$$

then the amortized utility ranks u above v only if

$$\mathrm{Gap}_\lambda(v; R, \nu) - \mathrm{Gap}_\lambda(u; R, \nu) > \mathcal{U}_\lambda(v; R) - \mathcal{U}_\lambda(u; R).$$

Proof. By definition,

$$\mathcal{A}_\lambda(d; R, \nu) = \mathcal{U}_\lambda(d; R) - \mathrm{Gap}_\lambda(d; R, \nu).$$

Applying this identity to u and v, then subtracting, gives

$$\mathcal{A}_\lambda(u) - \mathcal{A}_\lambda(v) = \mathcal{U}_\lambda(u) - \mathcal{U}_\lambda(v) + \mathrm{Gap}_\lambda(v) - \mathrm{Gap}_\lambda(u).$$

The strict reversal condition follows immediately. □

Proposition H.6 (No amortized landscape reshaping under affine residual maps)**.** *Assume*

$$R(\nu) = M\nu + r_0 \qquad \text{a.s.}$$

for some matrix M and vector r_0. Then for every unit direction d,

$$\mathcal{A}_\lambda(d; R, \nu) = \mathcal{U}_\lambda(d; R).$$

Therefore the amortized and oracle directional landscapes coincide exactly.

Proof. The oracle gate for direction d is

$$g_d^\star(\nu) = (d^\top R(\nu) - \lambda)_+.$$

If $R(\nu) = M\nu + r_0$, then

$$g_d^\star(\nu) = \left((M^\top d)^\top \nu + d^\top r_0 - \lambda\right)_+.$$

This is exactly a single affine-ReLU. By the exact bridge criterion,

$$\mathcal{A}_\lambda(d; R, \nu) = \mathcal{U}_\lambda(d; R).$$

Since this holds for every d, the full amortized directional landscape agrees with the oracle landscape. □

Remark H.7 (Scope of the affine no-reshaping result)**.** The result applies to the untied one-ReLU encoder class used in the bridge theorem. Tied encoders impose an additional restriction $e_r = d_r$, and therefore may have a nonzero amortization gap even when $R(\nu)$ is affine. Any ranking change in the tied case is due to this extra restriction, not to the generic one-ReLU amortization gap analyzed here.

Pair-only one-slot utility. Let $u_1, u_2, u_3 \in S^{d-1}$ be orthonormal and let $A > 0$. Suppose the data distribution is

$$x_{12} = A(u_1 + u_2), \qquad x_{13} = A(u_1 + u_3),$$

with probability $1/2$ each.

For the ground-truth direction u_1,

$$\langle u_1, x_{12} \rangle = A, \qquad \langle u_1, x_{13} \rangle = A.$$

Hence

$$\mathcal{U}_\lambda(u_1) = \frac{1}{2} \cdot \frac{1}{2}(A - \lambda)_+^2 + \frac{1}{2} \cdot \frac{1}{2}(A - \lambda)_+^2 = \frac{1}{2}(A - \lambda)_+^2.$$

For the pair direction

$$d_{12} := \frac{u_1 + u_2}{\sqrt{2}},$$

one has

$$\langle d_{12}, x_{12} \rangle = \sqrt{2}A, \qquad \langle d_{12}, x_{13} \rangle = \frac{A}{\sqrt{2}}.$$

Therefore

$$\mathcal{U}_\lambda(d_{12}) = \frac{1}{4}(\sqrt{2}A - \lambda)_+^2 + \frac{1}{4}(A/\sqrt{2} - \lambda)_+^2.$$

In particular, if

$$A < \lambda < \sqrt{2}A,$$

then

$$\mathcal{U}_\lambda(u_1) = 0, \qquad \mathcal{U}_\lambda(d_{12}) = \frac{1}{4}(\sqrt{2}A - \lambda)^2 > 0.$$

Thus the non-ground-truth pair direction is strictly preferred to the ground-truth direction u_1. Since $R(\nu) = \nu = x$ in the one-slot setting, Proposition H.6 implies that the untied one-slot SAE has exactly the same ranking.

Proposition H.8 (Exact sparse coding prefers pair features in the pair-only model)**.** *In the pair-only model above, define*

$$D_{\mathrm{GT}} = [u_1, u_2, u_3]$$

and

$$d_{12} = \frac{u_1 + u_2}{\sqrt{2}}, \qquad d_{13} = \frac{u_1 + u_3}{\sqrt{2}}, \qquad D_{\mathrm{pair}} = [d_{12}, d_{13}].$$

For an equal-budget comparison, also let $D_{\mathrm{pair}}^+ = [d_{12}, d_{13}, e]$ for any unit vector e. Let

$$\Phi_\lambda(D) = \mathbb{E}\left[\min_{z \geq 0} \frac{1}{2}\|x - Dz\|^2 + \lambda \mathbf{1}^\top z\right].$$

Then

$$\Phi_\lambda(D_{\mathrm{pair}}^+) \leq \Phi_\lambda(D_{\mathrm{pair}}) < \Phi_\lambda(D_{\mathrm{GT}}) \qquad \text{for every } 0 < \lambda < \sqrt{2}A,$$

and $\Phi_\lambda(D_{\mathrm{pair}}^+) = \Phi_\lambda(D_{\mathrm{pair}}) = \Phi_\lambda(D_{\mathrm{GT}}) = A^2$ for $\lambda \geq \sqrt{2}A$.

Proof. First consider D_{GT}. Since u_1, u_2, u_3 are orthonormal, the nonnegative Lasso separates coordinatewise. For $x_{12} = Au_1 + Au_2$, the optimal code is

$$z_1^\star = z_2^\star = (A - \lambda)_+, \qquad z_3^\star = 0.$$

The same loss is obtained for x_{13}. Hence

$$\Phi_\lambda(D_{\mathrm{GT}}) = \begin{cases} 2\lambda A - \lambda^2, & 0 \leq \lambda < A, \\ A^2, & \lambda \geq A. \end{cases}$$

Indeed, for $0 \le \lambda < A$, the residual on an active coordinate is λ, giving reconstruction cost λ^2 across the two active coordinates and sparsity cost $2\lambda(A-\lambda)$, hence total $2\lambda A - \lambda^2$. For $\lambda \ge A$, the zero code is optimal and the loss is

$$\frac{1}{2}\|x_{12}\|^2 = A^2.$$

Now consider D_{pair}. For x_{12},

$$x_{12} = \sqrt{2}A\, d_{12}.$$

The candidate code

$$z_{12}^\star = (\sqrt{2}A - \lambda)_+, \qquad z_{13}^\star = 0$$

satisfies the KKT conditions. When $0 \le \lambda < \sqrt{2}A$, the residual is

$$r = x_{12} - d_{12} z_{12}^\star = \lambda d_{12}.$$

The active KKT condition is $d_{12}^\top r = \lambda$. The inactive KKT inequality is

$$d_{13}^\top r = \lambda\, d_{13}^\top d_{12} = \frac{\lambda}{2} \le \lambda.$$

Thus the candidate is optimal. The same argument holds for x_{13} by symmetry. Therefore

$$\Phi_\lambda(D_{\mathrm{pair}}) = \begin{cases} \sqrt{2}\lambda A - \frac{1}{2}\lambda^2, & 0 \le \lambda < \sqrt{2}A, \\ A^2, & \lambda \ge \sqrt{2}A. \end{cases}$$

It remains to compare the two expressions. If $0 < \lambda < A$, then

$$\Phi_\lambda(D_{\mathrm{GT}}) - \Phi_\lambda(D_{\mathrm{pair}}) = (2\lambda A - \lambda^2) - \left(\sqrt{2}\lambda A - \frac{1}{2}\lambda^2\right) = \lambda A(2-\sqrt{2}) - \frac{1}{2}\lambda^2.$$

Since $\lambda < A$,

$$\lambda A(2-\sqrt{2}) - \frac{1}{2}\lambda^2 > \lambda A\left(2 - \sqrt{2} - \frac{1}{2}\right) > 0.$$

If $A \le \lambda < \sqrt{2}A$, then

$$\Phi_\lambda(D_{\mathrm{GT}}) = A^2$$

and

$$\Phi_\lambda(D_{\mathrm{pair}}) = \sqrt{2}\lambda A - \frac{1}{2}\lambda^2.$$

Thus

$$\Phi_\lambda(D_{\mathrm{GT}}) - \Phi_\lambda(D_{\mathrm{pair}}) = A^2 - \sqrt{2}\lambda A + \frac{1}{2}\lambda^2 = \frac{1}{2}(\lambda - \sqrt{2}A)^2 > 0.$$

Since D_{pair}^+ contains the two columns of D_{pair}, its optimal value is no larger than $\Phi_\lambda(D_{\mathrm{pair}})$, giving the strict equal-budget comparison for $0 < \lambda < \sqrt{2}A$. For $\lambda \ge \sqrt{2}A$, every unit column d satisfies $d^\top x \le \|x\| = \sqrt{2}A \le \lambda$ on both samples, so the zero code is KKT-optimal for all three dictionaries and the loss is A^2. This proves the claim. □

Proposition H.9 (Support-faithful pair-only firing need not recover the atom direction)**.** *In the pair-only model, let $a \ge 0$ be a gate satisfying*

$$a(x_{12}) > 0, \qquad a(x_{13}) > 0.$$

Then

$$\mu(a) := \mathbb{E}[xa]$$

is not parallel to u_1. In particular, a gate that fires on all events containing latent feature 1 does not recover the decoder direction u_1.

Proof. Write

$$a_{12} := a(x_{12}) > 0, \qquad a_{13} := a(x_{13}) > 0.$$

Then

$$\mu(a) = \frac{1}{2} A a_{12}(u_1 + u_2) + \frac{1}{2} A a_{13}(u_1 + u_3).$$

Equivalently,

$$\mu(a) = \frac{A}{2}(a_{12} + a_{13})u_1 + \frac{A}{2} a_{12} u_2 + \frac{A}{2} a_{13} u_3.$$

Since $a_{12}, a_{13} > 0$ and u_1, u_2, u_3 are orthonormal, the u_2 and u_3 components are nonzero. Hence $\mu(a) \not\parallel u_1$. □

Proposition H.10 (Generic pair-only support contamination)**.** *Assume the residualized atoms $\{\tilde{w}_i : i \notin B\}$ are linearly independent. Fix $j \notin B$, and suppose*

$$\mathbb{P}(E_{\{j\}}(B)) = 0.$$

Let $a \geq 0$, with $\mathbb{E}[\|R_B\| a] < \infty$, satisfy

$$a > 0 \qquad \text{a.s. on } \{\alpha_j > 0\}.$$

If there exists $i \notin B$, $i \neq j$, such that

$$\mathbb{P}(\alpha_i > 0, \alpha_j > 0) > 0,$$

then

$$\mu(a; B) \not\parallel \tilde{w}_j.$$

Proof. Since $a > 0$ a.s. on $\{\alpha_j > 0\}$, and since

$$\mathbb{P}(\alpha_i > 0, \alpha_j > 0) > 0,$$

we have

$$\mathbb{E}[\alpha_i a] > 0.$$

The gated moment is

$$\mu(a; B) = \sum_{k \notin B} \mathbb{E}[\alpha_k a] \tilde{w}_k.$$

By the moment characterization theorem, under linear independence this vector is parallel to $\tilde{w}_j$ only if all coefficients on $\tilde{w}_i$, $i \neq j$, vanish. The i-th coefficient is strictly positive, so $\mu(a; B) \not\parallel \tilde{w}_j$. □

Remark H.11 (Special cancellation is an exception)**.** The negative pair-only statements above are generic for linearly independent positive geometry. They do not exclude special cancellation geometries. For example, in a regular simplex with exchangeable off-target gated moments, the off-target contributions can cancel exactly, as in Proposition G.5. Such cases are structured exceptions rather than consequences of vanilla sparse reconstruction.

I Residual-field transfer and bias effects

The residual-field transfer results connect the positive latent-ray oracle field

$$\mathcal{U}_\lambda(\cdot; R_B)$$

to the SAE leave-one-out field

$$\mathcal{U}_\lambda(\cdot; R_{-r}).$$

The transfer is not automatic: it holds under an explicit residual-compatibility condition and a strict margin.

When the latent residual field and the SAE leave-one-out residual differ, biases and residual mismatch spend recovery margin.

I.1 Shifted support-event decomposition

Proposition I.1 (Shifted support-event decomposition)**.** *Let*

$$x = \sum_{i=1}^{F} \alpha_i w_i, \qquad \alpha_i \geq 0,$$

and let

$$R_B = \pi_B x = \sum_{i \notin B} \alpha_i \tilde{w}_i, \qquad \tilde{w}_i := \pi_B w_i,$$

be the projected residual. For a deterministic shift $b \in \mathbb{R}^d$, define

$$R_{B,b} := \pi_B(x - b) = R_B - \beta_B, \qquad \beta_B := \pi_B b.$$

Then, for every unit vector v,

$$\mathcal{U}_\lambda(v; R_{B,b}) = \sum_{T \subseteq [F] \setminus B} \frac{1}{2} \mathbb{E}\left[\left(\sum_{i \in T} \alpha_i \langle v, \tilde{w}_i \rangle - \langle v, \beta_B \rangle - \lambda \right)_+^2 \mathbf{1}_{E_T(B)} \right].$$

Equivalently, the deterministic shift turns the original scalar threshold λ into the direction-dependent threshold

$$\lambda_v^b := \lambda + \langle v, \beta_B \rangle.$$

In particular, the empty support event $T = \varnothing$ contributes

$$\frac{1}{2} \mathbb{E}\left[(-\langle v, \beta_B \rangle - \lambda)_+^2 \mathbf{1}_{E_\varnothing(B)} \right].$$

Proof. On the event $E_T(B)$,

$$R_B = \sum_{i \in T} \alpha_i \tilde{w}_i.$$

Therefore

$$R_{B,b} = R_B - \beta_B = \sum_{i \in T} \alpha_i \tilde{w}_i - \beta_B,$$

and

$$\langle v, R_{B,b} \rangle - \lambda = \sum_{i \in T} \alpha_i \langle v, \tilde{w}_i \rangle - \langle v, \beta_B \rangle - \lambda.$$

Substituting into

$$\mathcal{U}_\lambda(v; R_{B,b}) = \frac{1}{2} \mathbb{E}[(\langle v, R_{B,b} \rangle - \lambda)_+^2]$$

and summing over the full partition

$$\{E_T(B) : T \subseteq [F] \setminus B\}$$

gives the formula. □

This differs structurally from Theorem 3.2, where the sum may be restricted to nonempty residual supports because $R_B = 0$ on $E_\varnothing(B)$. After a shift, $R_{B,b} = -\beta_B$ on the empty support event, so absence events can contribute positive utility whenever

$$-\langle v, \beta_B \rangle > \lambda.$$

I.2 Uniform perturbation bounds for utility and gradient

Lemma I.2 (Uniform utility-field perturbation, expanded)**.** *Let $R, S \in L^2(\mathbb{R}^d)$ and $\lambda \geq 0$. Then*

$$\sup_{\|v\|=1} \left| \mathcal{U}_\lambda(v; R) - \mathcal{U}_\lambda(v; S) \right| \leq \frac{1}{2} \|R - S\|_{L^2} \left(\|R\|_{L^2} + \|S\|_{L^2} \right) =: \eta(R, S).$$

Proof of Lemma I.2. For real numbers a, b,

$$|a_+^2 - b_+^2| = |a_+ - b_+|(a_+ + b_+) \le |a - b|(a_+ + b_+).$$

Set

$$a = \langle v, R\rangle - \lambda, \qquad b = \langle v, S\rangle - \lambda.$$

Since $\lambda \ge 0$,

$$a_+ \le |\langle v, R\rangle|, \qquad b_+ \le |\langle v, S\rangle|.$$

Therefore

$$\left|\mathcal{U}_\lambda(v; R) - \mathcal{U}_\lambda(v; S)\right| \le \frac{1}{2}\mathbb{E}\left[|\langle v, R - S\rangle|\left(|\langle v, R\rangle| + |\langle v, S\rangle|\right)\right].$$

By Cauchy–Schwarz,

$$\mathbb{E}[|\langle v, R - S\rangle||\langle v, R\rangle|] \le \|\langle v, R - S\rangle\|_{L^2}\|\langle v, R\rangle\|_{L^2} \le \|R - S\|_{L^2}\|R\|_{L^2},$$

where the final inequality uses $\|v\| = 1$. Similarly,

$$\mathbb{E}[|\langle v, R - S\rangle||\langle v, S\rangle|] \le \|R - S\|_{L^2}\|S\|_{L^2}.$$

Combining the two estimates gives the stated uniform bound. □

Lemma I.3 (Gradient-field perturbation)**.** *Let $R, S \in L^2(\mathbb{R}^d)$. Then, for every unit vector v,*

$$\|\nabla_v \mathcal{U}_\lambda(v; R) - \nabla_v \mathcal{U}_\lambda(v; S)\| \le \left(\|R\|_{L^2} + \|S\|_{L^2}\right)\|R - S\|_{L^2}.$$

The same bound holds after projecting onto $v^\perp$.

Proof of Lemma I.3. Using

$$\nabla_v \mathcal{U}_\lambda(v; R) = \mathbb{E}[(\langle v, R\rangle - \lambda)_+ R],$$

write

$$\nabla_v \mathcal{U}_\lambda(v; R) - \nabla_v \mathcal{U}_\lambda(v; S) = \mathbb{E}[(\langle v, R\rangle - \lambda)_+(R - S)] + \mathbb{E}[((\langle v, R\rangle - \lambda)_+ - (\langle v, S\rangle - \lambda)_+)\, S].$$

For the first term,

$$(\langle v, R\rangle - \lambda)_+ \le |\langle v, R\rangle| \le \|R\|,$$

so Cauchy–Schwarz gives the bound

$$\|\mathbb{E}[(\langle v, R\rangle - \lambda)_+(R - S)]\| \le \|R\|_{L^2}\|R - S\|_{L^2}.$$

For the second term, the positive-part map is 1-Lipschitz, hence

$$|(\langle v, R\rangle - \lambda)_+ - (\langle v, S\rangle - \lambda)_+| \le |\langle v, R - S\rangle| \le \|R - S\|.$$

Another Cauchy–Schwarz application gives

$$\|\mathbb{E}[((\langle v, R\rangle - \lambda)_+ - (\langle v, S\rangle - \lambda)_+)\, S]\| \le \|R - S\|_{L^2}\|S\|_{L^2}.$$

Adding the two estimates proves the claim. Orthogonal projection onto $v^\perp$ cannot increase norm. □

I.3 Margin transfer theorem

For the SAE leave-one-out residual define

$$\varepsilon_{B,-r} := \|R_{-r} - R_B\|_{L^2} = \left\| P_B x - b_{\mathrm{dec}} - \sum_{s \ne r} d_s z_s \right\|_{L^2},$$

and

$$\eta_{B,-r} := \eta(R_B, R_{-r}) = \frac{1}{2}\|R_{-r} - R_B\|_{L^2}\left(\|R_B\|_{L^2} + \|R_{-r}\|_{L^2}\right).$$

Theorem I.4 (Residual-field transfer to amortized SAE utility)**.** *Let $\mathcal{C} \subseteq S^{d-1}$ be a competitor set, and suppose a direction u has latent oracle margin*

$$\gamma := \mathcal{U}_\lambda(u; R_B) - \sup_{v \in \mathcal{C}} \mathcal{U}_\lambda(v; R_B) > 0.$$

Let

$$\mathrm{Gap}_\lambda(u; R_{-r}, \nu) = \mathcal{U}_\lambda(u; R_{-r}) - \mathcal{A}_\lambda(u; R_{-r}, \nu)$$

be the amortization gap of u on the SAE residual. Then

$$\mathcal{A}_\lambda(u; R_{-r}, \nu) - \sup_{v \in \mathcal{C}} \mathcal{A}_\lambda(v; R_{-r}, \nu) \geq \gamma - 2\eta_{B,-r} - \mathrm{Gap}_\lambda(u; R_{-r}, \nu).$$

In particular, if

$$\gamma > 2\eta_{B,-r} + \mathrm{Gap}_\lambda(u; R_{-r}, \nu),$$

then the amortized SAE utility ranks u above every direction in $\mathcal{C}$.

Proof of Theorems 5.3 and I.4. By Lemma I.2,

$$\mathcal{U}_\lambda(u; R_{-r}) \geq \mathcal{U}_\lambda(u; R_B) - \eta_{B,-r},$$

and

$$\sup_{v \in \mathcal{C}} \mathcal{U}_\lambda(v; R_{-r}) \leq \sup_{v \in \mathcal{C}} \mathcal{U}_\lambda(v; R_B) + \eta_{B,-r}.$$

Therefore

$$\mathcal{U}_\lambda(u; R_{-r}) - \sup_{v \in \mathcal{C}} \mathcal{U}_\lambda(v; R_{-r}) \geq \gamma - 2\eta_{B,-r}.$$

By definition,

$$\mathcal{A}_\lambda(u; R_{-r}, \nu) = \mathcal{U}_\lambda(u; R_{-r}) - \mathrm{Gap}_\lambda(u; R_{-r}, \nu).$$

Moreover, by Theorem H.2,

$$\mathcal{A}_\lambda(v; R_{-r}, \nu) \leq \mathcal{U}_\lambda(v; R_{-r}) \qquad \forall v.$$

Hence

$$\sup_{v \in \mathcal{C}} \mathcal{A}_\lambda(v; R_{-r}, \nu) \leq \sup_{v \in \mathcal{C}} \mathcal{U}_\lambda(v; R_{-r}).$$

Combining the last three displays gives the stated inequality. □

Applying this theorem with

$$\mathcal{C}_\varepsilon = \{v \in S^{d-1} : \angle(v, u_j) \geq \varepsilon\}$$

gives the transferred singleton-localization statement used in Section 5.1: if

$$\mathcal{U}_\lambda(u_j; R_B) - \sup_{v \in \mathcal{C}_\varepsilon} \mathcal{U}_\lambda(v; R_B) > 2\eta_{B,-r} + \mathrm{Gap}_\lambda(u_j; R_{-r}, \nu),$$

then the amortized SAE utility ranks u_j above every direction at angle at least ε from u_j.

I.4 Transfer of local merge pressure

Corollary I.5 (Transferred local merge pressure)**.** *Let $u \in S^{d-1}$, $h \in u^\perp \cap S^{d-1}$, and suppose the latent oracle gradient satisfies*

$$\rho := \langle \nabla_v \mathcal{U}_\lambda(u; R_B), h \rangle = \mathbb{E}[(\langle u, R_B \rangle - \lambda)_+ \langle h, R_B \rangle] > 0.$$

If

$$\rho > \big(\|R_B\|_{L^2} + \|R_{-r}\|_{L^2}\big)\|R_{-r} - R_B\|_{L^2},$$

then

$$\langle \nabla_v \mathcal{U}_\lambda(u; R_{-r}), h \rangle > 0.$$

Consequently, u is not a local maximizer of

$$v \mapsto \mathcal{U}_\lambda(v; R_{-r})$$

on the sphere.

Proof. By Lemma I.3,

$$|\langle \nabla_v \mathcal{U}_\lambda(u; R_{-r}) - \nabla_v \mathcal{U}_\lambda(u; R_B), h\rangle| \le \big(\|R_B\|_{L^2} + \|R_{-r}\|_{L^2}\big)\|R_{-r} - R_B\|_{L^2}.$$

The assumed strict inequality implies

$$\langle \nabla_v \mathcal{U}_\lambda(u; R_{-r}), h\rangle > 0.$$

For the great-circle perturbation

$$v_\theta = \cos\theta\, u + \sin\theta\, h,$$

we have

$$\frac{d}{d\theta}\mathcal{U}_\lambda(v_\theta; R_{-r})\Big|_{\theta=0} = \langle \nabla_v \mathcal{U}_\lambda(u; R_{-r}), h\rangle > 0.$$

Thus the utility increases along a feasible tangent direction, so u is not a local maximizer on the sphere. □

I.5 Special case: bias-only mismatch

Suppose the previously learned slots exactly reconstruct the V_B-component,

$$\sum_{s\neq r} d_s z_s = P_B x.$$

Then

$$P_{V_B^\perp} R_{-r} = R_B - \pi_B b_{\mathrm{dec}}.$$

Consequently,

$$\|P_{V_B^\perp} R_{-r} - R_B\|_{L^2} = \|\pi_B b_{\mathrm{dec}}\|.$$

If $b_{\mathrm{dec}} \in V_B$, then $\pi_B b_{\mathrm{dec}} = 0$, so the projected residual seen by directions in $V_B^\perp$ coincides with R_B, up to the amortization gap.

J Direction–support separation proofs

Decoder direction is a statement about a gated first moment, whereas support recovery is a statement about the event on which the gate fires. The proofs below show why the two coincide only under extra fixed-direction and exact-support assumptions.

Direction recovery and support recovery separate through exact counterexamples; they coincide under rank-one event support.

J.1 Proof of Theorem 6.1

Expand

$$\mu(a; B) = \sum_{i\notin B} \mathbb{E}[\alpha_i a]\, \tilde{w}_i.$$

Because $\mu(a; B) \neq 0$ and $\tilde{w}_j \neq 0$, direction alignment means $\mu(a; B) \in \mathrm{span}(\tilde{w}_j)$. This holds if and only if the sum of the off-j contributions lies in $\mathrm{span}(\tilde{w}_j)$, which proves the first claim. Under linear independence of the residualized atoms, all coefficients except possibly the j-th must vanish. Since $\alpha_i, a \ge 0$, the zero-moment condition $\mathbb{E}[\alpha_i a] = 0$ is equivalent to $\alpha_i a = 0$ almost surely.

Proposition J.1 (Direction without support)**.** *Let $x = \alpha_j w_j$ with $\mathbb{E}[\alpha_j] < \infty$. For any $G \subseteq \{\alpha_j > 0\}$ with positive conditional probability, the gate $a = \mathbf{1}_G$ gives $\mu(a) \parallel w_j$. If the conditional law admits positive-probability subsets of arbitrarily small mass, then support recall can be made arbitrarily low while maintaining exact direction recovery.*

Proposition J.2 (Support without direction)**.** *Let $x = \alpha_1 w_1 + \alpha_2 w_2$ with $w_1 \perp w_2$ and both support events $\{1\}$ and $\{1, 2\}$ having positive probability. Set $a = \mathbf{1}_{\{\alpha_1 > 0\}}$, which recovers the presence event of atom 1 exactly. Then $\mu(a) = \mathbb{E}[\alpha_1 a]w_1 + \mathbb{E}[\alpha_2 a]w_2$; whenever $\mathbb{E}[\alpha_2 a] > 0$ the decoder direction is rotated away from w_1.*

J.2 Proofs of Propositions J.1 and J.2

For Proposition J.1, in the one-atom world one has

$$\mu(a) = \mathbb{E}[\alpha_j a] w_j,$$

so the decoder direction is exactly w_j. The support recall is $\mathbb{P}(G \mid \{\alpha_j > 0\})$; if the conditional law admits positive-probability subsets of arbitrarily small mass, this recall can be made arbitrarily low while preserving direction alignment.

For Proposition J.2, the activation recovers the support event $\{\alpha_1 > 0\}$ exactly by construction, but

$$\mu(a) = \mathbb{E}[\alpha_1 a] w_1 + \mathbb{E}[\alpha_2 a] w_2,$$

and the second coefficient is positive whenever the event $\{1, 2\}$ has positive probability and α_2 is positive there.

J.3 Proof of Theorem 6.2

Since $a = 0$ almost surely on A^c and $R_B = \xi u$ almost surely on A,

$$\mu(a; B) = \mathbb{E}[R_B a] = \mathbb{E}[\xi a \mathbf{1}_A] u.$$

The scalar prefactor is positive by assumption, so the normalized moment direction is exactly u.

K Sparsity, frequency matching, and support-event geometry

The support-event view makes sparsity matching and marginal-frequency matching precise but limited. Expected sparsity and marginal frequencies are low-order summaries of the latent support law. The sparse-reconstruction objective depends on the full event law $\mathbb{P}(S = T)$, together with the amplitudes and geometry on each support event.

> Matching global sparsity or marginal frequencies is not enough. The missing object is the full joint support-event law, especially pair and burst events that can dominate thresholded utility.

K.1 Cardinality matching and support recovery

Let

$$Z_i := \mathbf{1}_{\{\alpha_i > 0\}}, \qquad S := \{i : Z_i = 1\}, \qquad N := |S| = \sum_{i=1}^{F} Z_i.$$

Let a learned code have active set

$$\widehat{S}(x) := \{r : z_r(x) > 0\}, \qquad \widehat{N}(x) := |\widehat{S}(x)|.$$

The ground-truth expected sparsity is

$$s_\star := \mathbb{E}N = \sum_{i=1}^{F} p_i, \qquad p_i := \mathbb{P}(Z_i = 1).$$

Support recovery, samplewise sparsity recovery, and expected sparsity matching satisfy

$$\widehat{S} = S \text{ a.s.} \quad \Longrightarrow \quad \widehat{N} = N \text{ a.s.} \quad \Longrightarrow \quad \mathbb{E}\widehat{N} = \mathbb{E}N.$$

The converses fail in general.

Lemma K.1 (Cardinality mismatch lower-bounds support error)**.** *For any two finite sets A, B,*

$$|A \triangle B| \geq \big||A| - |B|\big|.$$

Consequently, if a learned code has fixed samplewise sparsity $\widehat{N} = k$, then

$$\mathbb{E}|\widehat{S} \triangle S| \geq \mathbb{E}|N - k|.$$

In particular, exact support recovery under a deterministic k-sparse code requires

$$N = k \qquad \text{a.s.}$$

Proof. The identity

$$|A \triangle B| = |A| + |B| - 2|A \cap B|$$

and the inequalities $|A \cap B| \le |A|$, $|A \cap B| \le |B|$ imply

$$|A \triangle B| \ge |A| - |B|, \qquad |A \triangle B| \ge |B| - |A|.$$

Taking the maximum gives the deterministic inequality. Applying it samplewise gives the expectation bound. If $\widehat{S} = S$ a.s. and $\widehat{N} = k$ a.s., then $N = k$ a.s. □

Thus expected sparsity matching fixes only a first moment, and fixed samplewise sparsity fixes only a cardinality. Neither specifies which support event occurred.

K.2 Thresholding under the correct orthogonal dictionary

Even if the dictionary directions are correct, the ℓ_1 threshold can alter support semantics.

Proposition K.2 (Amplitude thresholding under an orthogonal teacher dictionary)**.** *Assume $w_1, \dots, w_F$ are orthonormal and*

$$x = \sum_{i=1}^{F} \alpha_i w_i, \qquad \alpha_i \ge 0.$$

For the teacher dictionary $D = (w_1, \dots, w_F)$, the exact nonnegative ℓ_1 code is

$$z_i^\star(x) = (\alpha_i - \lambda)_+.$$

Hence

$$\operatorname{supp} z^\star(x) = \{i : \alpha_i > \lambda\}.$$

Exact support recovery with the teacher dictionary holds if and only if

$$\mathbb{P}(0 < \alpha_i \le \lambda) = 0 \qquad \textit{for every } i.$$

Moreover,

$$\mathbb{E}|\operatorname{supp} z^\star(x)| = \sum_{i=1}^{F} \mathbb{P}(\alpha_i > \lambda) \le \sum_{i=1}^{F} \mathbb{P}(\alpha_i > 0) = s_\star,$$

with equality if and only if $\mathbb{P}(0 < \alpha_i \le \lambda) = 0$ for every i.

Proof. For an orthonormal dictionary, the objective separates:

$$\frac{1}{2}\left\| x - \sum_i z_i w_i \right\|^2 + \lambda \sum_i z_i = \sum_i \left[\frac{1}{2}(\alpha_i - z_i)^2 + \lambda z_i \right].$$

The nonnegative scalar minimizer is $z_i^\star = (\alpha_i - \lambda)_+$. The support and expectation claims follow immediately. □

This proposition isolates an amplitude-threshold mismatch. The rest of the appendix concerns a different issue: even when sparsity statistics are matched, the event geometry can make non-teacher directions preferable.

K.3 Marginal frequencies do not determine oracle preference

The support-event decomposition depends on the full law of S, not only on the marginals $p_i = \mathbb{P}(i \in S)$. The following proposition gives an exact two-feature separation.

Proposition K.3 (Same marginals, different oracle-preferred directions)**.** *Let $u, w \in \mathbb{R}^d$ be orthonormal, let $A > 0$, and set*

$$x = A(Z_1 u + Z_2 w).$$

Assume

$$\mathbb{P}(Z_1 = 1) = \mathbb{P}(Z_2 = 1) = p, \qquad \mathbb{P}(Z_1 = Z_2 = 1) = q.$$

Then

$$\mathbb{P}(S = \{1\}) = p - q, \qquad \mathbb{P}(S = \{2\}) = p - q, \qquad \mathbb{P}(S = \{1, 2\}) = q.$$

Let $m = (u + w)/\sqrt{2}$. For any $\lambda \geq 0$,

$$\mathcal{U}_\lambda(u; x) = \frac{1}{2} p (A - \lambda)_+^2,$$

and

$$\mathcal{U}_\lambda(m; x) = (p - q)\frac{1}{2}(A/\sqrt{2} - \lambda)_+^2 + q\frac{1}{2}(\sqrt{2}A - \lambda)_+^2.$$

In particular, if

$$A/\sqrt{2} < \lambda < A,$$

then

$$\mathcal{U}_\lambda(m; x) = \frac{1}{2} q (\sqrt{2}A - \lambda)^2.$$

For every $0 < p \leq 1/2$, both $q = 0$ and $q = p$ are feasible with the same marginals (p, p) and the same expected sparsity $2p$. For $q = 0$,

$$\mathcal{U}_\lambda(u; x) > \mathcal{U}_\lambda(m; x),$$

whereas for $q = p$,

$$\mathcal{U}_\lambda(m; x) > \mathcal{U}_\lambda(u; x).$$

Proof. The formula for $\mathcal{U}_\lambda(u; x)$ follows because $\langle u, x \rangle = A$ exactly when $Z_1 = 1$, which has probability p. For m, singleton events have projection $A/\sqrt{2}$, and the joint event has projection $\sqrt{2}A$, giving the displayed expression.

If $A/\sqrt{2} < \lambda < A$, singleton contributions to m vanish, while teacher contributions remain. When $q = 0$, the merge has zero utility and the teacher has positive utility. When $q = p$,

$$\mathcal{U}_\lambda(m; x) - \mathcal{U}_\lambda(u; x) = \frac{1}{2} p \left[(\sqrt{2}A - \lambda)^2 - (A - \lambda)^2 \right] > 0,$$

because $\sqrt{2}A - \lambda > A - \lambda > 0$. □

Thus marginal frequency matching does not determine recovery. The two laws have the same (p_1, p_2) and the same expected sparsity, but different support-event laws.

K.4 Variance, covariance, and pair-event mass

Expected sparsity is the first moment of the support size. Pairwise mixed-event mass is controlled by second-order structure.

Lemma K.4 (Support-size variance and expected pair mass)**.** *Let* $N = \sum_{i=1}^{F} Z_i$*, with* $p_i = \mathbb{P}(Z_i = 1)$*. Then*

$$\mathrm{Var}(N) = \sum_{i=1}^{F} p_i(1-p_i) + 2\sum_{i<j} \mathrm{Cov}(Z_i, Z_j).$$

Moreover,

$$\mathbb{E}\binom{N}{2} = \sum_{i<j} \mathbb{P}(Z_i = 1, Z_j = 1) = \frac{1}{2}\left(\mathrm{Var}(N) + s_\star^2 - s_\star\right),$$

where $s_\star = \mathbb{E}N$*.*

Proof. The variance identity follows from expanding $\mathrm{Var}(\sum_i Z_i)$. For the pair count,

$$\binom{N}{2} = \frac{1}{2}N(N-1).$$

Since

$$\mathbb{E}[N(N-1)] = \mathbb{E}[N^2] - \mathbb{E}[N] = \mathrm{Var}(N) + s_\star^2 - s_\star,$$

the formula follows. □

For fixed expected sparsity $s_\star$, larger $\mathrm{Var}(N)$ implies larger expected pair co-activation mass. Positive covariance produces bursty co-activation and increases mixed-event mass. Negative covariance can stabilize the cardinality. However, variance is still only a summary: it does not identify which pairs co-activate, their geometry, their amplitudes, or whether their above-threshold contributions dominate singleton events. Conversely, zero variance does not imply singleton dominance: if $N = 2$ almost surely, then $\mathrm{Var}(N) = 0$ but $\mathbb{E}\binom{N}{2} = 1$.

K.5 Power-law marginal frequencies

Assume independent Bernoulli support indicators with

$$p_i = Ci^{-\alpha}, \qquad 0 < C \le 1, \qquad \alpha > 0, \qquad i = 1, \dots, F.$$

Let

$$N_F := \sum_{i=1}^{F} Z_i, \qquad s_F := \mathbb{E}N_F = CH_{F,\alpha}, \qquad \rho_F := \frac{s_F}{F},$$

where

$$H_{F,\alpha} := \sum_{i=1}^{F} i^{-\alpha}.$$

Proposition K.5 (Power-law sparsity asymptotics)**.** *As* $F \to \infty$*,*

$$s_F \to C\zeta(\alpha), \qquad \rho_F \sim \frac{C\zeta(\alpha)}{F} \to 0, \qquad \textit{if } \alpha > 1,$$

$$s_F \sim C\log F, \qquad \rho_F \sim \frac{C\log F}{F} \to 0, \qquad \textit{if } \alpha = 1,$$

and

$$s_F \sim \frac{C}{1-\alpha}F^{1-\alpha}, \qquad \rho_F \sim \frac{C}{1-\alpha}F^{-\alpha} \to 0, \qquad \textit{if } 0 < \alpha < 1.$$

Thus $\rho_F \to 0$ *for every* $\alpha > 0$*, but* $s_F \to \infty$ *whenever* $\alpha \le 1$*.*

Proof. These are the standard integral-test asymptotics for generalized harmonic sums. □

Under independence,

$$\mathrm{Var}(N_F) = \sum_{i=1}^{F} p_i(1-p_i) = CH_{F,\alpha} - C^2 H_{F,2\alpha}.$$

Proposition K.6 (Power-law variance and singleton disappearance)**.** *Under the independent power-law model above, if $\alpha \le 1$, then*

$$\frac{\mathrm{Var}(N_F)}{s_F^2} \to 0 \qquad \textit{and hence} \qquad \frac{N_F}{s_F} \to 1 \quad \textit{in probability}.$$

In particular,

$$\mathbb{P}(N_F = 1) \to 0.$$

Moreover,

$$\mathbb{E}\binom{N_F}{2} = \frac{1}{2}\left(s_F^2 - C^2 H_{F,2\alpha}\right) \sim \frac{1}{2}s_F^2 \qquad \textit{for } \alpha \le 1.$$

Proof. For $0 < \alpha < 1$, $s_F \asymp F^{1-\alpha}$. Also

$$H_{F,2\alpha} = \begin{cases} O(F^{1-2\alpha}), & 0 < \alpha < 1/2, \\ O(\log F), & \alpha = 1/2, \\ O(1), & 1/2 < \alpha < 1. \end{cases}$$

Thus $H_{F,2\alpha} = o(s_F^2)$, and

$$\mathrm{Var}(N_F) = CH_{F,\alpha} - C^2 H_{F,2\alpha} \sim s_F.$$

For $\alpha = 1$, $s_F \sim C \log F$ and $H_{F,2} \to \zeta(2)$, so again $\mathrm{Var}(N_F)/s_F^2 \to 0$. Chebyshev's inequality gives $N_F/s_F \to 1$ in probability. Since $s_F \to \infty$, the event $\{N_F = 1\}$ is contained in

$$\left\{ \left| \frac{N_F}{s_F} - 1 \right| \ge 1 - \frac{1}{s_F} \right\},$$

whose probability tends to zero.

Finally, by independence,

$$\mathbb{E}\binom{N_F}{2} = \sum_{i<j} p_i p_j = \frac{1}{2}\left[\left(\sum_i p_i \right)^2 - \sum_i p_i^2 \right] = \frac{1}{2}\left(s_F^2 - C^2 H_{F,2\alpha}\right).$$

The same comparison gives the asymptotic equivalence. □

Thus, for $\alpha \le 1$,

$$\rho_F \to 0, \qquad s_F \to \infty, \qquad \mathbb{P}(N_F = 1) \to 0, \qquad \mathbb{E}\binom{N_F}{2} \sim \frac{1}{2}s_F^2.$$

The representation is sparse as a fraction of all possible features, but it is not singleton-dominated.

K.6 Load transition at the hidden dimension

Let d_F be the hidden-state dimension and define the active-load ratio

$$\chi_F := \frac{s_F}{d_F}.$$

This compares the expected number of active features with the ambient dimension.

If $d_F = d$ is fixed, then for $\alpha = 1$,

$$\chi_F \sim \frac{C \log F}{d},$$

so the scale $s_F \approx d$ occurs at

$$F_c \approx \exp(d/C).$$

For $0 < \alpha < 1$,

$$\chi_F \sim \frac{C}{(1-\alpha)d} F^{1-\alpha},$$

so the scale $s_F \approx d$ occurs at

$$F_c \approx \left(\frac{(1-\alpha)d}{C} \right)^{1/(1-\alpha)}.$$

For $\alpha > 1$, $s_F \to C\zeta(\alpha)$, so there is no F-driven transition; the limiting load is

$$\chi_\infty = \frac{C\zeta(\alpha)}{d}.$$

More generally, if

$$d_F \asymp DF^\beta$$

with $D > 0$, then for $0 < \alpha < 1$,

$$\chi_F \asymp \frac{C}{(1-\alpha)D} F^{1-\alpha-\beta}.$$

Thus the load transition occurs at

$$\beta = 1 - \alpha.$$

If $\beta > 1 - \alpha$, then $\chi_F \to 0$; if $\beta < 1 - \alpha$, then $\chi_F \to \infty$.

This load transition is distinct from the sparsity-ratio limit. One can have

$$\rho_F = s_F/F \to 0$$

while simultaneously

$$s_F/d_F \to \infty.$$

Thus a representation may be sparse relative to the number of possible features but overloaded relative to the observed dimension.

K.7 Pairwise-only supports

We now analyze a regime in which sparsity matching is exact but singleton dominance is impossible. Assume

$$|S| = 2 \qquad \text{almost surely.}$$

Let

$$\mathbb{P}(S = \{i, j\}) = \pi_{ij}, \qquad \sum_{i<j} \pi_{ij} = 1.$$

Then

$$s_\star = 2, \qquad \mathrm{Var}(|S|) = 0, \qquad s_\star/F = 2/F \to 0.$$

Nevertheless,

$$\mathbb{P}(S = \{j\}) = 0 \qquad \text{for every } j.$$

Therefore the singleton margin used in Theorem 4.1 is zero for every teacher feature.

Proposition K.7 (Pairwise-only supports create local merge pressure)**.** *Assume $w_1, \dots, w_F$ are orthonormal unit vectors. On the pair event $S = \{i, j\}$, suppose*

$$R = \alpha_i w_i + \alpha_j w_j, \qquad \alpha_i, \alpha_j > 0.$$

Fix $i \neq j$. If

$$\mathbb{E}\big[(\alpha_i - \lambda)_+ \alpha_j \, \mathbf{1}_{\{S=\{i,j\}\}}\big] > 0,$$

then w_i is not a local maximizer of $v \mapsto \mathcal{U}_\lambda(v; R)$ on the sphere.

Proof. Let

$$v_\theta = \cos\theta\, w_i + \sin\theta\, w_j.$$

The sphere directional derivative is

$$\frac{d}{d\theta}\mathcal{U}_\lambda(v_\theta; R)\bigg|_{\theta=0} = \mathbb{E}[(\langle w_i, R\rangle - \lambda)_+ \langle w_j, R\rangle]\,.$$

By orthonormality, the event $S = \{i, j\}$ contributes

$$\mathbb{E}\big[(\alpha_i - \lambda)_+ \alpha_j\, \mathbf{1}_{\{S=\{i,j\}\}}\big]\,.$$

Other pair events contribute zero to this tangent derivative: if $i \notin S$, then $\langle w_i, R\rangle = 0$, and if $j \notin S$, then $\langle w_j, R\rangle = 0$. Hence the derivative is positive, so w_i is not a local maximizer. □

If amplitudes vary on a pair event, then the event need not be rank-one. On $S = \{i, j\}$,

$$R = \alpha_i w_i + \alpha_j w_j.$$

The normalized residual direction depends on the ratio α_j/α_i. Hence the pair event is rank-one if and only if this ratio is almost surely constant on the event.

For directions in the pair plane,

$$v_\theta = \cos\theta\, w_i + \sin\theta\, w_j,$$

the conditional pair-event utility is

$$U_{ij}(\theta) = \frac{1}{2}\pi_{ij}\mathbb{E}\big[(\alpha_i\cos\theta + \alpha_j\sin\theta - \lambda)_+^2 \mid S = \{i, j\}\big]\,.$$

At the teacher endpoint,

$$U'_{ij}(0) = \pi_{ij}\mathbb{E}[(\alpha_i - \lambda)_+\alpha_j \mid S = \{i, j\}]\,.$$

Thus, whenever the pair event has positive above-threshold mass, the best one-slot compressor of the pair event is pushed into the pair cone rather than remaining at a teacher ray.

Finally, exact two-sparsity does not by itself identify teacher rays even if exact reconstruction is possible. Suppose $F = 2$ and

$$x = \alpha_1 w_1 + \alpha_2 w_2, \qquad \frac{\alpha_2}{\alpha_1} \in [r_-, r_+] \subset (0, \infty) \quad \text{almost surely.}$$

Then the observed residual directions lie in a truncated cone strictly inside $\mathrm{cone}(w_1, w_2)$. Any two rays whose positive cone contains

$$\{w_1 + r w_2 : r \in [r_-, r_+]\}$$

can reconstruct every sample using at most two nonnegative coefficients. Thus exact two-sparsity and zero reconstruction error do not identify the teacher rays unless the observed cone exposes its boundary rays.

K.8 Summary

Sparsity matching fixes a cardinality scale. Frequency matching fixes marginal activation rates. Neither fixes the support-event law. In the oracle sparse-reconstruction objective, recovery is governed by thresholded eventwise geometry:

$$\mathbb{P}(S = T), \qquad \{\alpha_i : i \in T\}, \qquad \{\tilde{w}_i : i \in T\}.$$

Correct sparsity can therefore enable recovery in regimes where it coincides with rank-one or singleton-event dominance. Outside those regimes, the same matched sparsity can coexist with merge pressure, pair-cone compression, splitting, or direction/support mismatch.

L What can be certified by an observer?

If direction recovery does not imply support recovery, then decoder alignment alone cannot certify a trained SAE feature, especially in real models with no ground-truth supports. Any population certificate that uses only observed activations must be a functional of the observed activation law. The observable substitute is residual-direction geometry relative to a known observer-chosen subspace.

The observer diagnostics separate direction drift, poor cap precision, poor recall, intrinsic cap ambiguity, gate mismatch, and residual-path sensitivity.

L.1 Observable atoms, caps, and directional margins

Let $H \subseteq \mathbb{R}^d$ be a known observer-chosen subspace, for example the span of selected learned decoder directions or any other subspace specified without using latent supports. Since H is known, $P_{H^\perp}$ is computable, and $R_H := P_{H^\perp} x$ is observable from x. On $\{R_H \neq 0\}$, define $U_H := R_H/\|R_H\|$. The exact observable analogue of a rank-one event is the residual-direction atom $A_u(H) := \{R_H \neq 0,\ U_H = u\}$.

The observer cannot ask whether an SAE feature recovered the hidden latent event, because that event is not observed. The replacement question is geometric: is there an observed residual-direction atom or cap, and does sparse coding prefer directions near it? This is weaker than latent support recovery, but it is an audit target that depends only on observed activations.

Proposition L.1 (maximal observable rank-one event)**.** *Fix H and $u \in S^{d-1}$. If $A \in \sigma(R_H)$, $A \subseteq \{R_H \neq 0\}$, and $R_H = \xi u$ a.s. on A for some $\xi > 0$ on A, then $A \subseteq A_u(H)$ up to null sets. Conversely, $A_u(H)$ is itself a rank-one residual event whenever it has positive probability.*

Proof. If $A \subseteq \{R_H \neq 0\}$ and $R_H = \xi u$ with $\xi > 0$ on A, then $U_H = u$ almost surely on A, so $A \subseteq A_u(H)$ up to null sets. Conversely, if $\mathbb{P}(A_u(H)) > 0$, then on $A_u(H)$ we have

$$R_H = \|R_H\|\ u \qquad \text{a.s. on } A_u(H),$$

so $A_u(H)$ is a rank-one event for $\mathbb{R}_+ u$. □

Thus $A_u(H)$ is the maximal observable event on which the residual lies on the ray $\mathbb{R}_+ u$. Since exact atoms may have zero mass in the population, use positive-width caps

$$A_{u,\varepsilon}(H) := \{R_H \neq 0,\ \angle(U_H, u) \leq \varepsilon\}.$$

These caps are observable relaxations of rank-one residual events.

For $0 < \varepsilon < \phi \leq \pi$, define

$$M_{\mathrm{dir}}(u, \varepsilon, \phi, \lambda; H) := \underline{W}_\lambda(u, \varepsilon; H) - \overline{W}_\lambda(u, \varepsilon, \phi; H) - L(u, \varepsilon; H),$$

where

$$\underline{W}_\lambda(u, \varepsilon; H) := \frac{1}{2}\mathbb{E}\left[(\|R_H\| \cos\varepsilon - \lambda)_+^2 \mathbf{1}_{A_{u,\varepsilon}(H)}\right],$$

$$\overline{W}_\lambda(u, \varepsilon, \phi; H) := \frac{1}{2}\mathbb{E}\left[(\|R_H\| \cos(\phi - \varepsilon) - \lambda)_+^2 \mathbf{1}_{A_{u,\varepsilon}(H)}\right],$$

and

$$L(u, \varepsilon; H) := \frac{1}{2}\mathbb{E}\left[\|R_H\|^2 \mathbf{1}_{A_{u,\varepsilon}(H)^c}\right].$$

The margin M_{dir} is an observer-side version of singleton dominance. The first term lower-bounds how much the cap supports u. The second asks how much a direction at least ϕ away could still get from the same cap. The third gives every sample outside the cap to the competitor as a conservative budget.

Theorem L.2 (observer-only directional certificate)**.** *If $M_{\mathrm{dir}}(u, \varepsilon, \phi, \lambda; H) > 0$, then every global maximizer of $v \mapsto \mathcal{U}_\lambda(v; R_H)$ lies in the spherical ϕ-ball around u.*

Proof. On the cap $A_{u,\varepsilon}(H)$, one has

$$\langle u, R_H \rangle \geq \|R_H\| \cos \varepsilon,$$

which gives the lower bound $\underline{W}_\lambda(u, \varepsilon; H)$ for the utility of u. If $\angle(v, u) \geq \phi$, then on the same cap

$$\angle(v, U_H) \geq \phi - \varepsilon,$$

hence

$$\langle v, R_H \rangle \leq \|R_H\| \cos(\phi - \varepsilon),$$

which yields the upper bound $\overline{W}_\lambda(u, \varepsilon, \phi; H)$. On the complement, competitor utility is bounded above by $L(u, \varepsilon; H)$, while the utility of u remains nonnegative. Subtracting gives the result. □

L.2 Feature-level audit quantities

An observer can certify an objective-level statement without latent supports: the observed residual distribution's sparse-coding utility field prefers a direction near u. A trained feature r, with decoder d_r and activation a_r, can then be audited against the certified cap. Direction alignment is $D_{r,u} := \langle d_r, u \rangle$. Cap precision and recall are

$$\pi_{r,u} := \mathbb{P}(A_{u,\varepsilon}(H) \mid a_r > 0), \qquad \rho_{r,u} := \mathbb{P}(a_r > 0 \mid A_{u,\varepsilon}(H)).$$

Intrinsic cap ambiguity is

$$\kappa_{\text{in}}(u, \varepsilon; H) := \frac{\mathbb{E}[\|P_{u^\perp} R_H\|^2 \mathbf{1}_{A_{u,\varepsilon}(H)}]}{\mathbb{E}[\|R_H\|^2 \mathbf{1}_{A_{u,\varepsilon}(H)}]}.$$

The one-slot gate gap is $G_r := \mathbb{E}[(a_r - g_r^\star)^2] / (\mathbb{E}[(g_r^\star)^2] + \delta_0)$, where $g_r^\star(\nu) := (d_r^\top R_{-r}(\nu) - \lambda)_+$ and $\delta_0 > 0$ is a fixed stabilizer. Ratios are interpreted only when their denominators are positive.

These quantities separate observable failure modes. Low $D_{r,u}$ indicates direction drift. Low precision indicates that the feature fires off the certified cap, as in merging or contamination. Low recall indicates that the feature covers only part of the certified cap, as in splitting, partial coverage, or absorption-like false negatives. Large κ_{in} indicates that the cap itself is not close to rank-one, so even a perfect cap detector would have ambiguous direction semantics. Large G_r indicates that the learned activation fails to realize the sample-wise oracle gate for its decoder and leave-one-out residual.

Proposition L.3 (observable contamination floor)**.** *Let $A = A_{u,\varepsilon}(H)$, assume $\mathbb{E}[\|R_H\|^2 \mathbf{1}_A] > 0$, and suppose $A = G \sqcup C$, where the clean part satisfies*

$$\|P_{u^\perp} R_H\|^2 \leq \sin^2 \varepsilon_0 \, \|R_H\|^2 \qquad \text{on } G.$$

Define the energy fraction of the contaminating part by

$$\omega_C(u, \varepsilon; H) := \frac{\mathbb{E}[\|R_H\|^2 \mathbf{1}_C]}{\mathbb{E}[\|R_H\|^2 \mathbf{1}_A]}.$$

Then

$$\omega_C(u, \varepsilon; H) \geq \frac{\kappa_{\text{in}}(u, \varepsilon; H) - \sin^2 \varepsilon_0}{1 - \sin^2 \varepsilon_0}.$$

In particular, for an exact rank-one refinement $\varepsilon_0 = 0$,

$$\omega_C(u, \varepsilon; H) \geq \kappa_{\text{in}}(u, \varepsilon; H).$$

Proof. On the clean part G,

$$\|P_{u^\perp} R_H\|^2 \leq \sin^2 \varepsilon_0 \, \|R_H\|^2,$$

while on C,

$$\|P_{u^\perp} R_H\|^2 \leq \|R_H\|^2.$$

Taking expectations over $A_{u,\varepsilon}(H)$ gives

$$\mathbb{E}[\|P_{u^\perp} R_H\|^2 \mathbf{1}_A] \leq \sin^2 \varepsilon_0 \, \mathbb{E}[\|R_H\|^2 \mathbf{1}_G] + \mathbb{E}[\|R_H\|^2 \mathbf{1}_C].$$

Divide by $\mathbb{E}[\|R_H\|^2 \mathbf{1}_A]$ and rearrange. □

L.3 Recursive and feature-level refinements

The observer-only theory suggests a recursive recovery procedure. Start with the residualization subspace $H_0 = \{0\}$. At stage t, form the projected residual R_{H_t}, estimate the residual-direction law U_{H_t}, identify candidate caps $A_{u,\varepsilon}(H_t)$, certify them using M_{dir}, and add certified directions to H_{t+1}. This is the residualized analogue of separable NMF: one searches for anchor directions in projected residual space rather than anchor points in data space.

Proposition L.4 (recursive residual-direction recovery)**.** *Suppose there exist directions $u_1, \dots, u_m$ and scales $(\varepsilon_t, \phi_t, \lambda_t)$ such that for each stage t,*

$$M_{\mathrm{dir}}(u_t, \varepsilon_t, \phi_t, \lambda_t; H_{t-1}) > 0, \qquad H_{t-1} := \mathrm{span}\{u_1, \dots, u_{t-1}\}.$$

Then at every stage t, every global maximizer of the oracle utility field

$$v \mapsto \mathcal{U}_{\lambda_t}(v; R_{H_{t-1}})$$

lies in the ϕ_t-ball around u_t.

Proof. Apply Theorem L.2 at each stage t with residualization subspace H_{t-1}. The positivity of the directional margin implies that every global maximizer of the stage-t oracle field lies inside the corresponding ϕ_t-ball around u_t. □

Remark L.5 (global sparsity summaries are not support certificates)**.** There exist two disjoint witness families A and C with the same probability mass such that the gates

$$a = \mathbf{1}_A, \qquad b = \mathbf{1}_C$$

satisfy

$$\mathbb{E}[a] = \mathbb{E}[b], \qquad \mathrm{Var}(a) = \mathrm{Var}(b),$$

while recovering disjoint support families.

Proof. Take any two disjoint measurable sets A and C with $\mathbb{P}(A) = \mathbb{P}(C) > 0$. The Bernoulli gates $a = \mathbf{1}_A$ and $b = \mathbf{1}_C$ then have identical means and variances, but are supported on disjoint event families. □

For the feature-level angle bound, fix a target direction u, a target event A such as an observable residual-direction cap, and a trained activation $a_r \geq 0$. Define the in-event longitudinal signal and the in/out transverse loads by

$$\beta_{r,u}(B) := \mathbb{E}[\langle u, R_B \rangle \, a_r \mathbf{1}_A],$$

$$\eta^{\mathrm{in}}_{r,u}(B) := \mathbb{E}[\|P_{u^\perp} R_B\| \, a_r \mathbf{1}_A], \qquad \eta^{\mathrm{out}}_{r,u}(B) := \mathbb{E}[\|R_B\| \, a_r \mathbf{1}_{A^c}].$$

The bound below is meaningful when $\beta_{r,u}(B) > 0$ and the displayed denominator is positive.

Proposition L.6 (feature-level angle bound)**.** *Let*

$$\mu_r(B) := \mathbb{E}[R_B a_r].$$

Then

$$\|P_{u^\perp}\mu_r(B)\| \le \eta^{\mathrm{in}}_{r,u}(B) + \eta^{\mathrm{out}}_{r,u}(B),$$

and

$$\langle u, \mu_r(B)\rangle \ge \beta_{r,u}(B) - \eta^{\mathrm{out}}_{r,u}(B).$$

Hence, if

$$\beta_{r,u}(B) > \eta^{\mathrm{in}}_{r,u}(B) + 2\eta^{\mathrm{out}}_{r,u}(B),$$

then

$$\tan\angle(\mu_r(B), u) \le \frac{\eta^{\mathrm{in}}_{r,u}(B) + \eta^{\mathrm{out}}_{r,u}(B)}{\beta_{r,u}(B) - \eta^{\mathrm{out}}_{r,u}(B)}.$$

When the decoder is fit by least squares to the gate, the same bound controls the angle of d_r. At a stationary point of the normalized untied SAE, d_r is aligned with the gated residual moment from Proposition H.1, so the same signal-versus-contamination interpretation continues to apply.

Proof. Decompose

$$\mu_r(B) = \mathbb{E}[R_B a_r \mathbf{1}_A] + \mathbb{E}[R_B a_r \mathbf{1}_{A^c}].$$

The first term has longitudinal component at least $\beta_{r,u}(B)$ and transverse norm at most $\eta^{\mathrm{in}}_{r,u}(B)$. The second term has norm at most $\eta^{\mathrm{out}}_{r,u}(B)$. Therefore

$$\|P_{u^\perp}\mu_r(B)\| \le \eta^{\mathrm{in}}_{r,u}(B) + \eta^{\mathrm{out}}_{r,u}(B)$$

and

$$\langle u, \mu_r(B)\rangle \ge \beta_{r,u}(B) - \eta^{\mathrm{out}}_{r,u}(B).$$

Taking the ratio yields the tangent bound. □

Observable contamination-floor lower bound. For convenience one may define

$$\underline{\omega}(u, \varepsilon; \varepsilon_0, H) := \max\left\{0,\ \frac{\kappa_{\mathrm{in}}(u, \varepsilon; H) - \sin^2\varepsilon_0}{1 - \sin^2\varepsilon_0}\right\}.$$

This is the smallest in-cap contamination fraction consistent with an ε_0-clean refinement. The quantity is interpreted only when $\mathbb{E}[\|R_H\|^2 \mathbf{1}_{A_{u,\varepsilon}(H)}] > 0$.

Reporting protocol. For each certified residual-direction cap, report

$$q := \mathbb{P}(A_{u,\varepsilon}(H)), \qquad \kappa_{\mathrm{in}}, \qquad M_{\mathrm{dir}}, \qquad \underline{\omega}.$$

For each matched feature, report

$$D_{r,u}, \qquad \pi_{r,u}, \qquad \rho_{r,u}, \qquad \eta^{\mathrm{in}}_{r,u}, \qquad \eta^{\mathrm{out}}_{r,u}, \qquad G_r.$$

This cap-wise audit gives population evidence for faithful recovery, direction-only recovery, intrinsic cap ambiguity, splitting, merging, absorption, and amortization failure.

L.4 Failure-mode taxonomy from observer-only quantities

For completeness, we restate the observer-only taxonomy implied by the quantities above.

Faithful cap recovery. A feature matched to a certified cap has high direction score, high cap precision and recall, low cap contamination, and small gate gap.

Direction-only recovery. A feature has high decoder-cap alignment but poor cap precision or recall. The decoder aligns with the cap direction, but the gate does not fire exactly on the observable cap.

Intrinsic cap ambiguity. A cap has positive directional margin but large contamination floor. Poor support scores in this regime should not be overinterpreted as an SAE failure.

Merged features. A feature has no dominant matched cap or draws comparable gate-weighted signal from several certified caps. The feature is supported by mixed events rather than by a single rank-one event.

Splitting. Several features align with the same residual-direction cap, each with moderate recall, while their union covers the cap.

Absorption. A certified residual-direction cap has positive directional margin, but its residual utility drops sharply after subtracting learned slots. One natural observable score is

$$\mathrm{Absorb}(u; H) := \mathcal{U}_\lambda(u; R_H) - \mathcal{U}_\lambda\left(u;\, R_H - \sum_r d_r a_r\right).$$

Large positive absorption indicates that learned features suppress observable residual utility along u, consistent with absorption of the cap-supporting event.

Amortization failure. A cap is directionally certified, yet no matched feature has small gate gap.

Residual-path sensitivity. Small margins or comparable cap utilities flag possible sensitivity to perturbations; they are not by themselves an SGD seed-instability certificate.

M Synthetic two-stage SAE illustrations

This appendix gives small synthetic illustrations for the realizability and direction/support separations in the main text. The experiments are not used as evidence for an SGD theory, nor do they validate the population oracle path. They are sanity checks for three narrower points: a directly observed learned code is easier to recover than the original GT support law; high decoder alignment need not imply support recovery; and learned SAE supports can have a different joint support law from independent GT supports.

Data and architecture. We generate data from a positive latent-ray model with $d = 3$ observed dimensions and $F = 7$ ground-truth features. Supports are independent Bernoulli variables with probability $p = 0.2$, active magnitudes are independent $\mathrm{Unif}(0, 1)$, and teacher directions are random unit vectors in $\mathbb{R}^3$. Thus the expected GT sparsity is $Fp = 1.4$. Evaluation uses $50{,}000$ fresh examples.

Each SAE has an untied encoder and decoder. In the row-wise implementation,

$$z = \mathrm{ReLU}((x - b_{\mathrm{dec}})W_{\mathrm{enc}} + b_{\mathrm{enc}}), \qquad \hat{x} = zW_{\mathrm{dec}} + b_{\mathrm{dec}},$$

with normalized decoder rows. The objective is

$$\frac{1}{2}\,\mathbb{E}\|x - \hat{x}\|_2^2 + \lambda\,\mathbb{E}\|z\|_1.$$

At each training step, decoder gradients are projected to remove radial components, and decoder rows are renormalized to unit norm after the optimizer update.

Training protocol. All SAEs are trained with Adam using learning rate 3×10^{-3}, betas $(0.9, 0.999)$, $\epsilon = 10^{-8}$, no weight decay, and batch size 2048. SAE1 is trained for up to 4000 steps and SAE2 for up to 3000 steps. Validation uses a fixed batch of 8192 samples, evaluated every 500 steps. Early stopping is enabled after 2500 steps; we track EMA validation explained variance and EMA L_1 with decay 0.95, and stop if both plateau for 20 validation checks, with tolerances 10^{-4} for EV and 10^{-3} relative improvement for L_1. This early-stopping criterion did not materially affect the reported runs, which typically reached the maximum step budget.

For each trained SAE, λ is selected by a small binary search over $[10^{-5}, 10^{-1}]$, with up to two upper-bound expansions and four geometric midpoint rounds. Each candidate trains a fresh SAE. The

Table 2: Aggregate recovery in the two-stage SAE experiments. Code-student variants train SAE2 on SAE1 activations; output-student variants train SAE2 on SAE1 reconstructed hidden states. Direction-matched support F1 is the main metric for testing whether direction recovery carries the intended activation pattern.

Experiment	Pair	Direction	Direction-matched F1	Support-matched F1	Support-matched direction
Baseline code-student	GT → SAE1	0.953	0.583	0.605	0.952
Baseline code-student	GT → SAE2	0.971	0.600	0.600	0.971
Baseline code-student	SAE1 → SAE2	0.993	0.937	0.937	0.993
Output-student	GT → SAE1	0.953	0.583	0.605	0.952
Output-student	GT → SAE2	0.945	0.514	0.538	0.911
Output-student	SAE1 → SAE2	0.956	0.804	0.804	0.956
Fixed-decoder code-student	GT → SAE1	1.000	0.571	0.577	0.994
Fixed-decoder code-student	GT → SAE2	1.000	0.569	0.575	0.994
Fixed-decoder code-student	SAE1 → SAE2	1.000	0.995	0.995	1.000
Fixed-decoder output-student	GT → SAE1	1.000	0.571	0.577	0.994
Fixed-decoder output-student	GT → SAE2	1.000	0.391	0.518	0.940
Fixed-decoder output-student	SAE1 → SAE2	1.000	0.592	0.711	0.935
Random teacher	GT → SAE1	0.933	0.490	0.557	0.618
Random teacher	GT → SAE2	0.899	0.421	0.518	0.663
Random teacher	SAE1 → SAE2	0.948	0.783	0.818	0.808

selected candidate is the one whose final validation L_0 is closest to the target, breaking ties by higher explained variance. SAE1 targets the GT expected sparsity Fp, while SAE2 targets the measured teacher L_0. Default seeds are fixed: toy distribution seed 0, GT dictionary seed 1, SAE1 seed 2, SAE2 seed 20, and evaluation seed 4.

Variants. We compare five variants. In the *baseline code-student* variant, SAE1 is trained on GT hidden states and SAE2 is trained directly on SAE1 activations $z_1(x) = \mathrm{SAE1.encode}(x)$. The second-stage target is therefore a nonnegative coordinate code. In the *output-student* variant, SAE2 is trained on SAE1 reconstructed hidden states

$$\hat{x}_1(x) = \mathrm{SAE1.decode}(\mathrm{SAE1.encode}(x)),$$

so the teacher code is not given directly. In the *fixed-decoder code-student* variant, SAE1's decoder is fixed to the GT directions, SAE2 is trained on SAE1 activations, and SAE2's decoder is fixed to the identity in SAE1-code space; after composition through SAE1, SAE2's hidden-space directions are GT-aligned. In the *fixed-decoder output-student* variant, SAE1's decoder is fixed to the GT directions, SAE2 is trained on SAE1 reconstructed hidden states, and SAE2's decoder is fixed to the same GT directions in hidden space. In the *random-teacher* variant, SAE1 is left at random initialization and SAE2 is trained on the resulting random teacher activations.

Metrics. For a source representation A and target representation B, $A \to B$ means that each source feature of A is matched to its best counterpart in B. Direction is the mean best raw cosine between source and target decoder directions. Direction-matched F1 first matches by closest decoder direction and then computes support F1. Support-matched F1 instead matches by activation support. Unless stated otherwise, support F1 in the text means direction-matched F1, since this directly tests whether direction recovery carries the intended activation pattern.

All support comparisons use the same underlying evaluation examples. GT supports come from the latent generator. SAE1 supports are computed from $z_1(x)$. Code-student SAE2 supports are computed from SAE2 applied to $z_1(x)$. Output-student SAE2 supports are computed from SAE2 applied to $\hat{x}_1(x)$. For SAE1→SAE2 in code-student experiments, directions are compared in SAE1-code space. For GT→SAE2 in code-student experiments, SAE2 decoder directions are composed back into hidden space through SAE1's decoder: if D_1 is the SAE1 decoder in hidden space and D_2 is the SAE2 decoder in SAE1-code space, the composed hidden-space directions are $D_1 D_2$.

The code-student baseline is a positive control. Since SAE2 is trained on SAE1's nonnegative code, the target support representation is directly coordinate-ReLU representable. SAE2 recovers SAE1's learned code much more faithfully than either SAE recovers the original GT supports:

$$\mathrm{F1}(\mathrm{SAE1} \to \mathrm{SAE2}) = 0.937, \qquad \mathrm{F1}(\mathrm{GT} \to \mathrm{SAE1}) = 0.583, \qquad \mathrm{F1}(\mathrm{GT} \to \mathrm{SAE2}) = 0.600.$$

Table 3: Training and support-correlation diagnostics. "Selection L_0" is the final validation L_0 used by the L_1-selection procedure; "Eval L_0" is measured on the $50{,}000$-sample evaluation set used for support correlations. GT supports are nearly independent on the evaluation set, while learned SAE supports have much larger absolute off-diagonal correlations. Approximate L_0 targeting does not determine the support-event law.

Experiment	Representation	L_1	Target L_0	Selection L_0	Eval L_0	EV	Mean abs. off-diag. corr.	Max abs. off-diag. corr.
Baseline code-student	GT	–	–	–	1.395	–	0.003	0.017
Baseline code-student	SAE1	0.237	1.400	1.266	1.260	0.843	0.205	0.561
Baseline code-student	SAE2	0.018	1.273	1.117	1.115	0.995	0.184	0.454
Output-student	GT	–	–	–	1.395	–	0.003	0.017
Output-student	SAE1	0.237	1.400	1.266	1.260	0.843	0.205	0.561
Output-student	SAE2	0.100	1.273	1.039	1.025	0.960	0.161	0.581
Fixed-decoder code-student	GT	–	–	–	1.395	–	0.003	0.017
Fixed-decoder code-student	SAE1	0.487	1.400	1.170	1.158	0.624	0.212	0.561
Fixed-decoder code-student	SAE2	10^{-5}	1.175	1.179	1.166	1.000	0.210	0.561
Fixed-decoder output-student	GT	–	–	–	1.395	–	0.003	0.017
Fixed-decoder output-student	SAE1	0.487	1.400	1.170	1.158	0.624	0.212	0.561
Fixed-decoder output-student	SAE2	0.056	1.175	1.153	1.144	0.968	0.147	0.756
Random teacher	GT	–	–	–	1.395	–	0.003	0.017
Random teacher	SAE1	–	–	3.473	3.468	–	0.386	0.816
Random teacher	SAE2	0.003	3.473	3.303	3.303	0.999	0.249	0.646

The fixed-decoder code-student variant reaches nearly exact recovery, with

$$\mathrm{F1}(\mathrm{SAE1} \to \mathrm{SAE2}) = 0.995.$$

This confirms that the metric and training setup can detect near-perfect recovery when the target code is directly presented.

The output-student variants are harder because SAE2 is trained on $\hat{x}_1(x)$, not on $z_1(x)$. With learned decoders, recovery of the SAE1-induced representation is weaker than in the direct code-student case but remains higher than GT recovery:

$$\mathrm{F1}(\mathrm{SAE1} \to \mathrm{SAE2}) = 0.804, \qquad \mathrm{F1}(\mathrm{GT} \to \mathrm{SAE2}) = 0.514.$$

The fixed-decoder output-student ablation is more stringent. Here both SAE1 and SAE2 have the correct GT decoder directions fixed in hidden space. Nevertheless,

$$\mathrm{F1}(\mathrm{GT} \to \mathrm{SAE2}) = 0.391, \qquad \mathrm{F1}(\mathrm{SAE1} \to \mathrm{SAE2}) = 0.592.$$

Thus even when decoder directions are correct by construction, the student does not recover the teacher or GT support events. The high support-matched directions show that the best support matches are still geometrically nearby, but the corresponding support overlaps remain far from exact. Direct code recovery and hidden-state recovery are different tasks.

The fixed-decoder ablations give the cleanest direction/support separation. Decoder directions are correct by construction, but GT support recovery remains poor. In the code-student fixed-decoder variant,

$$\mathrm{direction}(\mathrm{GT} \to \mathrm{SAE2}) = 1.000, \qquad \mathrm{F1}(\mathrm{GT} \to \mathrm{SAE2}) = 0.569.$$

In the output-student fixed-decoder variant, the separation is even sharper:

$$\mathrm{direction}(\mathrm{GT} \to \mathrm{SAE2}) = 1.000, \qquad \mathrm{F1}(\mathrm{GT} \to \mathrm{SAE2}) = 0.391.$$

Thus correct decoder directions do not certify recovery of the intended activation pattern: the learned gates remain geometrically related to the correct rays, but their support events differ substantially.

The fixed-decoder output-student ablation also separates approximate sparsity targeting from support recovery. SAE2's final L_0 is 1.153, close to its target 1.175, and its explained variance is 0.968. Yet its GT support F1 is only 0.391. Thus a model can have correct decoder directions, approximately matched sparsity, and high reconstruction quality while still using the wrong support-event law.

The correlation check shows the same point at the level of joint support statistics. GT supports are independent to finite-sample precision, with mean absolute off-diagonal support correlation 0.0034 and maximum 0.0168. Learned supports are substantially correlated. In the baseline, SAE1 has mean 0.2048 and maximum 0.5607, while SAE2 has mean 0.1838 and maximum 0.4544. In

the fixed-decoder output-student variant, SAE2 has mean 0.1473 and maximum 0.7560. Thus the mismatch is not merely a direction-matching artifact; learned SAE codes can have different joint support statistics from the GT latent code even when decoder directions are fixed.

These experiments are intentionally limited. They use one synthetic geometry and one evaluation seed, and they do not include confidence intervals. The formal claims in the paper are the population statements and counterexamples proved in the main text and preceding appendices.

N LLM Usage

Large language models were used for editing, wording, and non-authoritative review-style feedback on presentation and claim calibration. All theoretical claims, proofs, experiments, and final text were checked and are the responsibility of the authors.

O Claim ledger

Table 4 summarizes results and perspective.

Table 4: Claim ledger and relation to prior work. The formal results concern population one-slot or residual-greedy sparse reconstruction, transfer to one-row SAE realizability, and the separation between decoder directions and latent support semantics.

Topic	What we prove	What we interpret	Connection to prior work
Population objective	For a fixed residual law and unit direction v, the optimal nonnegative one-slot code is $(\langle v, R\rangle - \lambda)_+$, with gain $U_\lambda(v; R) = \frac{1}{2}\mathbb{E}[(\langle v, R\rangle - \lambda)_+^2]$. For positive latent-ray residuals, this gain decomposes exactly over residual support events.	The one-slot sparse-reconstruction objective scores thresholded residual mass, not teacher-feature identity directly. Singleton, mixed, and broader support events contribute separately.	A population, free-direction analogue of thresholded matching pursuit, complementary to sparse recovery and dictionary-learning guarantees under incoherence, separability, or sparse-use assumptions. [22, 30, 14, 27, 2]
Teacher recovery	If the utility lost by rotating away from a singleton rank-one teacher event exceeds a conservative off-singleton competition budget, then all one-slot global maximizers lie near the residualized teacher direction.	This certifies recovery when clean singleton events dominate off-singleton contamination above threshold. Failure of the certificate is not claimed to imply failure of recovery.	Explains why restrictive sparsity, frequency, geometry, or hyperparameter alignment assumptions in SAE recovery theory help: they make teacher-identifying events dominate thresholded utility. [9, 6, 11, 7]
Merging and hedging	We give local and explicit population certificates under which rotating from a teacher direction toward another residual direction increases one-slot utility. In two-feature co-occurrence models, a mixed direction can beat individual teachers once joint events cross threshold.	Merging can arise from thresholded mixed-event mass before finite samples, amortized encoders, or optimization enter. Correlation is one source; the formal driver is above-threshold transverse residual mass.	Feature hedging and non-canonical-feature work show that trained SAEs can learn mixed latents. We isolate an oracle-level eventwise mechanism that rewards such directions. [5, 20]
Splitting	When a broad event family has heterogeneous residual directions, event-partition certificates show that a multi-direction specialized representation can obtain larger residual-greedy utility than forcing the family through one shared direction. We also separate this from encoder-induced splitting.	Splitting can reflect heterogeneous residual evidence, not only semantic oversegmentation or training instability. A semantic feature may arrive through several residual rays.	Connects empirical splitting and matching-pursuit-style residual decompositions to a population utility comparison. [8**?** , 10]
Absorption and residual paths	Sequential residual subtraction can suppress above-threshold evidence for a later teacher direction on part of its support. We also construct exact tied population residual paths where early choices lead to inequivalent later choices.	Absorption-like holes and residual-path alternatives can be substrates of the population residual-greedy objective. We do not claim a convergence theorem for SGD or a full account of empirical seed sensitivity.	Complements absorption, seed-instability, and SDL non-identifiability studies by isolating residual suppression and tie-breaking mechanisms. [8, 26, 28]
Encoder realizability	For a fixed decoder direction and leave-one-out residual, one encoder row performs squared-error regression onto the oracle target $d^\top R - \lambda$ with a nonnegative affine-ReLU output. The one-row value matches the oracle iff the oracle gate lies in the L^2-closure of affine-ReLU gates; exact realization is the attained case.	Oracle preference and SAE realizability are separate layers. A direction may be oracle-preferred while a one-ReLU row cannot realize its gate exactly.	Identifies a representability gap between unrestricted sparse-coding gates and one-row ReLU SAE encoders. [3, 12, 9, 11]
Direction vs. support	Decoder direction alignment is a gated residual-moment condition. Under residualized linear independence, exact alignment with a teacher ray requires zero gated co-activation with all other residualized atoms. Counterexamples show that direction and support recovery do not imply each other.	A recovered decoder direction alone does not certify that the latent fires on the intended support. Directional semantics and activation-pattern semantics are distinct.	Refines latent-recovery work showing that reconstruction or decoder recovery need not identify latent variables. [28, 31]
Sparsity, frequency, diagnostics	Exact support recovery implies sample-wise sparsity and expected sparsity matching, but converses fail. Without latent supports, we prove observer-only residual-cap certificates and define direction, precision, recall, ambiguity, and gate-consistency diagnostics.	Sparsity and firing frequency are useful shadows of support recovery, but the objective depends on joint support events, amplitudes, thresholds, and geometry. Observer diagnostics certify residual geometry, not latent support semantics.	Explains what sparsity/frequency alignment captures in toy recovery studies and what it misses; gives population diagnostics for settings without ground-truth supports. [6, 9, 11, 3, 29]

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state theoretical claims about the population sparse-reconstruction objective, support-event decomposition, dominance regimes for recovery, one-ReLU realizability, and direction/support separation. The paper does not claim empirical validation or a full theory of trained SAE optimization dynamics.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses structural limitations including the positive latent-ray model, nonnegative coefficients, population rather than finite-sample analysis, fixed-residual or residual-greedy scope, non-vacuous residual charts, absence of an SGD convergence theory, lack of empirical validation on pretrained-model SAEs, and the limited illustrative role of the synthetic experiments.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper states the modeling assumptions, including finite-moment assumptions, the positive latent-ray model, nonnegative coefficients, residual-chart conditions, and one-ReLU encoder restrictions where relevant. The theoretical results are accompanied by theorem statements, separation constructions, and complete proofs in the appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The synthetic experiments are illustrative rather than central to the main theoretical claims. The appendix specifies the data-generating process, SAE architecture, objective, optimizer, training steps, validation protocol, L_1 selection procedure, seeds, evaluation set size, matching metrics, and aggregate results needed to reproduce the two-stage SAE experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The paper does not currently provide an anonymized code release. The experiments use synthetic data generated from fully specified distributions, and the appendix provides the information needed to reimplement them.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The appendix gives the experimental setting: synthetic generator, dimensionality, number of features, support and magnitude distributions, evaluation sample size, SAE architecture, reconstruction and L_1 objective, Adam hyperparameters, training schedule, validation protocol, decoder normalization, L_1 binary search, seeds, and matching metrics.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The reported experiments are small synthetic illustrations and do not include formal confidence intervals or error bars.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: The experiments are lightweight synthetic SAE runs, but the current manuscript does not fully specify wall-clock time, hardware type, memory, or total compute. Since the experiments are illustrative and not the main contribution, this omission does not affect the reproducibility of the theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: The work is theoretical and uses only synthetic toy experiments. It does not involve human subjects, private data, scraped datasets, deployed systems, or release of high-risk models.

Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper is foundational work on interpretability and sparse representations. Its positive impact is improved understanding of when SAE features should or should not be interpreted as ground-truth features; a potential negative impact is misplaced confidence in interpretability tools if their structural limitations are ignored.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: The paper does not release pretrained models, image generators, scraped datasets, or other high-risk assets.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not rely on external datasets, pretrained models, or proprietary assets. Existing scholarly work is credited through citations.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not introduce a new dataset, benchmark, pretrained model, or reusable software package as a contribution.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not involve crowdsourcing, human-subject studies, user studies, or participant compensation.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not involve human subjects, crowdsourcing, or participant data, so IRB or equivalent approval is not applicable.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The paper declares LLM usage for editing, wording, and non-authoritative review-style feedback.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.