WHEN DO SPARSE AUTOENCODERS RECOVER FEATURES, AND WHAT DO THEY LEARN INSTEAD?

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) reconstruct activations as nonnegative sums of learned directions, and each latent's activation is read as the strength of one feature of the data. Training sees only reconstruction and sparsity, never the features, and learned latents are known to split, absorb and merge features and to change between runs. One explanation is limited inference: a one-layer encoder can provably fall short of the best sparse code chosen separately for each input, and more expressive encoders have been reported to recover such codes better and to yield more interpretable latents. We remove that limitation, building synthetic data whose features are identifiable and whose strengths a one-layer ReLU encoder computes exactly, and vary only how the strengths of co-occurring features depend on each other while directions, frequencies, marginal strength laws and covariance stay fixed. For independent exponential strengths, every sufficiently near-optimal affine-ReLU SAE of arbitrary finite width recovers each feature's additive contribution through fixed groups of latents as the L_1 penalty vanishes, whereas every such near-optimum retains nonvanishing contribution error when co-occurring strengths are often nearly equal. Adding a second encoder layer makes a cheaper non-recovering code available even under independence, while penalizing the number of active latents with one-layer responses restores individual recovery as its price vanishes. On a 16-dimensional synthetic source, L_1 SAEs trained on the coupled data reach a lower loss than any code on the generating directions. Recovery failures therefore need not come from limited inference, and when inference is already exact, a more expressive encoder can cause them.

1 INTRODUCTION

Sparse autoencoders (SAEs) are used to identify variables represented in neural-network hidden states, on the hypothesis that some variables are carried along fixed directions whose strengths vary between inputs (Elhage et al., 2022; Park et al., 2024). An SAE learns candidate directions and a nonnegative activation for each, then reconstructs the hidden state by summing their products (Huben et al., 2024; Bricken et al., 2023). Its latents are read from the examples that activate them and are used to intervene by changing those activations (Templeton et al., 2024; Marks et al., 2025), so both uses take a latent to stand for a feature of the data and its decoded term to be that feature's contribution to the hidden state.

Training never sees an individual feature. The loss rewards reconstruction and penalizes the code, so an SAE can reconstruct almost perfectly while its latents split, mix or reassign features, and interpretations and interventions built on those latents inherit the error. Evaluations bear this out. Simple baselines outperform SAE latents for model control (Wu et al., 2025), although modern SAEs can be competitive with supervised features on one circuit task (Makelov et al., 2025); sparse probes do not consistently beat simple baseline ensembles (Kantamneni et al., 2025); and improvements on proxy metrics do not reliably transfer to practical evaluations (Karvonen et al., 2025; Korznikov et al., 2026).

Empirical studies show that SAE decompositions are not canonical. Latents found at one width decompose further at another (Leask et al., 2025), and they split and absorb, so a general latent misses cases a more specific latent has taken over (Chanin et al., 2025b). Correlated features make narrow SAEs merge (Chanin et al., 2025a), the chosen sparsity level changes which features appear (Chanin

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

& Garriga-Alonso, 2026a), and two runs on the same data learn different features (Paulo & Belrose, 2026; Gerasimov et al., 2026). Many observations establish pathologies without identifying what caused them; in particular, calling a latent absorbed or merged requires knowing which contributions the generator treats as separate.

Recovery studies supply this by fixing a generator whose features are known and asking when the learned dictionary and code recover it. They give recovery conditions and procedures (Chen et al., 2026; Nelson et al., 2026; Zhu et al., 2026), characterize when sparse objectives admit non-generating or polysemantic optima (Dorrell, 2026; Mencattini et al., 2026; Cui et al., 2026), and provide benchmarks with known generators (Chanin & Garriga-Alonso, 2026b). Classical dictionary learning gives analogous conditions (Spielman et al., 2012; Arora et al., 2015), and Wang et al. (2020, Section 3.2) show that a rotated dictionary can have lower L_1 cost than the generating one. A recovery failure can therefore mean that training missed the generating solution or that the objective prefers another decomposition, and a comparison of two dictionaries leaves open whether a wider SAE could keep the generating directions alongside a cheaper code.

Recovery can also fail before the objective ranks decompositions: the encoder may be unable to compute the generating strengths. An SAE encoder applies one affine map followed by a gate, creating a provable gap to per-input sparse coding (O'Neill et al., 2025). O'Neill et al. (2025) also report higher synthetic-code correlation and automated-interpretability scores for multilayer encoders, making limited inference a natural explanation to test. If the strengths lie outside the response class, failure cannot distinguish objective preference from inaccessible inference.

Our contributions are threefold.

- A framing that makes recovery failures attributable. We define recovery as supplying each feature's decoded contribution on every input, possibly through a fixed group of latents, and separate the four conditions it requires: the data identify the features, the encoder can compute their strengths, the objective prefers them, and training reaches that preference. Our sources satisfy the first two by construction, so any failure we observe belongs to the objective or to training (Sections 2–3).
- Class-wide results for the objective. Any exact code that uses less activation than the generating code excludes recovery for every sufficiently near-optimal SAE in its class, at any width and with free biases (Lemma 1). A certificate proves recovery for independent exponential strengths (Theorem 2). With these tools we show that the dependence between co-occurring strengths, the sign structure of features, the encoder depth and the sparsity penalty each change whether recovery holds, while the true strengths remain identifiable and computable (Section 3).
- Measurements that separate objective preference from training. Certified bounds on coding at fixed decoders show that coupled L1 SAEs on a 16-factor source attain a lower objective than any code on the generating directions, and continued training under a switched strength law moves the same models as the objective predicts. The same bound applies to public GPT-2 SAEs without labels, and reduction bounds show that strengths can remain readable when their contributions cannot be removed separately (Sections 4–5).

2 What recovery means, and what the loss rewards

2.1 What a recovered contribution must supply

We construct hidden states from known feature directions and strengths:

$$X = \sum_{i=1}^{F} \alpha_i w_i, \qquad \alpha_i \geq 0, \quad \|w_i\| = 1. \tag{1}$$

Feature i is present when $\alpha_i > 0$ and contributes the vector $\alpha_i w_i$. The SAE is trained on the summed hidden states through the reconstruction and sparsity objective without strength labels, the generating quantities are retained only to evaluate the frozen model, to select hyperparameters on separate independent-strength populations, and to build ablation comparisons. Orthogonality and independence are assumed only where a particular result requires them. The SAE reconstructs

$$f(X) = b + \sum_{j=1}^{K} d_j g_j(X), \qquad \|d_j\| = 1, \quad g_j(X) \geq 0, \tag{2}$$

where b is one shared bias. Width K counts available latents, whereas activity counts the outputs that are positive on a particular input.

A contribution comparison tests the learned direction and activation together. For unit directions $\|d_j g_j - \alpha_i w_i\|^2 = (g_j - \alpha_i)^2 + 2g_j\alpha_i(1 - d_j^\top w_i)$, so $d_j^\top w_i = 1$ makes the second term vanish and leaves the first untouched: a correct direction leaves the whole coefficient error unchecked. Because the comparison is made input by input, it also depends on the joint law of the strengths: coefficients correct on every input are correct jointly, whereas matching each activation's distribution separately says nothing about which input received which value.

Even exact reconstruction can assign those vectors differently. For perpendicular unit u, v, take $X = Z_1 u + Z_2 v$ with $Z_i \in \{0, 1\}$. Writing $(t)_+ = \max(t, 0)$, use decoder directions $u, v, (u + v)/\sqrt{2}$ and responses

$$g_{10} = (Z_1 - Z_2)_+, \qquad g_{01} = (Z_2 - Z_1)_+, \qquad g_{11} = \sqrt{2}(Z_1 + Z_2 - 1)_+. \tag{3}$$

Since $Z_1 = u^\top X$ and $Z_2 = v^\top X$, each response is an affine function of X followed by ReLU, and the code reconstructs every input exactly. Figure 1(d) shows its non-generating allocation, and Section 5 shows that both strengths remain readable.

Scoring a trained model. A faithful SAE need not use one latent per feature. Wide SAEs split features across several latents, for instance latents that share a feature's direction but fire on different subsets of its occurrences (Bricken et al., 2023), and the objective cannot distinguish such splits: parallel latents that together carry the same activation leave reconstruction and L_1 cost unchanged (see Appendix C.3). We therefore credit each feature with a fixed group of latents, reused on every input, which is the case $\Omega_{ij} \in \{0, 1\}$. Fractions $\Omega_{ij} \geq 0$ with $\sum_i \Omega_{ij} \leq 1$ also cover parallel copies that each carry a fixed share, and allowing a latent to be shared between features only makes the score more lenient. Writing $v_i(X) = \sum_j \Omega_{ij} d_j g_j(X)$, we report contribution normalized mean squared error (NMSE)

$$\widehat{\mathcal{R}_{\text{frac}}} = \frac{\sum_i \mathbb{E}\|\alpha_i w_i - v_i(X)\|^2}{\mathbb{E}\sum_i \alpha_i^2}. \tag{4}$$

Setting $\Omega = 0$ gives NMSE one for any angles between the generating directions: its denominator sums individual contribution energies, which cannot cancel as they can in the input norm. The column constraint keeps the shares assigned from one latent from adding up to more than its decoded vector. Each assignment also defines an edit: multiplying g_j by $1 + (s - 1)\Omega_{ij}$ changes the reconstruction by $(s - 1)v_i$, so contribution error measures the error of this change relative to scaling the intended feature (Appendix C.3). We use this contribution error throughout. Theorem 2 controls it in mean absolute error, the norm its certificate bounds, while experiments report the squared NMSE (4), which is stricter because small squared error implies small absolute error. The obstruction holds in both norms (Appendix E.5), and Appendix E.1 states what squared-error recovery would additionally require.

2.2 What the loss pays for a response

The objective we analyze first is

$$\mathcal{L}(D, b, g) = \mathbb{E}\left[\tfrac{1}{2}\|X - b - Dg(X)\|^2 + \lambda \sum_j g_j(X)\right], \qquad \lambda > 0. \tag{5}$$

Because activations are nonnegative, the second term is the L_1 norm of the code. We first use untied affine-ReLU encoders, $g_j(X) = (e_j^\top X + a_j)_+$, whose weights need not equal the decoder directions. We fix decoder norms to one because replacing d_j by cd_j and g_j by g_j/c for $c > 1$ would otherwise lower the penalty without changing any reconstruction. To isolate the role of activation magnitude, Section 3.3 replaces the L_1 penalty with a count of active latents.

At a fixed dictionary D and bias b, classical sparse coding optimizes the activations separately for each input (Lee et al., 2006). An SAE instead produces all codes with a single encoder from a response class $\mathcal{G}$, amortizing inference across inputs (Ranzato et al., 2006; Gregor & LeCun, 2010). Its objective therefore cannot fall below the per-input optimum. We call the difference the encoder gap $A_{\mathcal{G}}(D, b) \geq 0$; it vanishes when the per-input optimum is realizable by $\mathcal{G}$, while training may add further optimization error (Appendix D.1).

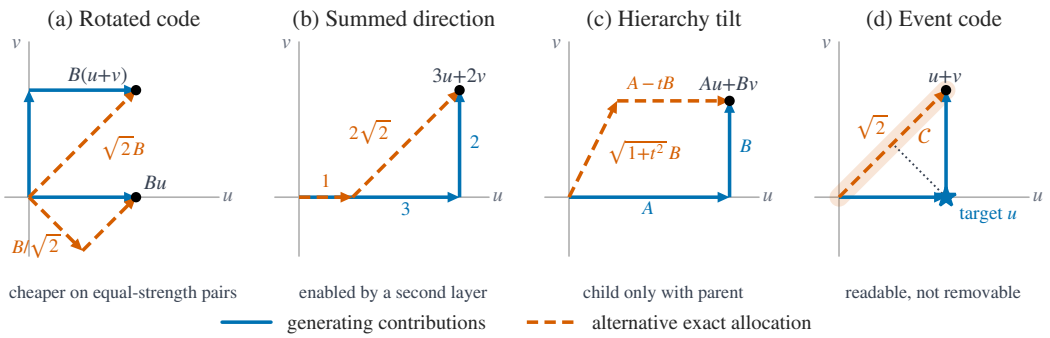


Figure 1: **Alternative exact decompositions of the same input.** Blue paths are generating contributions; orange dashes are exact alternatives. (a) Rotation spends $\sqrt{2}B$ instead of B on a singleton, but $\sqrt{2}B$ instead of $2B$ on an equal pair. (b) On $3u + 2v$, a summed-direction latent lowers activation from 5 to $1 + 2\sqrt{2}$; it needs $m = \min(a_u, a_v)$, which one affine-ReLU layer cannot compute. (c) The hierarchy construction gains at first order while normalization costs $O(t^2 B)$; the full code uses $-u$ when $A < tB$. (d) Reading and removal differ: $Z_1 = g_{10} + g_{11}/\sqrt{2}$, but reduction-only edits stay in $\mathcal{C}$ and cannot isolate u on the pair.

For a fixed source and response class, the gap can differ between dictionaries: an encoder may code optimally at one dictionary and not at another. Changing the encoder can therefore change which dictionary is cheapest under joint training. We show this in Section 3.2.

3 What determines contribution recovery?

Which exact allocation in Figure 1 is cheaper depends on which inputs occur, what the encoder can compute, and what the penalty charges; the following subsections vary each in turn.

3.1 How co-occurring strengths move together

We begin with a population whose generating strengths an affine-ReLU encoder can compute exactly. Let u_1, u_2 be orthonormal and draw

$$X = x_1 u_1 + x_2 u_2, \qquad x_i = Z_i \epsilon_i B_i. \tag{6}$$

Here $Z_i \in \{0, 1\}$ records occurrence, $B_i > 0$ is magnitude, and the balanced signs $\epsilon_i \in \{-1, 1\}$ are mutually independent and independent of occurrences and magnitudes. Each signed factor is represented by two nonnegative features, with contributions $(x_i)_+ u_i$ and $(-x_i)_+ (-u_i)$, and the four responses $(\pm u_i^\top X)_+$ recover their strengths exactly. The orthogonal coordinates make the strengths directly readable, and the balanced signs let us change how magnitudes depend on one another while holding the input covariance fixed.

Inputs containing a single factor (singletons) put positive probability on that factor's line, which distinguishes the generating directions from mixtures within the source family of Appendix C.1; Appendix C.4 shows that the same geometry can be learned. Our experiments use $p = \Pr(Z_i = 1) = 1/4$ and pair rate $c = \Pr(Z_2 = 1 \mid Z_1 = 1) = 3/4$, so each factor still occurs alone with probability $p(1 - c) > 0$, as both the identification argument and the recovery proof require. Pairs are nonetheless frequent enough for a shared direction to save activation; since $s_B \le 1$, the rotated code introduced below can win only when $c > 2 - \sqrt{2}$. Magnitudes are exponential with mean $\mu = 1.5$, because an exponential magnitude with a balanced sign is a Laplace variable, the simplest case the certificate of Appendix E.1 covers (Appendix E.2); μ only sets the units.

We compare independent paired magnitudes with a mixture that uses the nearby-quantile coupling of Appendix E.8 with probability η and independent draws otherwise. Under the coupling, paired magnitudes lie at nearby quantiles of their common law, so co-occurring strengths are often nearly equal. We use $\eta = .95$, which preserves each marginal and the occurrence process while balanced signs keep cross-coordinate covariance zero. The two laws assign positive probability to the same

input events, so a code reconstructs one exactly if and only if it reconstructs the other, although its expected activation cost changes. Trained SAEs already differ: at width eight and $\lambda = .1$, median contribution NMSE over 32 starts is .00223 under independence and .464 under coupling, compared with .00213 for the faithful reference, which soft-thresholds the true strengths on the generating directions. Coupled decoders leave the generating axes for their sums and differences (Appendix G.6, Figure 5). Related synthetic benchmarks use correlated firing and parent-scaled hierarchical magnitudes (Chanin & Garriga-Alonso, 2026b), while our matched comparison isolates the joint magnitude law.

The competing code. The rotated code replaces the signed generating axes by $\pm(u_1 + u_2)/\sqrt{2}$ and $\pm(u_1 - u_2)/\sqrt{2}$. Taking the positive part of the projection onto each direction gives a fixed affine-ReLU encoder that reconstructs every input in (6) exactly (Figure 1(a)).

Write $M = \mathbb{E}\sum_j g_j(X)$ for the expected sum of a code's activations, the quantity the penalty in (5) charges, and let $s_B = \mathbb{E}[\min(B_1, B_2) \mid Z_1 = Z_2 = 1]/\mu$. Averaging over singleton and pair inputs gives (Appendix E.4)

$$\frac{M_{\mathrm{rot}}}{M_{\mathrm{gen}}} = \sqrt{2}\left(1 - \frac{c\, s_B}{2}\right), \qquad M_{\mathrm{gen}} = 2p\mu. \tag{7}$$

The rotated code is cheaper exactly when $c\, s_B > 2 - \sqrt{2}$, a condition on the product rather than on either factor alone. Write $\kappa = c\, s_B$ for the share of a feature's expected strength held in common with a partner. Here c is the probability that the feature is paired, and s_B is the fraction of its magnitude that the partner matches when it is. Figure 2(a) shows where rotation saves or costs activation on each pair input, and Figure 2(b) shows the population average as κ varies. Independence gives $s_B = 1/2$, so the generating code is cheaper, whereas the coupling above gives $s_B = .9236$, so the rotated code is cheaper. Appendices F.6 and G.8 sweep the pair rate and the penalty.

A cheaper code leaves open whether a wider SAE could keep the generating directions alongside it and still recover their contributions. The next lemma rules this out for every representation in the class, including hybrids.

Lemma 1 (Activation budget)**.** *Let $f = b + \sum_j d_j g_j$ have unit decoder columns and nonnegative responses, write $M_{\mathrm{gen}} = \mathbb{E}\sum_i \alpha_i$, and let fixed fractions $\Omega_{ij} \geq 0$ with $\sum_i \Omega_{ij} \leq 1$ assign its latents to generating features. Then*

$$\sum_i \mathbb{E}\|\alpha_i w_i - v_i(X)\| \geq M_{\mathrm{gen}} - \mathbb{E}\sum_j g_j(X). \tag{8}$$

If a response class contains an exact code with expected activation $M_{\mathrm{alt}} < M_{\mathrm{gen}}$, every sequence within $o(\lambda)$ of that class's infimum satisfies $\liminf_{\lambda\to 0}\sum_i \mathbb{E}\|\alpha_i w_i - v_i(X)\| \geq M_{\mathrm{gen}} - M_{\mathrm{alt}}$.

Supplying $v_i(X)$ from unit directions costs at least $\|v_i(X)\|$ of activation, and the column budget prevents one activation from being counted twice, so the triangle inequality gives (8). A solution within δ of the infimum costs at most $\lambda M_{\mathrm{alt}} + \delta$ and hence uses at most $M_{\mathrm{alt}} + \delta/\lambda$ expected activation. Thus one feasible saving excludes recovery throughout the class, regardless of width or latent reassignment.

Theorem 2 (Joint strengths reverse contribution recovery)**.** *Consider* (6) *with exponential marginal magnitudes, occurrence rate $p \leq 1/2$, and pair rate $2 - \sqrt{2} < c < 1$. Under independent paired magnitudes, every affine-ReLU SAE sequence within $o(\lambda)$ of the class infimum over arbitrary finite widths and free biases recovers the four generating contributions in mean absolute error after grouping as $\lambda \to 0$. There are continuous coupled magnitude laws with the same occurrence probabilities, marginal magnitude laws, generating directions, and input covariance for which every such sequence has contribution error bounded away from zero after any disjoint grouping, and equally after any fractional allocation with $\sum_i \Omega_{ij} \leq 1$.*

The negative half applies Lemma 1 to the rotated code when $\kappa > 2 - \sqrt{2}$.

To prove recovery under independence, we bound every affine-ReLU response separately. Inputs where a factor occurs alone are reweighted to stand in for all occurrences of that factor; for independent exponential strengths, what any single response supplies along a factor under this weighting is at most its mean activation, with equality only on the signed generating axes. The bound holds response by response, so it needs no limit on width and forces near-optimal SAEs onto the axes; balanced signs

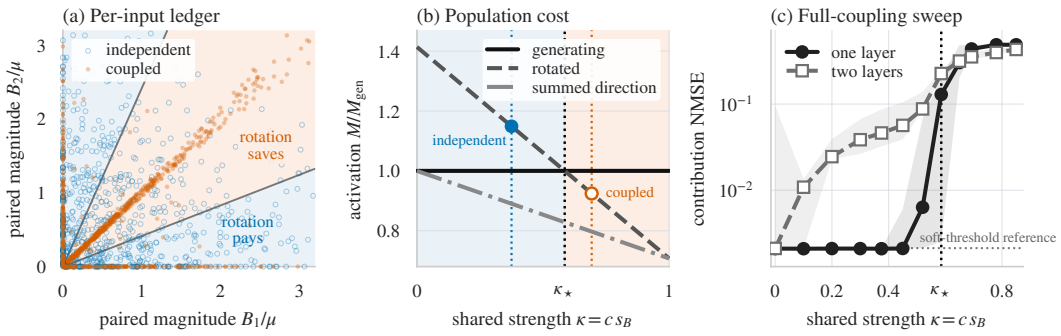


Figure 2: **The pair distribution decides which code is cheaper, and the encoder which code is available.** Panels (a)–(b) connect the input-level activation ledger to its population average: rotation becomes cheaper at $\kappa_\star = 2 - \sqrt{2}$, while the summed-direction code enabled by a second layer is cheaper for every $\kappa > 0$. (c) Median contribution NMSE over 16 width-eight starts at $\lambda = .1$, with interquartile bands and the soft-threshold reference. This full-coupling sweep varies the pair rate and is distinct from the reference laws in (b); Appendix F.6 gives the sweep setup and matched-κ comparisons.

and small reconstruction error then fix the bias and the strengths (Appendix E.1). Proposition E.1 extends the conclusion to finitely many orthonormal factors whose coordinates are independent Gaussians multiplied by positive scales. The scales may depend on one another arbitrarily, provided they are independent of occurrence and satisfy the occurrence conditions stated there. The required optimization accuracy is $o(\lambda)$ because competing exact codes differ by order λ.

Independence alone does not fix κ: Appendix E.3 gives finite-variance independent laws on both sides of the threshold, including two with the same tail exponent.

Without balanced signs, Figure 1(c)'s tilted code is cheaper for every hierarchy with finite mean child strength in which a child occurs only with its parent, even with independent strengths. The child latent carries part of the parent's contribution, as in feature absorption (Chanin et al., 2025b). Appendix E.6 gives the code, and Lemma 1 excludes recovery for every sufficiently near-optimal affine-ReLU SAE.

3.2 What the encoder can compute

Return to the signed source with independent exponential strengths, where Theorem 2 guarantees recovery, and keep that population fixed while adding an encoder layer. Richer encoders can narrow the gap to per-input coding when one layer cannot compute the codes (O'Neill et al., 2025). Here the true strengths are already computable, so the question is which competing responses the additional computation makes available.

Why the second layer changes the result. For nonparallel unit directions, set $m = \min(a_u, a_v)$ and write

$$a_u u + a_v v = (a_u - m)u + (a_v - m)v + \|u + v\| m\, d_{\text{sh}}, \qquad d_{\text{sh}} = \frac{u + v}{\|u + v\|}. \tag{9}$$

Figure 1(b) realizes (9), saving $(2 - \|u + v\|)m$ without increasing the active count. A second ReLU layer computes $m = a_u - (a_u - a_v)_+$ and the remainders, whereas one affine score cannot: $(x_1 - |x_2|)_+$ bends along $x_2 = 0$, which no ReLU of an affine function does (Appendix F.3).

The same construction applies to any two learned responses without access to the generator. Replacing them by their positive differences and one response along the normalized sum $d_j + d_k$ preserves reconstruction without increasing the active count, while lowering the L_1 penalty by $\lambda(2 - \|d_j + d_k\|)\mathbb{E}\min(g_j, g_k)$. This saving, studied by Dorrell (2026), is available only when the encoder can compute the minimum. Any class supporting the construction therefore has no finite L_1 optimum with two nonparallel latents active together with positive probability (Appendix F.3), whereas the one-layer class does not support the construction here.

The two-layer encoder with four axial and four diagonal outputs lowers expected activation from .75 to .668, so Lemma 1 excludes recovery throughout any class containing this encoder. At $\lambda = .1$, the exact construction also scores .0668, below the .0726 reached by optimal coding on the generating directions with optimized shared bias (Appendix F.3), and Figure 2(b) compares all three codes.

At the generating dictionary, the one-layer encoder can implement the optimal soft-thresholded code exactly. At the summed-direction dictionary it cannot attain that dictionary's optimal coding objective (Appendix F.3), while the extra layer makes that code available and thereby changes which dictionary is cheapest under joint training.

Trained recovery across co-occurrence rates. Figure 2(c) uses the fully coupled law ($\eta = 1$) and varies the pair rate c. The one-layer median remains near the soft-threshold reference through $\kappa = .45$, whereas the two-layer median departs from it by $\kappa = .10$. In a separate comparison on the two reference laws, the two-layer encoder reaches contribution NMSE .00209 under independence and .00210 under coupling when the generating directions are frozen. It loses recovery only once the directions are learned jointly. Appendices F.6 and F.3 give the sweep and the reference-population comparison.

3.3 What sparsity charges

Return to one-layer responses and replace the charge on activation magnitude by a charge for each positive output:

$$J_\beta = \mathbb{E}\big[\tfrac{1}{2}\|X - b - Dg(X)\|^2 + \beta n_g(X)\big], \qquad n_g(X) = \sum_j \mathbf{1}\{g_j(X) > 0\}. \tag{10}$$

Under this penalty, shrinking an activation that stays positive no longer lowers the cost. We allow affine-ReLU responses or the affine-Jump responses $g_j = q_j \mathbf{1}\{q_j > \theta_j\}$ used by JumpReLU SAEs, with $q_j = e_j^\top X + a_j$ and $\theta_j > 0$ learned (Rajamanoharan et al., 2024b).

For the codes of Figure 1(a), a count penalty charges each code two outputs on almost every pair, but charges the rotated code two outputs on a singleton where the generating code uses one; the generating code is therefore cheaper as $\beta \to 0$ under both laws. At finite β, a threshold that silences the weaker rotated response on a near-equal pair recovers part of rotation's saving (Appendix F.4); Appendix F.1 gives the width-uniform gate bound used below.

Theorem 3 (Activity costs recover individual contributions)**.** *Fix either magnitude law in Theorem 2 and use affine-ReLU or affine-Jump responses. Let C_β be the infimum of* (10) *over arbitrary finite widths, unit decoder columns, and free biases. Every sequence with $J_\beta \le C_\beta + o(\beta)$ contains distinct rows whose directions converge to the four generating directions and whose responses converge to the matching strengths in mean square as $\beta \to 0$. Moreover, $C_\beta/\beta \to \mathbb{E}|S|$ and $\mathbb{E}|n_g - |S|| \to 0$, where S is the set of generating features present.*

Table 1: **Objective-level outcomes on the reference population.** The first three rows give vanishing-penalty limits; TopK has no penalty, so its row records an objective tie. Fixed-group recovery permits several rows per contribution, individual recovery identifies one.

Response class and penalty	Independent strengths	Coupled strengths
One-layer affine ReLU, L_1	Fixed-group recovery	Persistent contribution error
Two-layer encoders supporting (9), L_1	Persistent contribution error	Persistent contribution error
One-layer affine ReLU/JumpReLU, L_0	Individual recovery	Individual recovery
One-layer affine TopK, quota $k \ge 2$	Generating and rotated exact codes tie	Same

What count penalties leave unresolved. With tied strengths, one shared output replaces two, so a count penalty prefers it, and even the L_1 objective does. For independent Bernoulli$(1/4)$ features, $(Z_1 + Z_2 - 1)_+$ detects joint occurrence with one affine-ReLU response, and Proposition I.1 proves that every global L_1 optimum at sufficiently small λ uses such event directions, possibly duplicated along parallel rows at larger widths. Here $\kappa = 1/4$, below the rotation threshold: κ governs only the comparison with the rotated code, and with constant magnitudes event latents can win even with

one layer. By contrast, Figure 1(c)'s hierarchy tilt leaves the count unchanged, so the count penalty is indifferent to it. JumpReLU trains at a finite count price, TopK and BatchTopK impose quotas, and Gated SAE charges a detection proxy (Gao et al., 2025; Bussmann et al., 2024; Rajamanoharan et al., 2024a). Appendices F.2, F.4, E.6, I.3, and B give the proof, boundaries, and gate-specific comparisons.

4 WHICH REPRESENTATIONS DOES TRAINING LEARN?

Where two exact codes have equal objective values, a wider SAE can improve on both by averaging their differently shrunken reconstructions, so an analytic crossing need not mark a training transition (Appendix G.8). We therefore measure attained objectives and change the distribution while continuing from the same checkpoint.

4.1 COMPLETE AND OVERCOMPLETE CONFIRMATION

We test 16 interacting factors in 16 dimensions, where the strengths remain exactly computable, and 64 factors in 32 dimensions, where no global affine inverse exists. In both sources we vary only dependence between partner magnitudes, transforming Gaussian pairs with correlation zero or .99 to exponential marginals while holding directions, occurrences, marginals and covariance fixed. In neither source does any input contain only one active factor. Appendix C.2 identifies the complete source from co-occurrence subspaces, and Appendix G.2 gives the construction, prespecified recovery control, and all 288 models.

Figure 4 in Appendix G.2 reports the paired comparison. On the complete source, independent-strength L_1 models reach median contribution NMSE $.063$, compared with $.0175$ for the faithful reference at the same $\lambda = .2$. Coupling raises the median to $.513$, worsens all eight sampled source dictionaries, and places 90.2% of activation within 15° of a pair sum or difference. This is the only setting that passes the prespecified recovery control, which it does in 23 of 24 starts. JumpReLU, Top4 and the overcomplete source fail the independent-law control, so their coupling results establish worsening from an already poor allocation rather than loss of prior recovery.

4.2 DISTINGUISHING DICTIONARY CHOICE, INFERENCE, AND TRAINING

Could better activations at the generating directions match what the coupled model attains? We solve the nonnegative coding problem at both dictionaries, optimizing one shared bias and bracketing the optimum between a feasible code and a feasible dual point, so the comparison does not depend on solver convergence.

On four preselected complete-source dictionaries, the coupled-trained encoders have mean objective $.535$, and each lies below the certified lower bound for any generating-direction code, whose optimum averages $.576$. Better inference at those directions therefore cannot match the coupled model. Under independence the trained encoder remains above that optimum, $.593$ versus $.575$, and near the generating rays. Recoding the learned dictionary beats the generating one by $.008$ under independence and $.058$ under coupling, so the laws differ in whether the trained encoder itself beats every generating-direction code, not in dictionary quality alone. Appendix G.3 gives the full audit and explains why the small independent advantage lies outside Theorem 2.

Changing the law redirects a trained model. We continue each checkpoint on either the same law or the other law, then evaluate both continuations on identical inputs from the destination law. Adding coupling raises contribution NMSE by $.1358$ while lowering the objective by $.02147$. Removing coupling lowers NMSE by $.4218$ and the objective by $.07266$. All 24 L_1 starts move as predicted when coupling is added, and removing it returns 18 of 24 starts below $.10$ (Appendix G.4).

Why the trained count-gate models still miss contributions. Under Top4, faithful and mixed exact codes both reach zero loss. With generating directions frozen, aligned starts reach $.012$ error while random starts remain near $.40$, so training misses available faithful responses. JumpReLU trains at finite β, outside Theorem 3. At its calibrated price $\beta = .16$, a threshold can silence the weaker rotated response on a near-equal pair and recover part of rotation's saving on pairs. The trained responses concentrate near pair sums and differences and nearly match the rotated code's number of active outputs (Appendices G.5 and F.4).

Similar decoder mixtures therefore occur when another decomposition is cheaper, when several exact decompositions tie, and when training misses faithful responses available at the same directions. Decoder geometry does not tell these cases apart, so we use coding and initialization controls. Appendix H adds an exact example where a recovered solution is locally stable despite a cheaper same-width alternative, so persistence alone does not establish global preference either.

5 What a learned SAE supports and what an observer can check

A failed contribution assignment does not imply that the full code has lost the generating strengths. We distinguish two operations: reading a feature's strength from the code, and removing its decoded contribution. Figure 1(d) gives the exact counterexample. Appendix I.3 proves that a small-penalty version of this code is globally optimal (Proposition I.1) and gives a rerouting edit that raises one activation while lowering another, outside the reduction class below.

Ablation and attenuation can only reduce existing activations. We therefore bound every such edit at once: the distance from $\alpha_i w_i$ to $\mathcal{C}(X) = \{Dz : 0 \le z_j \le g_j(X)\}$ lower-bounds the error of any reduction, even one chosen separately for each input. This class is broader than deleting a preassigned group. On coupled complete-source inputs, readers using the known generating inverse of the reconstruction score .0343, .0929 and .0126 for L_1, JumpReLU and Top4, while every edit that only reduces the same activations has certified error at least .352, .351 and .314. On coupled overcomplete inputs, full-code affine readers also outperform all reduction-only edits. Accurate estimates can therefore coexist with inaccurate selective reductions. Appendix I.1 gives the fits, certificates and independent-law results.

Without generating labels, neither reconstruction error nor nonlinear residual energy detects that allocation change. Reconstruction error moves by a few percent where contribution error moves by 843%. Under exact selective omission on the balanced-sign source, the fraction of residual energy unexplained by the best affine fit determines the retained fraction of squared magnitude. The same retained-energy fraction does not determine how many feature occurrences remain covered (Appendices J.3 and J.4).

The fixed-decoder test needs no generating dictionary. On four public GPT-2 SAEs, it certifies that better codes at the same decoder would lower checkpoint loss by 5–15% on a fixed 512-position sample, despite 95.7–99.6% explained variance. High explained variance does not show that the encoder uses the dictionary optimally. Section 3.2 shows that a more expressive encoder can also change the preferred directions under joint training (Appendix K).

6 Conclusion

SAEs can recover features only when, within the encoder class, no exact reallocation of the same inputs costs less than the generating code. Under L_1 with one affine-ReLU layer, independent signed strengths meet this condition and recovery follows (Theorem 2), while co-occurring strengths that are often nearly equal violate it, although directions, frequencies, marginal laws and covariance are unchanged. What they learn instead are codes that spend one activation on several features' contributions: sums and differences, a summed direction carrying the shared strength, a child latent absorbing part of its parent, or a latent per co-occurrence event. These codes reconstruct exactly, and strengths often remain readable from them, but reducing their activations cannot remove one feature's contribution. Count penalties remove this incentive only as their weight β tends to zero (Theorem 3); at the finite weights used in training, or under quotas, the same allocations can win or tie.

An SAE latent is therefore best read as a unit of the cheapest allocation that the objective and encoder can realize. It coincides with a feature of the data under conditions that depend on how co-occurring strengths vary together, which reconstruction, explained variance and residual energy do not reveal. Two consequences follow. A more expressive encoder can make a cheaper non-recovering allocation computable, so encoder changes should be evaluated both at fixed directions and under joint training, measuring contributions alongside reconstruction and code correlation (O'Neill et al., 2025; Barin Pacela et al., 2026). Benchmarks should vary the joint law of co-occurring strengths, not only which features co-occur.

REPRODUCIBILITY STATEMENT

Appendix A maps the main-text results to their supporting derivations and experiments. Appendices F and G.1 give the experimental setups, training procedures, replication counts, and calibration and test sizes for the synthetic experiments. Appendix K specifies the checkpoints, activation stream, sampling rule, and preserved objective for the public-checkpoint audit.

ETHICS STATEMENT

This work studies controlled synthetic sources and publicly released SAE checkpoints. It involves no new human-subject data collection and releases no new language model. Our intervention results evaluate changes to hidden-state contributions, and downstream effects require separate validation before these procedures are used to control a model.

USE OF LANGUAGE MODELS

We have used language models to search for the relevant literature, non-authoritative review and challenge of obtained results, and iteration upon human-drafted hypotheses, results and experiments, and for preparing the submission content. All ideas in the paper are generated by human authors, and all theoretical claims, proofs, experiments, and final text were checked and are the responsibility of the authors.

REFERENCES

Sanjeev Arora, Rong Ge, Tengyu Ma, and Ankur Moitra. Simple, Efficient, and Neural Algorithms for Sparse Coding. In *Proceedings of the Conference on Learning Theory*, volume 40, pp. 113–149, 2015. URL https://proceedings.mlr.press/v40/Arora15.html.

Vitória Barin-Pacela, Shruti Joshi, Isabela Camacho, Simon Lacoste-Julien, and David Klindt. Stop Probing, Start Coding: Why Linear Probes and Sparse Autoencoders Fail at Compositional Generalisation. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337, pp. 364–412, 2026. URL https://proceedings.mlr.press/v337/barin-pacela26a.html.

Dan Braun, Jordan Taylor, Nicholas Goldowsky-Dill, and Lee Sharkey. Identifying Functionally Important Features with End-to-End Sparse Dictionary Learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 107286–107325, 2024. URL https://doi.org/10.52202/079017-3408.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nicholas L. Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E. Burke, Tristan Hume, Shan Carter, Tom Henighan, and Chris Olah. Towards Monosemanticity: Decomposing Language Models With Dictionary Learning. *Transformer Circuits Thread*, 2023. URL https://transformer-circuits.pub/2023/monosemantic-features/index.html.

Bart Bussmann, Patrick Leask, and Neel Nanda. BatchTopK Sparse Autoencoders, 2024. URL https://arxiv.org/abs/2412.06410.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning Multi-Level Features with Matryoshka Sparse Autoencoders. In *Proceedings of the International Conference on Machine Learning*, volume 267, pp. 6077–6101, 2025. URL https://proceedings.mlr.press/v267/bussmann25a.html.

David Chanin and Adrià Garriga-Alonso. Sparse but Wrong: Incorrect L0 Leads to Incorrect Features in Sparse Autoencoders. In *Proceedings of the International Conference on Machine Learning*, 2026a. URL https://openreview.net/forum?id=eSIxjYeQhD.

David Chanin and Adrià Garriga-Alonso. SynthSAEBench: Evaluating Sparse Autoencoders on Scalable Realistic Synthetic Data, 2026b. URL https://arxiv.org/abs/2602.14687.

David Chanin, Tomáš Dulka, and Adrià Garriga-Alonso. Feature Hedging: Correlated Features Break Narrow Sparse Autoencoders. In *Mechanistic Interpretability Workshop at NeurIPS*, 2025a. URL https://openreview.net/forum?id=wg0TxhOKLg.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for Absorption: Studying Feature Splitting and Absorption in Sparse Autoencoders. In *Advances in Neural Information Processing Systems*, volume 38, pp. 91489–91526, 2025b. URL https://doi.org/10.52202/085713-2755.

Siyu Chen, Heejune Sheen, Xuyuan Xiong, Tianhao Wang, and Zhuoran Yang. Taming Polysemanticity in LLMs: Theory-Grounded Feature Recovery via Sparse Autoencoders. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=VtWkPIbAQ8.

Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the Limits of Sparse Autoencoders: A Theoretical Framework and Reweighted Remedy. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=DSOTgzeH3w.

William Dorrell. How Optimality Structures Sparse Dictionaries: A Theory for Understanding SAE Representations, 2026. URL https://arxiv.org/abs/2606.02385.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy Models of Superposition. *Transformer Circuits Thread*, 2022. URL https://transformer-circuits.pub/2022/toy_model/index.html.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not All Language Model Features Are One-Dimensionally Linear. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/d3221cdb27e49d9c1cd35ad254feccfe-Abstract-Conference.html.

Joshua Engels, Logan Riggs Smith, and Max Tegmark. Decomposing The Dark Matter of Sparse Autoencoders. *Transactions on Machine Learning Research*, 2025b. URL https://openreview.net/forum?id=sXq3Wb3vef.

Lucy Farnik, Tim Lawson, Conor Houghton, and Laurence Aitchison. Jacobian Sparse Autoencoders: Sparsify Computations, Not Just Activations. In *Proceedings of the International Conference on Machine Learning*, volume 267, pp. 16178–16236, 2025. URL https://proceedings.mlr.press/v267/farnik25a.html.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and Evaluating Sparse Autoencoders. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/42ef3308c230942d223c411adf182c88-Abstract-Conference.html.

Gleb Gerasimov, Timofei Rusalev, Nikita Balagansky, Daniil Laptev, Vadim Kurochkin, and Daniil Gavrilov. Unstable Features, Reproducible Subspaces: Understanding Seed Dependence in Sparse Autoencoders, 2026. URL https://arxiv.org/abs/2606.12138.

Karol Gregor and Yann LeCun. Learning Fast Approximations of Sparse Coding. In *Proceedings of the International Conference on Machine Learning*, pp. 399–406, 2010. URL https://icml.cc/Conferences/2010/papers/449.pdf.

Sai Sumedh R. Hindupur, Ekdeep Singh Lubana, Thomas Fel, and Demba E. Ba. Projecting Assumptions: The Duality Between Sparse Autoencoders and Concept Geometry. In *Advances in Neural Information Processing Systems*, volume 38, pp. 12986–13036, 2025. URL https://doi.org/10.52202/085713-0391.

Robert Huben, Logan Smith, Hoagy Cunningham, Aidan Ewart, and Lee Sharkey. Sparse Autoencoders Find Highly Interpretable Features in Language Models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/1fa1ab11f4bd5f94b2ec20e794dbfa3b-Abstract-Conference.html.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the Origins of Linear Representations in Large Language Models. In *Proceedings of the International Conference on Machine Learning*, volume 235, pp. 21879–21911, 2024. URL https://proceedings.mlr.press/v235/jiang24d.html.

Subhash Kantamneni, Joshua Engels, Senthooran Rajamanoharan, Max Tegmark, and Neel Nanda. Are Sparse Autoencoders Useful? A Case Study in Sparse Probing. In *Proceedings of the International Conference on Machine Learning*, volume 267, pp. 29018–29049, 2025. URL https://proceedings.mlr.press/v267/kantamneni25a.html.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Isaac Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum Stuart McDougall, Kola Ayonrinde, Demian Till, Matthew Wearden, Arthur Conmy, Samuel Marks, and Neel Nanda. SAEBench: A Comprehensive Benchmark for Sparse Autoencoders in Language Model Interpretability. In *Proceedings of the International Conference on Machine Learning*, volume 267, pp. 29223–29264, 2025. URL https://proceedings.mlr.press/v267/karvonen25a.html.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Rogov, Ivan Oseledets, and Elena Tutubalina. Sanity checks for sparse autoencoders: Do saes beat random baselines?, 2026. URL https://arxiv.org/abs/2602.14111.

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. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/84ca3f2d9d9bfca13f69b48ea63eb4a5-Paper-Conference.pdf.

Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient Sparse Coding Algorithms. In *Advances in Neural Information Processing Systems*, volume 19, pp. 801–808, 2006. URL https://papers.nips.cc/paper/2979-efficient-sparse-coding-algorithms.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. Gemma Scope: Open Sparse Autoencoders Everywhere All At Once on Gemma 2. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 278–300, 2024. URL https://aclanthology.org/2024.blackboxnlp-1.19/.

Yizhou Liu, Ziming Liu, and Jeff Gore. Superposition Yields Robust Neural Scaling. In *Advances in Neural Information Processing Systems*, volume 38, pp. 176757–176793, 2025. URL https://doi.org/10.52202/085713-5320.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the International Conference on Machine Learning*, volume 97, pp. 4114–4124, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards Principled Evaluations of Sparse Autoencoders for Interpretability and Control. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/53356aebeea8ffd40a8ac3bb66243162-Abstract-Conference.html.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse Feature Circuits: Discovering and Editing Interpretable Causal Graphs in Language Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3ba4d47a83e498c2b1a0868cba20f6de-Abstract-Conference.html.

Tommaso Mencattini, Francesco Montagna, and Francesco Locatello. The Rate-Distortion-Polysemanticity Tradeoff in SAEs, 2026. URL https://arxiv.org/abs/2605.14694.

Walter Nelson, Theofanis Karaletsos, and Francesco Locatello. Toward Identifiable Sparse Autoencoders, 2026. URL https://arxiv.org/abs/2605.31245.

Charles O'Neill, Alim Gumran, and David Klindt. Compute Optimal Inference and Provable Amortisation Gap in Sparse Autoencoders. In *Proceedings of the International Conference on Machine Learning*, volume 267, pp. 46877–46896, 2025. URL https://proceedings.mlr.press/v267/o-neill25a.html.

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models. In *Proceedings of the International Conference on Machine Learning*, volume 235, pp. 39643–39666, 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Gonçalo Paulo and Nora Belrose. Sparse Autoencoders Trained on the Same Data Learn Different Features. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3c1fe56b043848b211030c202764c6a7-Abstract-Conference.html.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, János Kramár, Rohin Shah, and Neel Nanda. Improving Sparse Decomposition of Language Model Activations with Gated Sparse Autoencoders. In *Advances in Neural Information Processing Systems*, volume 37, pp. 775–818, 2024a. URL https://doi.org/10.52202/079017-0024.

Senthooran Rajamanoharan, Tom Lieberum, Nicolas Sonnerat, Arthur Conmy, Vikrant Varma, János Kramár, and Neel Nanda. Jumping Ahead: Improving Reconstruction Fidelity with JumpReLU Sparse Autoencoders, 2024b. URL https://arxiv.org/abs/2407.14435.

Marc'Aurelio Ranzato, Christopher Poultney, Sumit Chopra, and Yann LeCun. Efficient Learning of Sparse Representations with an Energy-Based Model. In *Advances in Neural Information Processing Systems*, volume 19, pp. 1137–1144, 2006. URL https://papers.nips.cc/paper_files/paper/2006/hash/87f4d79e36d68c3031ccf6c55e9bbd39-Abstract.html.

Daniel A. Spielman, Huan Wang, and John Wright. Exact Recovery of Sparsely-Used Dictionaries. In *Proceedings of the Conference on Learning Theory*, volume 23, pp. 37.1–37.18, 2012. URL https://proceedings.mlr.press/v23/spielman12.html.

Xiaoqing Sun, Alessandro Stolfo, Joshua Engels, Ben Wu, Senthooran Rajamanoharan, Mrinmaya Sachan, and Max Tegmark. Dense SAE Latents Are Features, Not Bugs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 155583–155613, 2025. URL https://doi.org/10.52202/085713-4686.

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, Alex Tamkin, Esin Durmus, Tristan Hume, Francesco Mosconi, 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. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Demian Till. Do Sparse Autoencoders Find "True Features"?, 2024. URL https://www.lesswrong.com/posts/QoR8noAB3Mp2KBA4B/do-sparse-autoencoders-find-true-features.

Yu Wang, Siqi Wu, and Bin Yu. Unique Sharp Local Minimum in L1-Minimization Complete Dictionary Learning. *Journal of Machine Learning Research*, 21:1–52, 2020. URL https://jmlr.org/papers/v21/19-169.html.

Zhengxuan Wu, Aryaman Arora, Atticus Geiger, Zheng Wang, Jing Huang, Dan Jurafsky, Christopher D. Manning, and Christopher Potts. AxBench: Steering LLMs? Even Simple Baselines Outperform Sparse Autoencoders. In *Proceedings of the International Conference on Machine Learning*, volume 267, pp. 67035–67080, 2025. URL https://proceedings.mlr.press/v267/wu25a.html.

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. URL https://openreview.net/forum?id=Hl3rEn7S4P.

A WHERE TO FIND THE SUPPORTING ARGUMENT

The appendices follow the main argument. Appendix B places the question in prior work; Appendices C–E define the target and prove the objective comparisons; Appendices F–H treat encoder access, penalties, training and stability; and Appendices I–K treat observer tests. Appendix G.1 first records the shared implementation and cohort sizes; later experimental sections state their departures and evaluation data.

Table 2: Where each main-text question is answered.

Question	Direct entry points
Why are the sources identifiable, and what does the recovery score allow?	Appendices C.1, C.2, and C.3: identification, fractional allocation, and splitting invariance.
Can the reference geometry arise from learning?	Appendix C.4: source objective, optimum, curriculum, and imperfect learned checkpoint.
What is optimized when the decoder is fixed?	Appendices D.1 and D.2: encoder-class gap, coding conditions, and primal–dual bounds.
Why does independence force recovery at any finite width?	Appendices E.1 and E.2: response inequality, directional strictness, and recovery of the strengths.
Why does a cheaper code exclude every faithful near-optimum?	Appendices E.4, E.5, E.8, and E.6: competing codes, the class-wide activation bound, and the one-sided boundary.
What changes with encoder depth, count penalties, and co-occurrence?	Appendices F.3, F.1–F.5, and F.6: summed-direction responses, the count result and its boundaries, and the sweep in Figure 2(c).
How were training, coding, and optimization separated?	Appendices G.1–G.8 record the cohorts, coding audit, continuation, initialization tests, robustness, scale, and finite-price controls; Appendix H gives the stability counterexample.
What remains readable, removable, or auditable without labels?	Appendices I.1–J.4 give the readers, reduction bounds, residual identities and coverage tests; Appendix K gives the public-checkpoint audit.

A.1 THE FOUR CONDITIONS, AND WHAT TESTS EACH

Section 1 gives four conditions that must hold before a trained SAE supplies a feature's contribution. The paper tests them separately.

Identification. On the reference source, inputs where one factor occurs alone put positive probability on each generating line, while pairs with continuously varying strengths put zero probability on any fixed mixed line (Appendix C.1). The 16- and 64-factor sources of Section 4.1 have no global singleton events. Appendix C.2 identifies the complete source through intersections of event planes; the overcomplete source remains an evaluation target without a corresponding identification claim.

Encoder access. On the reference source, $(\pm u_i^\top X)_+$ returns all four strengths exactly. Appendix C.4 gives the same reference when the source geometry is learned imperfectly. Section 3.2 then enlarges the encoder and asks which competing responses become computable.

Objective preference. Appendix E.1 bounds every response in the class, and Appendix E.5 converts any feasible activation saving into unavoidable contribution error. Appendix G.3 compares trained encoders with the best codes on the generating directions at the operating penalty.

Optimization. Appendix G.5 freezes the generating directions and varies initialization to isolate a Top4 optimization failure. Appendix G.4 changes only the strength law from the same checkpoint to test whether coupling changes the learned representation.

A.2 SCOPE OF THE CONCLUSIONS

Signed sources. The positive half of Theorem 2 uses factors that occur alone on both sides of each generating axis. Appendix E.6 removes the signs while retaining the occurrence process, independent exponential strengths and exact affine inverse, and obtains a cheaper tilted code. The guarantee therefore does not extend to one-sided features as stated. Dense language-model latents can occur in near-opposite pairs (Sun et al., 2025), but we have not established the balanced law used here; extending the result requires a response bound without the two-sided constraint.

Vanishing and operating penalties. Theorems 2 and 3 take $\lambda \to 0$ or $\beta \to 0$, whereas training uses L_1 penalties between .1 and .4. Appendix G.3 compares attained and generating-direction objectives at the operating penalty, and Appendix G.4 checks that movement toward pair combinations lowers the destination objective. Appendix G.8 shows why a crossing of two named codes need not be a training boundary: at $c = .63$ their crossing is $\lambda = .0269$, a mixture is cheaper there, and random-start recovery changes smoothly across it.

Encoder regime. Our reversal adds capacity where one affine-ReLU layer already computes the true strengths. Where it cannot, richer encoders can instead improve inference (O'Neill et al., 2025). Their reported code-correlation metric does not test contribution allocation, and we do not determine which effect dominates when co-occurring strengths are shared.

Scale and external validity. The primary sources reach 64 factors in 32 dimensions; Appendix G.7 adds a 256-factor control in the same dimension. Known synthetic contributions make recovery measurable, while the public checkpoints of Appendix K supply no such targets. The paper therefore does not estimate how often these mechanisms explain allocation failures in language-model SAEs. That test requires contributions fixed independently of the SAE, for example by a circuit model. Without such targets, the checkpoint audit asks only how much the coding objective can improve at a fixed decoder.

B Relation to prior work

From linear representations to a recovery target. SAEs entered language-model interpretation as dictionaries of directions with sparse, input-dependent activations (Huben et al., 2024; Bricken et al., 2023; Templeton et al., 2024). The linear representation hypothesis connects such directions to variables used by a network; Park et al. (2024) distinguish representation, probing, and steering through counterfactual definitions, while Jiang et al. (2024) derive linear representations from a latent-variable model of language and properties of the training objective. Toy superposition models make feature frequency and representational geometry explicit (Elhage et al., 2022; Liu et al., 2025). Dense antipodal SAE latents also occur in language models (Sun et al., 2025), motivating our treatment of signed factors. We take this literature as the starting point for a controlled additive source and state its source family explicitly rather than treating its directions as a universal feature inventory. General disentanglement results make such source assumptions essential to identification (Locatello et al., 2019). The source-learning calculation in Appendix C.4 verifies these identification conditions for our source. Multidimensional representations found by Engels et al. (2025a) and the geometry-dependent detection analysis of Hindupur et al. (2025) mark further modeling choices beyond the fixed-direction setting we study.

Coding a dictionary and learning one. Classical sparse coding separates coefficient inference from dictionary optimization (Lee et al., 2006). Predictive sparse encoders and learned approximations to iterative inference address the cost of obtaining coefficients (Ranzato et al., 2006; Gregor & LeCun, 2010). For modern SAEs, O'Neill et al. (2025) prove an amortisation gap for shallow encoders and examine richer inference. Under compositional shift, Barin-Pacela et al. (2026) find that per-sample FISTA at a learned SAE decoder leaves a recovery gap when the learned directions remain inaccurate. These results motivate our fixed-dictionary audit of remaining objective improvement and our joint-training encoder comparison on an identifiable source. There, the affine class codes the generating dictionary optimally but pays more at a dictionary with shared directions, whereas an explicit second layer makes that competing decomposition affordable.

Recovery guarantees and competing optima. Classical guarantees recover sparsely used full-rank dictionaries (Spielman et al., 2012) or incoherent overcomplete dictionaries under their sampling conditions (Arora et al., 2015). SAE-specific results exploit activation-frequency control (Chen et al., 2026), characterize a tied unbiased objective (Cui et al., 2026), or connect accurate codes and dictionary conditioning to identifiability (Nelson et al., 2026). We adopt this generator-based view while changing the recovery target and model class: the target permits one generating contribution to be split across a fixed group, and the positive theorem controls unrestricted minimization within the untied affine-ReLU class on its stated source, with extra latents and free biases.

Sparse coding can reward combinations rather than the generating factors. Wang et al. (2020) give a correlated Gaussian example where rotation reduces the L_1 coding objective although the generating

dictionary is a sharp local minimum. Dorrell (2026) derive constraints at nonnegative dictionary-learning optima and use them to explain splitting, absorption, and residual structure. Mencattini et al. (2026) formalize a reconstruction–activity–polysemanticity tradeoff. These results motivate our comparison of two matched laws. We hold directions, occurrence statistics, marginal magnitudes, and covariance fixed, then obtain an all-response recovery guarantee under one law and a contribution budget excluding recovery under the other. The budget extends a competing-code calculation to every sufficiently near-optimal SAE, including dictionaries that retain the original directions alongside their competitors.

Splitting, absorption, and collective representations. Chanin et al. (2025b) document context-dependent missing responses and partial absorption in language-model SAEs. Chanin et al. (2025a) analyze how limited width and feature correlation favor mixed directions, while Till (2024) discusses the broader ambiguity of interpreting sparsity-selected features. These observations motivate a recovery target tolerant to splitting. Individual-latent variation across seeds (Paulo & Belrose, 2026) and non-canonical learned units (Leask et al., 2025) likewise motivate evaluating more than one-to-one matches, while reproducible subspaces (Gerasimov et al., 2026) offer a collective geometric target. Our grouped contribution criterion asks what vector the activations in such a collection supply on each input, because a stable span alone leaves those coefficients undetermined. The continuous hierarchy in Appendix E.6 then connects absorption to an exact affine code and the class-wide activation budget.

Other objectives change the comparison. TopK constrains activity directly (Gao et al., 2025); JumpReLU separates its positive threshold from the magnitude passed to the decoder (Rajamanoharan et al., 2024b), and is used in Gemma Scope (Lieberum et al., 2024). BatchTopK distributes a quota over a batch (Bussmann et al., 2024). These designs distinguish a per-input cap from a penalty and from mean achieved activity, a distinction also emphasized by Chanin & Garriga-Alonso (2026a). Gated SAEs penalize a detection proxy and include an auxiliary reconstruction (Rajamanoharan et al., 2024a); that forward objective requires its own total-loss comparison. Matryoshka training evaluates nested partial reconstructions (Bussmann et al., 2025), end-to-end sparse dictionary learning weights errors by downstream effects (Braun et al., 2024), and Jacobian SAEs directly sparsify the Jacobian connecting features across a model component (Farnik et al., 2025). Identical full reconstructions can differ under partial-reconstruction or code-dependent losses. We therefore keep the conclusions objective-specific: each theorem and experiment refers to the objective actually analyzed or trained.

What evaluation establishes. Interpretability and control evaluations (Makelov et al., 2025), sparse probing (Kantamneni et al., 2025), sparse feature circuits (Marks et al., 2025), and SAEBench (Karvonen et al., 2025) test different uses of SAE latents. SynthSAEBench supplies known directions and firings at scale, including correlated and hierarchical supports (Chanin & Garriga-Alonso, 2026b). We use these evaluations as motivation for separating questions that are often bundled: what an activation lets us read, what vector the code allocates, what a reduction-only edit can remove, and what fixed-decoder coding improvements are observable. Decode–encode consistency can improve latent recovery under suitable decoder conditions (Zhu et al., 2026); the exact event code here shows why consistency alone does not supply separately removable contributions. Finally, Engels et al. (2025b) distinguish affine-predictable from nonlinear reconstruction error. Our retention calculation connects that measurement to selective omission, then shows why the same retained energy leaves occurrence coverage unresolved. The checkpoint audit uses observable objective values without assigning the checkpoint a generating semantics.

C Sources, identification, and the recovery target

The reference source identifies its directions through isolated occurrences; the complete confirmation source identifies them through event-plane intersections. We give those arguments, define the empirical assignment once, and then examine the separately trained source network.

C.1 What observations can identify

Let P_X be the observed hidden-state law. Two latent models can induce the same P_X while assigning different meanings to the same point. In the four-point law $\{0, u, c, u + c\}$, one model uses two rays u, c and assigns $u + c$ to their coactivation; another uses three mutually exclusive rays $u, c, u + c$.

No procedure using only samples from P_X, the SAE loss, and its checkpoint can distinguish them. The observations identify a target only relative to a stated family of sources or an external anchor. This is the distinction between source identification and the existence of other codes for the same observations (Locatello et al., 2019).

For an externally supplied dictionary A, the compatible nonnegative codes are $\mathcal{P}_A(x) = \{z \geq 0 : Az = x\}$. The minimum and maximum feasible value of coordinate i determine whether the anchored feature is absent in every compatible code, present in every code, or support-ambiguous. These linear programs certify semantics relative to A; they cannot certify that A is the unique latent ontology behind P_X.

For the signed reference source, assume a finite set of distinct rays, positive probability of an isolated occurrence on every ray, and continuously varying coefficients on higher-dimensional events. Isolated occurrences produce atoms in the angular law of $X/\|X\|$. A full-dimensional continuous pair event puts zero mass on any fixed ray, so these atoms identify the generating rays within that family. Splitting a signed factor into its two oriented rays then gives its nonnegative contributions. This argument concerns the source distribution; the SAE is not given occurrence labels.

C.2 Why the interacting source identifies its rays

The main complete-source experiment uses pair blocks. Within each block the two unit factors are perpendicular; blocks have independently drawn orientations. Each input selects two distinct blocks and chooses both members of a selected block with probability .6, or either singleton with probability .2, so the conditional partner probability is $.6/(.6 + .2) = .75$, matching the reference population of Section 3.1. Within-block strengths are $B_i = -\log \Phi(H_i)$ for correlated standard normals; balanced signs and different blocks are independent.

Exactly two-active inputs occur with positive probability and lie in planes spanned by factors from different blocks. Events with three or four active factors have continuous densities in higher-dimensional spans and place zero mass on any fixed plane. The positive-mass planes are therefore observable from the population. Two such planes sharing one factor intersect in its line, whereas planes on four distinct factors intersect only at zero. Every factor belongs to several planes, so these intersections identify all generating lines even though no input contains a single globally isolated factor. Observation noise or deterministic magnitude ties would invalidate this exact population argument; it provides no finite-sample identification rate.

In the complete source the generating matrix is invertible. Its inverse, followed by splitting signed coordinates into positive and negative parts, realizes the generating code with an affine-ReLU encoder. Rotating each factor pair before taking signed parts realizes the competing pair code as well. The experiment therefore keeps both representations inside the tested architecture. Appendix G records the finite-sample protocol and the separate empirical recovery control for the nonorthogonal complete geometry.

The same source also retains the activation obstruction under coupling. With two selected blocks and within-block pair probability .6, the generating code has expected activation $M_{\mathrm{gen}} = 3.2$, whereas the exact rotated code has

$$M_{\mathrm{rot}} = 2\sqrt{2}\{.4 + .6\mathbb{E}\max(B_1, B_2)\}.$$

For $B = -\log \Phi(G)$ and within-pair Gaussian correlation ρ, Gaussian integration by parts gives $\mathbb{E}\max(B_1, B_2) \leq 1 + \sqrt{(1-\rho)K/2}$ with $K = 3/2 + \sqrt{2/\pi}$. At $\rho = .99$, $M_{\mathrm{rot}} \leq 3.01034 < 3.2$. Combining this feasible code with the fractional contribution budget yields $\liminf R_{\mathrm{frac}} \geq .0017565$ for sequences within $o(\lambda)$ of the class optimum. This source-specific bound establishes a persistent negative result under coupling; it does not predict the finite-price plateau or supply an affine inverse for the overcomplete source.

C.3 Fractional assignment and its implementation

The main-text score assigns decoded vectors, with one fixed assignment reused on every input. Put $t_i(X) = \alpha_i w_i$ and $c_j(X) = d_j g_j(X)$. For $\mathcal{A} = \{\Omega \geq 0 : \sum_i \Omega_{ij} \leq 1\}$ define the unnormalized

risk

$$\mathcal{E}_{\mathrm{frac}} = \min_{\Omega \in \mathcal{A}} \sum_i \mathbb{E} \left\| t_i - \sum_j \Omega_{ij} c_j \right\|^2 . \tag{C.1}$$

The normalized population score is $\mathcal{E}_{\mathrm{frac}}/T$, with $T = \mathbb{E}\sum_i \alpha_i^2$. Reported empirical scores fit Ω on calibration inputs and evaluate that fixed matrix on different inputs. This distinction separates a population minimum from a finite fitted score.

Proposition C.1 (Proportional splitting and fractional assignment)**.** *For K learned latents and F targets with finite second moments, $\mathcal{E}_{\mathrm{frac}}$ equals the minimum disjoint-group error over all finite proportional splittings of the learned responses. At most $(F+1)K$ refined outputs suffice. Its value is unchanged by proportional duplication.*

Proof. A grouping of split copies induces Ω_{ij} by adding the fractions of original response j assigned to target i. Conversely, replace response g_j by $\Omega_{ij} g_j$ for each target and an unused remainder $(1 - \sum_i \Omega_{ij}) g_j$. These copies realize the fractional assignment as disjoint groups. The assignment set is compact and the risk is a continuous quadratic, so its minimum is attained. For ReLU, $t(e^\top X + a)_+ = (te^\top X + ta)_+$ for $t \geq 0$, keeping the response in the same class and preserving its total L_1 activation sum. Duplicating a response therefore changes neither minimum. Simultaneous duplication can change active count; the proposition concerns contribution error and L_1 equivalence. □

The fit is a convex quadratic over column simplices with an unused remainder. It can be evaluated from the calibration moments $\mathbb{E}[c_j^\top c_k]$, $\mathbb{E}[t_i^\top c_j]$, and $\mathbb{E}\|t_i\|^2$. We use projected-gradient fitting and an independent quadratic evaluation; the source-specific calibration sizes and numerical checks are in Appendix G. Decoder directions, response weights, and output bias are held fixed. The bias is a shared offset and is not assigned to generating features.

The column budget also makes the assignment an operation on the original code: scaling target i by $s \geq 0$ multiplies response j by $1 + (s-1)\Omega_{ij}$. The multiplier is nonnegative, and the change is exactly $(s-1)v_i$. This operation does not require constructing the split network used in the proof.

The score therefore measures a hidden-state change that the code itself can deliver. A model scoring well admits the intended edit, while a model scoring badly does not. The score does not predict how later computation responds, which is a separate question about the network downstream of the hidden state. For target i, the error of the intended change is $(s-1)^2 \mathbb{E}\|v_i - \alpha_i w_i\|^2$, since the delivered change is $(s-1)v_i$ where $(s-1)\alpha_i w_i$ was wanted; $s = 0$ is ablation and $s = 2$ doubling. Summing over targets and dividing by T returns the normalized score (4), making it the error of this family of edits. The reduction certificate in Appendix I.2 gives an even more permissive test by choosing reductions separately on each input.

C.4 What the source reconstruction objective forces

Let $x_i = Z_i \epsilon_i B_i$ for $i = 1, \ldots, r$, and $a_{i,+} = (x_i)_+$, $a_{i,-} = (-x_i)_+$. The signs are mutually independent, uniform on $\{-1, 1\}$, and independent of the support and magnitudes. Write $S = \{i : Z_i = 1\}$, $p_i = \mathbb{P}(Z_i = 1)$ and $\tau_{ij} = \mathbb{P}(Z_i = Z_j = 1)$. These supports index signed *factors*; a support event for the $2r$ nonnegative features also specifies their signs. Magnitudes are independent of S and have finite second moments. The source objective is

$$\mathcal{L}_{\mathrm{src}}(W, b^{\mathrm{src}}) = \mathbb{E} \sum_{i,s} I_{i,s} \left[a_{i,s} - (W_{i,s}^\top W a + b_{i,s}^{\mathrm{src}})_+ \right]^2, \qquad I_{i,s} > 0, \quad W \in \mathbb{R}^{r \times 2r}. \tag{C.2}$$

This objective belongs to the source network, which receives the prescribed coefficients as reconstruction targets. The downstream SAE receives only $X = Wa$ and minimizes its own reconstruction and sparsity loss. Source weights and bias are unrestricted; the main-text experiments use $I_{i,s} = 1$.

Proposition C.2 (Source geometry)**.** *If $p_i > 0$, $\tau_{ij} > 0$ for every distinct pair, and the positive magnitudes are mutually independent and nonconstant, the global minimum of* (C.2) *is zero. Its complete set of minimizers is $W_{i,\pm} = \pm u_i$, with $u_i^\top u_j = \delta_{ij}$ and $b^{\mathrm{src}} = 0$. The same conclusion holds for dependent magnitudes if every signed singleton and every signed pair-only event has positive probability and each singleton magnitude is nonconstant.*

Proof. An orthonormal set of antipodal pairs with zero bias reconstructs a exactly. Thus a global minimizer must reconstruct every target. Write $G = W^\top W$. Conditional on a support and signs, the equality for an active target is

$$(G_{is,is} - 1)B_i + \sum_{j \in S \setminus \{i\}} G_{is,j\epsilon_j} B_j + b_{is}^{\mathrm{src}} = 0 \quad \text{almost surely.} \tag{C.3}$$

Independent nonzero magnitude variances force every coefficient of this affine expression to vanish. Hence $G_{is,is} = 1$, $b_{is}^{\mathrm{src}} = 0$, and cross-factor Gram entries are zero. In the dependent-magnitude alternative, singleton variation first pins the diagonal and bias. On a pair-only event the remaining equation is $G_{is,jt}B_j = 0$, which again pins the cross entry.

The r subspaces $\mathrm{span}\{W_{i,+}, W_{i,-}\}$ are nonzero and mutually orthogonal in dimension r, so each is a line. The inactive opposite-state readout forces $G_{i+,i-} \le 0$. Two unit vectors in a line with nonpositive inner product are antipodal. This gives the full set of minimizers, up to common hidden rotation. □

Magnitude variation and hidden dimension do distinct work. For constant $B_i = 1$, the choices $W_{i,s} = s t_{i,s} u_i$ and $b_{i,s}^{\mathrm{src}} = 1 - t_{i,s}^2$, with $t_{i,s} \ge 1$, also give zero loss. Variation removes that norm–bias freedom. If the hidden dimension exceeds r, orthogonal factor subspaces need not be lines, so antipodality is no longer forced.

Why the support family matters for exact source reconstruction. The same active-target argument gives a restriction for a general positive source a with Gram matrix $G = W^\top W$. Conditional on an event $S = T$, suppose the topological support of a_T contains an open subset of $(0, \infty)^T$. Exact tied reconstruction requires $(G_{TT} - I)a_T + b_T^{\mathrm{src}} = 0$ almost surely. Continuity extends this equality to the conditional support; an affine function vanishing on an open set has zero linear and constant parts. Hence $G_{TT} = I$ and $b_T^{\mathrm{src}} = 0$. Inactive outputs additionally require $G_{jT}a_T + b_j^{\mathrm{src}} \le 0$ for $j \notin T$. If every pair occurs on such a full-dimensional pair event, these conditions force $G = I$, which is impossible for more features than hidden dimensions. The signed source avoids this obstruction through mutual exclusion. For general overcomplete positive sources, its imperfect-checkpoint reference below remains available even when zero source loss is impossible.

How coactivation supplies a restoring force. For independent magnitudes supported in $[b_-, b_+] \subset (0, \infty)$, sufficiently small perturbations of a source optimum keep every positive target's gate active. On each active target, squared error is at least the conditional variance of the affine expression in (C.3). Independent magnitudes make this variance a sum of squared Gram errors. The target sign occurs with probability $p_i/2$ and each coactive sign pair with probability $\tau_{ij}/4$, giving, with $\sigma_i^2 = \mathrm{Var}(B_i)$,

$$\mathcal{L}_{\mathrm{src}}(W, b^{\mathrm{src}}) \ge \sum_{i,s} I_{i,s} \left[\frac{p_i \sigma_i^2}{2} (G_{is,is} - 1)^2 + \sum_{j \ne i,t} \frac{\tau_{ij} \sigma_j^2}{4} G_{is,jt}^2 \right]. \tag{C.4}$$

The omitted squared conditional means and inactive-target errors are nonnegative. Bias adjustment cannot remove these variance terms. Rare coactivations therefore give weak cross-factor constraints even when the global optimum is unchanged.

A reference at an imperfect source checkpoint. For any W, remove zero columns, put $\rho_i = \|W_i\|$, and use unit decoders $d_i = W_i/\rho_i$. The source readout gives feasible affine-ReLU responses $g_i^{\mathrm{ref}}(X) = \rho_i (W_i^\top X + b_i^{\mathrm{src}})_+$, with target coefficient $\rho_i a_i$ and reconstruction $W\widehat{a}$. Weighted Cauchy–Schwarz and the operator norm give

$$\sum_i \mathbb{E}(g_i^{\mathrm{ref}} - \rho_i a_i)^2 \le \max_i (\rho_i^2 / I_i) \mathcal{L}_{\mathrm{src}}, \qquad \mathbb{E}\|X - W\widehat{a}\|^2 \le \|W \,\mathrm{diag}(I)^{-1/2}\|_{\mathrm{op}}^2 \mathcal{L}_{\mathrm{src}},$$

$$\left| \sum_i \mathbb{E} g_i^{\mathrm{ref}} - \sum_i \mathbb{E}\rho_i a_i \right| \le \left(\sum_i \rho_i^2 / I_i \right)^{1/2} \sqrt{\mathcal{L}_{\mathrm{src}}}. \tag{C.5}$$

Thus a trained source supplies an explicit SAE encoder as a reference, with coefficient and hidden-state errors controlled by the measured source loss. This reference is available even when the source has not reached the exact optimum; the downstream objective must then compare both its reconstruction error and activation sum.

Learning the source reference.

Source losses sum coefficient squared errors; SAE losses use (5). Both average over examples without dividing by width. Computation is double precision. The downstream SAEs use untied affine-ReLU encoders, unit decoders, free biases, and projected decoder gradients.

At $r = 8$, the independent source activates each factor with probability $1/8$; the clustered source has four independent pairs with the same marginals and conditional partner probability $3/4$. Balanced signs and uniform $[1, 2]$ magnitudes are independent of support. Eight paired starts per law train source weights from uniform $[-\sqrt{2/r}, \sqrt{2/r}]$ with zero bias, fresh batches of 256, and 5,000 Adam updates at rates $.015/.003/.0006$ for $3{,}000/1{,}000/1{,}000$ updates. Direct training uses the target law throughout. The curriculum first activates all factors for 3,000 updates, strengthening the cross-factor terms in (C.4) without changing the optimal Gram, then switches to the target law.

Recovery requires coefficient error below 10^{-8} on 100,000 fresh inputs and agreement with the antipodal Gram. Direct errors are .043–.099; the curriculum's maximum is 7.3×10^{-11}. We freeze the lowest-validation-error checkpoint without downstream selection. This establishes that the characterized geometry is reachable under the stated curriculum. It is a source-network experiment; the downstream paired SAE experiments use the explicit populations recorded in Appendix G.

When the learned source is only approximate. A feasible alternative with half-squared reconstruction error R_{alt} and expected activation M_{alt} gives $C_\lambda \le R_{\mathrm{alt}} + \lambda M_{\mathrm{alt}}$. Write δ_{opt} for the SAE objective gap and $\mathcal{E}_{\mathrm{contrib}}$ for its summed mean-norm generating-contribution error. The accounting in Appendix E.5 yields

$$\mathcal{E}_{\mathrm{contrib}} \ge M_{\mathrm{gen}} - M_{\mathrm{alt}} - \frac{R_{\mathrm{alt}} + \delta_{\mathrm{opt}}}{\lambda}.$$

Thus the alternative's reconstruction error must be compared with the price-weighted activation saving before inferring a recovery deficit. Applying sharing to a learned code preserves that code's reconstruction error exactly. This supplies a constructive test on imperfect learned geometries, while a fixed source error alone cannot support the $\lambda \to 0$ comparison.

D JOINT-LOSS AND FULL-DICTIONARY DERIVATIONS

Section 2.2 uses one response to expose how the joint loss prices a direction. One quadratic expansion gives both the improvement from replacing a single activation and the improvement from jointly optimizing a code. The weighted dual then certifies the attained coding comparisons. Throughout, D has finitely many unit-norm columns and the stated expectations are finite.

D.1 ONE GAP IDENTITY FOR ROWS, BLOCKS, AND COMPLETE CODES

Fix a set of rows $A \subseteq [K]$ and collect the bias and all rows outside A into the residual R^{-A}. For one input define the convex block objective

$$F_A(h) := \frac{1}{2}\|R^{-A} - D_A h\|^2 + \lambda \mathbf{1}^\top h, \qquad h \ge 0.$$

Let h_A^* minimize F_A and define its KKT slack

$$s_A^* := D_A^\top (D_A h_A^* - R^{-A}) + \lambda \mathbf{1}.$$

Optimality gives $s_A^* \ge 0$ and $(h_A^*)^\top s_A^* = 0$. Expanding around h_A^* yields the common identity

$$F_A(g_A) - F_A(h_A^*) = \frac{1}{2}\|D_A(g_A - h_A^*)\|^2 + (s_A^*)^\top g_A. \tag{D.1}$$

Both terms are nonnegative. The scalar and full-dictionary cases provide the comparisons used in the paper.

The gap charged by a response class. Taking $A = [K]$ and averaging over inputs defines the per-input coding optimum at a fixed dictionary,

$$\Phi(D, b) = \mathbb{E} \min_{h \ge 0} \Big[\tfrac{1}{2}\|X - b - Dh\|^2 + \lambda \mathbf{1}^\top h \Big],$$

in which the code for each input is chosen with no constraint tying it to the codes of other inputs. An SAE cannot choose that way, because one encoder has to produce every code, so let $\Psi_{\mathcal{G}}(D, b) = \inf_{g \in \mathcal{G}} \mathcal{L}(D, b, g)$ be the infimum of (5) over a response class $\mathcal{G}$. Since each encoder in $\mathcal{G}$ supplies one feasible code per input, $\Psi_{\mathcal{G}} \geq \Phi$, and we write

$$\Psi_{\mathcal{G}}(D, b) = \Phi(D, b) + A_{\mathcal{G}}(D, b), \qquad A_{\mathcal{G}}(D, b) \geq 0. \tag{D.2}$$

The gap $A_{\mathcal{G}}$ is the objective value the restriction costs at that dictionary, and a trained encoder can sit above $\Psi_{\mathcal{G}}$ by a further optimization error. Two facts about $A_{\mathcal{G}}$ are used in the body. It vanishes wherever the per-input optimum already lies in the class: at an orthonormal dictionary the minimizer of the inner problem is coordinatewise soft thresholding, $h_j^* = (d_j^\top (X - b) - \lambda)_+$ by (D.1) with $A = [K]$, which is an affine-ReLU response, so $A_{\text{aff}} = 0$ there. It is strictly positive at dictionaries whose optimal code needs a response the class cannot compute, which is what Appendix F.3 establishes for the dictionary carrying shared directions. Because the two values are attached to a dictionary rather than to the source, changing the response class changes which dictionary the joint problem finds economical, and this is the route by which an encoder change reaches the learned decomposition.

For a singleton block $A = \{j\}$, unit norm of d_j gives $h_j^* = (d_j^\top R^{-j} - \lambda)_+$. Define the value available along the current direction, the value delivered by its learned response, and their shortfall by

$$U_j = \frac{1}{2}\mathbb{E}[(d_j^\top R^{-j} - \lambda)_+^2], \tag{D.3}$$

$$V_j = \mathbb{E}\left[g_j(d_j^\top R^{-j} - \lambda) - \frac{1}{2}g_j^2\right], \tag{D.4}$$

$$Q_j := U_j - V_j = \mathbb{E}\left[\frac{1}{2}(g_j - h_j^*)^2 + g_j(\lambda - d_j^\top R^{-j})_+\right] \geq 0. \tag{D.5}$$

The full loss equals the loss with row j omitted minus V_j, while Q_j is the exact improvement left unrealized by its learned response. Equation (D.1) proves these identities. If a source model supplies active set S, partitioning the same expectation gives

$$U(d; R) = \sum_T \frac{1}{2}\mathbb{E}\left[(d^\top R - \lambda)_+^2 \mathbf{1}_{\{S=T\}}\right]. \tag{D.6}$$

For an intermediate block, write

$$q_A = D_A^\top R^{-A} - \lambda \mathbf{1}, \qquad H_A = D_A^\top D_A.$$

The value delivered by g_A is $V_A(g_A) = \mathbb{E}[g_A^\top q_A - \frac{1}{2} g_A^\top H_A g_A]$, and the optimal block value U_A replaces g_A by h_A^*. Taking expectations in (D.1) gives $U_A - V_A$. This comparison changes all rows in A together and therefore retains interactions that separate row replacements omit.

For $A = [K]$, every row gap vanishes exactly when every coordinate satisfies

$$z_j > 0 \Rightarrow d_j^\top r = \lambda, \qquad z_j = 0 \Rightarrow d_j^\top r \leq \lambda,$$

where $r = x - b - Dz$. These are the KKT conditions of the complete nonnegative-Lasso problem, so coordinatewise minimality and global code optimality coincide. No condition on dictionary rank is needed for this equivalence: it follows from convexity of the code objective.

Finally, let z^* be an exact complete code, $r^* = x - b - Dz^*$, and $s^* = \lambda \mathbf{1} - D^\top r^*$. The full-block instance of (D.1) is

$$\begin{aligned} &\frac{1}{2}\|x - b - Dg\|^2 + \lambda \mathbf{1}^\top g - \left(\frac{1}{2}\|r^*\|^2 + \lambda \mathbf{1}^\top z^*\right) \\ &\qquad = \frac{1}{2}\|D(g - z^*)\|^2 + (s^*)^\top g. \end{aligned} \tag{D.7}$$

Taking the infimum over a response class $\mathcal{G}$ gives

$$A_{\mathcal{G}}(D, b) = \inf_{g \in \mathcal{G}} \mathbb{E}\left[\frac{1}{2}\|D(g - z^*)\|^2 + (s^*)^\top g\right],$$

and hence $\Psi_{\mathcal{G}} = \Phi + A_{\mathcal{G}}$. The gap depends on the dictionary because both the exact code and its KKT slack depend on D, b.

D.2 One weighted coding dual for fixed-dictionary certificates

The synthetic and public-checkpoint audits use different weights and penalties but the same convex certificate. For samples x_i, positive weights ω_i, row penalties $p \geq 0$, a fixed dictionary D, and a shared bias b, consider

$$\min_{z_i \geq 0,\, b} \frac{1}{n} \sum_{i=1}^{n} \left[\frac{\omega_i}{2} \|x_i - b - Dz_i\|^2 + p^\top z_i \right]. \tag{D.8}$$

Introducing y_i for the reconstructed residual gives the dual

$$\max_{\substack{D^\top y_i \leq p \\ \sum_i y_i = 0}} \frac{1}{n} \sum_{i=1}^{n} \left[y_i^\top x_i - \frac{1}{2\omega_i} \|y_i\|^2 \right]. \tag{D.9}$$

The constraint $D^\top y_i \leq p$ follows from minimizing over each nonnegative code. Minimizing over the shared bias gives $\sum_i y_i = 0$. When the bias is fixed, replace x_i by $x_i - b$ and omit the balance constraint. Any feasible code and dual vector therefore bracket the exact fixed-dictionary optimum without assuming that the numerical solver has converged.

D.3 The radial reference used by the event-code example

For $r \geq 0$ define

$$\psi_\lambda(r) = \min_{s \geq 0} \{ \tfrac{1}{2}(r - s)_+^2 + \lambda s \} = \begin{cases} r^2/2, & r \leq \lambda, \\ \lambda r - \lambda^2/2, & r > \lambda. \end{cases}$$

Put $\nu = x - b$ and $s = \mathbf{1}^\top z$. Unit decoder columns imply $\|Dz\| \leq s$, hence

$$\tfrac{1}{2}\|\nu - Dz\|^2 + \lambda s \geq \tfrac{1}{2}(\|\nu\| - s)_+^2 + \lambda s \geq \psi_\lambda(\|\nu\|).$$

For $\|\nu\| > \lambda$, equality requires $s = \|\nu\| - \lambda$, reconstruction $(1 - \lambda/\|\nu\|)\nu$, and every active column pointing along ν. A finite input law can attain the bound by supplying each required event direction, provided the encoder can isolate its event. Appendix I.3 verifies this for the population it uses.

For an orthogonal event with $m \geq 2$ active strengths all equal to A and $0 < \lambda < A$, optimal generating-dictionary coding costs $m\lambda A - m\lambda^2/2$. Its excess over the radial reference is

$$\delta_m = \lambda A(m - \sqrt{m}) - \tfrac{1}{2}\lambda^2(m - 1) > 0. \tag{D.10}$$

This comparison supplies the event-code proof. It does not replace the response-class analysis needed for continuously varying strengths; the broad radial preference of ideal inference is also studied by Dorrell (2026).

E Magnitude dependence, response cost and recovery

The positive half of Theorem 2 requires a bound over every affine-ReLU response. The negative half needs an explicit cheaper code and the activation budget. We prove the general positive result first, then specify the magnitude laws and establish the negative constructions. The metric and its duplication invariance are in Appendix C.3; geometric reduction bounds are in Appendix I.2.

E.1 One certificate for a class of magnitude laws

Let C_λ be the infimum over all finite-width affine-ReLU SAEs with free biases and unit decoders. Work in orthonormal factor coordinates. Let S be the active-factor support, $p_i = \mathbb{P}(i \in S)$, $s_i = \mathbb{P}(S = \{i\}) > 0$, $P_0 = \mathbb{P}(S = \varnothing)$ and $M_0 = \sum_i p_i \mathbb{E}B_i$. Magnitudes and independent balanced signs are independent of S. Singletons reveal each coordinate without interference. We reweight them by the share of all occurrences they represent, producing a test that every response must satisfy. For the reference source, equality is possible only on signed generating axes; balanced signs give the test zero mean, so a free reconstruction bias cannot satisfy it. The bound is rowwise and therefore independent of width. Accurate reconstruction then fixes the grouped axial strengths,

while the activation budget rules out opposing responses that cancel in reconstruction but remain charged. Define

$$\phi(x) = \frac{p_i}{s_i}\operatorname{sign}(x_i)\mathbf{e}_i \text{ on } S = \{i\}, \qquad \phi = 0 \text{ otherwise.} \tag{E.1}$$

Here $\mathbf{e}_i$ is the standard coordinate vector and $s_i > 0$. This field has $\mathbb{E}\phi = 0$, $\mathbb{E}\phi^\top x = M_0$ and $V_\phi := \mathbb{E}\|\phi\|^2 = \sum_i p_i^2/s_i$. The weights make the singleton measurement reflect all occurrences of each factor. The task is to show that no allowed latent supplies a larger vector under this measurement than its expected activation across the population.

Once the row inequality is verified, for $f = b + \sum_j d_j g_j$ and $M = \sum_j \mathbb{E}g_j$, define the response slack $S_\phi = M - \mathbb{E}\phi^\top f = \sum_j(\mathbb{E}g_j - d_j^\top \mathbb{E}[\phi g_j]) \geq 0$. Unit decoders and $\mathbb{E}\phi = 0$ remove the output bias from this expression. Completing the square gives the exact identity behind the bound:

$$\mathcal{L} = \lambda M_0 - \tfrac{1}{2}\lambda^2 V_\phi + \tfrac{1}{2}\mathbb{E}\|x - f - \lambda\phi\|^2 + \lambda S_\phi. \tag{E.2}$$

The square controls reconstruction, and the slack records the sum of the row inequalities. To recover directions as well as bound the objective, we will need those inequalities to be strict away from the generating axes.

Proposition E.1 (Gaussian-scale-mixture protection)**.** *In r orthonormal factor coordinates, let $x_i = Z_i\sigma_i G_i$, where the G_i are independent standard normals and the strictly positive random vector σ is independent of G and Z. Scales may have arbitrary dependence across coordinates, and $\mathbb{E}\sigma_i^2 < \infty$. Set $p_i = \mathbb{P}(Z_i = 1)$, $s_i = \mathbb{P}(S = \{i\})$ and $P_0 = \mathbb{P}(S = \varnothing)$. Suppose every $s_i > 0$, $P_0 > 0$ and $\|p\|_2 < 1 + P_0$. For raw affine-ReLU encoders with nonnegative outputs, unit decoder columns and free biases, every sequence within $o(\lambda)$ of the infimum over all finite widths recovers $(x_i)_+$ and $(-x_i)_+$ after disjoint grouping in summed mean absolute error as $\lambda \to 0$.*

Proof. Work in factor coordinates and define $h_t(a) = \mathbb{E}(a|G| - t)_+$ for $a, t \geq 0$. Since $h_t(a)/a$ increases with a, putting $R = \|a\|_2$ gives

$$h_t(a_i) \leq \frac{a_i}{R}h_t(R), \qquad h_t(R) \geq \left(\sum_i h_t(a_i)^2\right)^{1/2}. \tag{E.3}$$

The zero-vector case is immediate. The certificate remains $\phi_i = (p_i/s_i)\operatorname{sgn}(x_i)\mathbf{1}_{S=\{i\}}$. For a row $g = (e^\top X - t)_+$ with $t \geq 0$, put $c_i = \mathbb{E}h_t(|e_i|\sigma_i)$ and $z = \mathbb{E}[\phi g]$. Conditional Gaussian closure, followed by (E.3) and Jensen, gives

$$2\mathbb{E}[g \mid S] = \mathbb{E}_\sigma h_t\left(\sqrt{\sum_{i\in S} e_i^2\sigma_i^2}\right) \geq \mathbb{E}_\sigma\|(h_t(|e_i|\sigma_i))_{i\in S}\|_2 \geq \|(c_i)_{i\in S}\|_2. \tag{E.4}$$

Independence of scales from the support makes the same c_i apply to every event. Singleton evaluation gives $z_i = (p_i/2)\operatorname{sgn}(e_i)c_i$. Consequently

$$\mathbb{E}g \geq \nu(z) := \mathbb{E}_S\left(\sum_{i\in S} z_i^2/p_i^2\right)^{1/2} \geq \|z\|_2.$$

The last step averages the vector with coordinates $\mathbf{1}_{i\in S}z_i/p_i$ before taking its norm.

We also need strictness to locate recovered directions. Write $q_i = s_i/p_i$ and $q = \min_i q_i > 0$. Separate singleton events, then apply the triangle inequality to the remaining average:

$$\nu(z) \geq \sum_i q_i|z_i| + \|((1 - q_i)z_i)_i\|_2 \geq q\|z\|_1 + (1 - q)\|z\|_2. \tag{E.5}$$

For the second inequality, $\sum_i(q_i - q)|z_i|$ bounds the norm of $((q_i - q)z_i)_i$, which combines with the remaining vector to give $(1 - q)\|z\|_2$. Thus normalized row moments can touch the unit sphere only on signed axes.

Positive intercepts cannot create another contact. For $g_a = (e^\top X + a)_+$, central symmetry and the empty atom give, almost everywhere for $a > 0$,

$$\frac{d}{da}\mathbb{E}g_a \geq (1 + P_0)/2, \qquad \left\|\frac{d}{da}\mathbb{E}[\phi g_a]\right\|_2 \leq \|p\|_2/2.$$

Their increments have moment-to-mass ratio at most $\rho = \|p\|_2/(1+P_0) < 1$. Combining an increment with the zero-intercept response therefore preserves the signed-axis contact restriction.

From strict response cost to the original coefficients. For $m = \mathbb{E}g > 0$, the normalized moment z/m lies in the compact convex set

$$K = \mathrm{conv}(K_- \cup \rho B_2), \qquad K_- = \{z : q\|z\|_1 + (1-q)\|z\|_2 \le 1\}.$$

Equation (E.5) places nonpositive-intercept moments in K_-; the positive-intercept derivative bound places their added increments in ρB_2. The only contacts of K with the unit sphere are signed coordinate axes. We now use this geometry to recover the original coefficients, including their magnitudes. The certificate bounds the optimum below by $\lambda M_0 - \lambda^2 V_\phi/2$; exact generating responses bound it above by λM_0. For $\mathcal{L} = C_\lambda + o(\lambda)$, these bounds imply $f \to x$ in L^2, $M \to M_0$, and total certificate slack $M - \mathbb{E}\phi^\top f \to 0$. Compactness gives a positive slack per unit of activation on every decoder a fixed distance from the signed axes. Its total activation must vanish. Assign every decoder to the nearest signed axis and sum its original responses into $H_{i,+}$ and $H_{i,-}$. Replacing decoders by their assigned axes changes reconstruction by $o(1)$ in L^1, so $Y_i + b_i \to x_i$, where $Y_i = H_{i,+} - H_{i,-}$.

Symmetry of each marginal gives $\mathbb{E}|x_i - b_i| - \mathbb{E}|x_i| \ge (1-p_i)|b_i|$. Since $\sum_i \mathbb{E}|Y_i| \le M \to M_0$, this forces $b \to 0$. Finally $2\sum_i \mathbb{E}\min(H_{i,+}, H_{i,-}) = M - \sum_i \mathbb{E}|Y_i| \to 0$. Hence

$$\sum_{i,s} \mathbb{E}|H_{i,s} - a_{i,s}| \longrightarrow 0. \tag{E.6}$$

Replacing the summed responses by their original decoder vectors changes each group contribution by $o(1)$ in L^1, because the activation carried off-axis vanishes. Thus the same convergence holds for the decoded groups. □

Singleton probabilities set the margin against off-axis directions, so the guarantee weakens as those events become rare. The conclusion is mean absolute coefficient recovery for a fixed population. Squared-error recovery would additionally require uniform integrability of the response errors across the encoder sequence: finite source moments alone do not prevent a near-optimal encoder from placing unbounded responses on a vanishing-probability region.

Finite optimization precision and certificate margin. The exact generating code costs λM_0. If $\delta_{\mathrm{opt}} = \mathcal{L} - C_\lambda$, the identity gives

$$S_\phi \le \lambda V_\phi/2 + \delta_{\mathrm{opt}}/\lambda, \qquad M_{\mathrm{outside}}(\varepsilon) \le \frac{\lambda V_\phi/2 + \delta_{\mathrm{opt}}/\lambda}{h_\varepsilon}. \tag{E.7}$$

Here $M_{\mathrm{outside}}(\varepsilon)$ is the activation on unit decoders at distance at least ε from every signed axis, and $h_\varepsilon = \inf_d[1 - \sup_{z\in K} d^\top z] > 0$ over those decoders. Compactness and the sphere-contact property prove positivity; summing each outside row's slack proves the second inequality. This isolates three requirements for concentration: a small penalty, excess objective small relative to that penalty, and an informative moment margin. The constant is population-dependent. The bias and cancellation steps above remain necessary to turn concentration into coefficient recovery. Absolute excess $o(1)$ alone is insufficient: the exact generating code has that accuracy even in a population where its expected activation is suboptimal by a fixed amount.

E.2 The exponential source is a special case

A centered Laplace variable with scale μ has the representation $\sqrt{V}G$, where V is exponential with rate $1/(2\mu^2)$ and G is standard normal: both characteristic functions equal $(1+\mu^2 t^2)^{-1}$. Independent exponential magnitudes with balanced signs therefore satisfy Proposition E.1. For any $p \le 1/2$ and $c < 1$, singletons have probability $p(1-c) > 0$ and $\sqrt{2}p < 1 + P_0$, verifying all its assumptions.

For two factors, $V_\phi = 2p/(1-c)$. The certificate and generating soft thresholds give

$$2p\mu\lambda - \frac{p}{1-c}\lambda^2 \le C_\lambda \le G_\lambda = 2p\mu^2(1 - e^{-\lambda/\mu}). \tag{E.8}$$

At the orthogonal generating dictionary, central symmetry makes zero an optimal shared bias: the optimized objective is convex and even in that bias. The upper expression integrates $\frac{1}{2}\min(B,\lambda)^2 + \lambda(B-\lambda)_+$ against the exponential density and sums the two occurrence probabilities. Thus $C_\lambda/\lambda \to 2p\mu$, and the preceding recovery argument proves the independent half of Theorem 2.

E.3 MAGNITUDE LAWS AND WHAT DEPENDENCE PRESERVES

The rotated-code comparison is governed by $\kappa = c\, s_B$, the share of a feature's expected strength held in common with its partner. Independence and the tail exponent do not determine this quantity. If $\kappa > 2 - \sqrt{2}$, the exact rotated competitor and the activation budget exclude recovery. At or below that threshold, this competitor supplies no recovery conclusion; recovery where it holds comes from the response certificate of Appendix E.1. The constructions below place independent finite-variance magnitude laws on both sides of the threshold, including laws with the same tail exponent, so neither property alone predicts the risk. The coupled law of Theorem 2 raises κ, although other source laws can raise it as well. The positive certificate extends to any finite number of orthonormal generating factors meeting its occurrence conditions, and the rotated comparison continues to depend on how strengths co-vary.

Constructing the heavy-tailed members. Student-t coordinates have $\sigma = \sqrt{\nu/U}$ with $U \sim \chi^2_\nu$, giving finite second moments for $\nu > 2$. A Lomax magnitude has survival $\mathbb{P}(B > b) = (1 + b/s)^{-\nu}$. Sample $R \sim \mathrm{Gamma}(\nu, \mathrm{rate} = s)$ and then $B \mid R \sim \mathrm{Exp}(R)$. Integrating over R gives that survival law. Adding a balanced sign gives a Laplace variable; conditional on R, it equals $\sqrt{V}G$ with $V \sim \mathrm{Exp}(\mathrm{rate} = R^2/2)$, since $\mathbb{E}e^{-t^2V/2} = R^2/(R^2 + t^2)$. Again $\nu > 2$ gives the required moment. All generating coefficients are the nonnegative positive and negative parts. Dependence among scales is allowed; correlated Gaussian coordinates conditional on the scales, which can concentrate coactive magnitudes near equality, are outside this construction.

Equal tail exponents, different sharing. For independent magnitudes with survival $\overline{F}$, integrate the events $\{B > t\}$ and $\{B_1 > t, B_2 > t\}$ to obtain

$$\frac{\mathbb{E}\min(B_1, B_2)}{\mathbb{E}B} = \frac{\int_0^\infty \overline{F}(t)^2\, dt}{\int_0^\infty \overline{F}(t)\, dt}.$$

Pareto-I with minimum b_0 has survival one below b_0 and $(b_0/t)^\nu$ above it. Lomax has the survival just given. Their shared fractions are $2(\nu - 1)/(2\nu - 1)$ and $(\nu - 1)/(2\nu - 1)$, respectively. Scaling each to mean one leaves these ratios unchanged. At $p = 1/4, c = 3/4, \nu = 3$:

Active law	Shared fraction	$M_{\mathrm{rot}}/M_{\mathrm{gen}}$	Affine-ReLU conclusion
Pareto-I	.8	.98995	Strict contribution obstruction
Lomax	.4	1.20208	Grouped recovery

The Lomax conclusion follows from Proposition E.1, not just from an unsuccessful rotation. Pareto-I's positive lower endpoint raises the expected minimum of independently sampled amplitudes. Its saving is small but strictly positive; the contribution budget converts it into a nonvanishing asymptotic deficit, without predicting a typical trained error. For $\nu \le 2$, these laws fall outside the finite-second-moment recovery and energy statements. This does not imply that every reconstruction has infinite error: exact reconstruction still has zero error.

For the scale-mixture family of Proposition E.1, dependent scales retain independent Gaussian fluctuations within each pair. If $B_i = \sigma_i|G_i|$ have equal means, conditional Gaussian closure gives

$$\mathbb{E}\max(B_1, B_2) = \frac{1}{2}\mathbb{E}(|\sigma_1 G_1 + \sigma_2 G_2| + |\sigma_1 G_1 - \sigma_2 G_2|) = \sqrt{\frac{2}{\pi}}\mathbb{E}\sqrt{\sigma_1^2 + \sigma_2^2}.$$

The norm inequality gives $\mathbb{E}\max(B_1, B_2) \ge \sqrt{2}\mu$, and therefore $s_B \le 2 - \sqrt{2}$. Correlating scales can make both magnitudes large together without concentrating them sufficiently near equality to make the specified rotation win at $c < 1$. Recovery still follows from the full response bound above.

E.4 THE EVENT COST OF ROTATING A FACTOR PAIR

Take activity p, pair probability $\tau = pc$, balanced signs and support-independent magnitudes of common mean μ. The four rotated signed projections reconstruct exactly. The identity $|x_1 + x_2| + |x_1 - x_2| = 2\max(|x_1|, |x_2|)$ gives

$$M_{\mathrm{rot}} = \sqrt{2}\{2p\mu - \tau\mathbb{E}\min(B_1, B_2)\}. \tag{E.9}$$

Together with $\mathbb{E}\min(B_1, B_2) = \mu - \mathbb{E}|B_1 - B_2|/2$, this proves (7). Writing $\mathcal{D} = 1 - s_B = \mathbb{E}|B_1 - B_2|/(2\mu)$, under independent identical magnitudes $\mathcal{D}$ is the marginal Gini coefficient G, so this rotation saves when $c(1 - G) > 2 - \sqrt{2}$. Under dependence, the marginal Gini can stay fixed while the joint disparity changes. Rejecting this rotated competitor does not by itself prove recovery.

E.5 From activation saving to contribution error

Write $M = \mathbb{E}\sum_j g_j$, $M_{\text{gen}} = \sum_i \mathbb{E}\alpha_i$, $\bar{n} = \sum_i \Pr(\alpha_i > 0)$, and $T = \mathbb{E}\sum_i \alpha_i^2$. For disjoint groups, or even input-dependent fractional assignments with column sums at most one,

$$\sum_i \|v_i(X)\| \le \sum_j g_j(X).$$

On a true occurrence, $\alpha_i \le \|\alpha_i w_i - v_i\| + \|v_i\|$. Summing expectations proves the mean-norm bound in (8). Cauchy–Schwarz restricted to those occurrences also gives

$$\sum_i \mathbb{E}\|\alpha_i w_i - v_i\|^2 \ge \frac{(M_{\text{gen}} - M)_+^2}{\bar{n}}, \qquad \mathcal{R}_{\text{frac}} \ge \frac{(M_{\text{gen}} - M)_+^2}{\bar{n}T}. \tag{E.10}$$

The first quantity is raw squared error; the second is normalized as in the main text. No generating inverse or orthogonal geometry enters either bound.

If an available code has half-squared reconstruction error R_{alt} and expected activation M_{alt}, then $C_\lambda \le R_{\text{alt}} + \lambda M_{\text{alt}}$. A solution with excess $\delta_{\text{opt}} = \mathcal{L} - C_\lambda$ obeys

$$\sum_i \mathbb{E}\|\alpha_i w_i - v_i\| \ge M_{\text{gen}} - M_{\text{alt}} - \frac{R_{\text{alt}} + \delta_{\text{opt}}}{\lambda}. \tag{E.11}$$

An exact competitor with a strict saving in expected activation therefore excludes recovery for every $o(\lambda)$-near-optimal sequence in any encoder class containing it. The argument includes free output biases because the bias is not assigned to individual targets. Applied to the coupled rotated code below, it proves the negative half of Theorem 2.

E.6 One-sided strengths and continuous hierarchies

Let $X = Au + Bv$ with perpendicular unit u, v, $A, B \ge 0$, and

$$B > 0 \implies A > 0. \tag{E.12}$$

For $t > 0$, use unit directions $d_t = (tu + v)/\sqrt{1 + t^2}$, u, and $-u$, with responses

$$g_t = \sqrt{1 + t^2}\, B, \qquad g_+ = (A - tB)_+, \qquad g_- = (tB - A)_+.$$

Since $A = u^\top X$ and $B = v^\top X$, these responses are affine-ReLU computable and reconstruct exactly: $d_t g_t = B(tu + v)$ and $u(g_+ - g_-) = (A - tB)u$. The tilted response supplies tB along the parent direction, leaving only $A - tB$ to the axial pair. Exactly one axial response is positive, so this transfer does not increase the active count. Its expected activation is

$$M(t) = \sqrt{1 + t^2}\mathbb{E}B + \mathbb{E}|A - tB|,$$

so

$$M(t) - M(0) = (\sqrt{1 + t^2} - 1 - t)\mathbb{E}B + 2\mathbb{E}(tB - A)_+. \tag{E.13}$$

The parent response saves the transferred amount linearly in t, whereas normalizing d_t costs only quadratically. The last term pays for the correction when the transfer exceeds the parent strength.

Proposition E.2 (Every integrable hierarchy admits a cheaper exact code)**.** *If* $\mathbb{E}A + \mathbb{E}B < \infty$, $\mathbb{E}B > 0$, *and* (E.12) *holds, then* $M'(0+) = -\mathbb{E}B < 0$. *In particular,* $M(t) < M(0)$ *for every sufficiently small* $t > 0$.

Proof. Divide (E.13) by t. The first term tends to $-\mathbb{E}B$. The second satisfies $0 \le (tB - A)_+/t \le B$ and converges pointwise to zero: where $A > 0$ a sufficiently small transfer is below it, and where $A = 0$ the hierarchy implies $B = 0$. Dominated convergence proves the derivative. □

Fix a tilt with saving $\Delta > 0$. Its exact code has objective $\lambda[M(0) - \Delta]$, so (E.11) gives limiting mean-norm contribution error at least Δ for every sufficiently near-optimal affine-ReLU SAE, including arbitrary finite widths, free biases, and fractional assignments.

For independent exponential strengths of mean μ on joint inputs, $\mathbb{E}(tB - A)_+ = \mu t^2/(1+t)$. The saving per joint occurrence is

$$\mu\left(1 + t - \sqrt{1+t^2} - \frac{2t^2}{1+t}\right),$$

which is $.1192\mu$ at $t = 1/4$, while the equal tilt $t = 1$ costs more than the generating code. The saving therefore does not require nearly equal magnitudes. Without the exact hierarchy, $M'(0+) = -\mathbb{E}B + 2\mathbb{E}[B\mathbf{1}_{A=0}]$: what matters for a small tilt is how much child strength occurs with its parent.

A count penalty on the same hierarchy. Assume additionally finite second moments and densities on the parent-only line and joint plane. Almost surely, $A \neq tB$ on the joint event. Faithful and tilted codes therefore each use one output on parent-only inputs and two on joint inputs. Set $\bar{n} = \Pr(A > 0, B = 0) + 2\Pr(B > 0)$. Both exact codes cost $\beta\bar{n}$. A near-optimal sequence has vanishing reconstruction error and bounded mean count; the width-uniform event-dimension bound in Appendix F.1 gives $\liminf \mathbb{E}n_g \geq \bar{n}$. Hence $C_\beta/\beta \to \bar{n}$, and both exact codes are $o(\beta)$-near-optimal. The tilted code retains a contribution deficit by its activation saving. The count penalty therefore permits both allocations here, whereas L_1 strictly favors an alternative to the generating code.

Removing signs with both isolated events retained. Keep the main occurrence process and independent exponential magnitudes, but let $A = Z_1B_1$ and $B = Z_2B_2$ be nonnegative. The same code is exact and

$$M(t) - M(0) = p\mu\left(\sqrt{1+t^2} - 1 + t - \frac{2ct}{1+t}\right).$$

Its derivative is $p\mu(1-2c)$, so $c > 1/2$ permits a strict saving even though both isolated events and the affine inverse remain. At $p = 1/4, c = 3/4, \mu = 1.5, t = 1/8$, the saving is .0127 against $M_{\text{gen}} = .75$. Child-only inputs now use two outputs, so this last construction is not count-neutral.

E.7 Finite-price comparisons in interacting pair subspaces

For the reference populations, independence gives $s_B = 1/2$, $\kappa = .375$, and $M_{\text{rot}}/M_{\text{gen}} = 1.149$; coupling at $\eta = .95$ gives $s_B = .9236$, $\kappa = .6927$, and ratio .9244. The comparison also survives overlapping pair planes. If each (u_b, v_b) is orthonormal, balanced signs make cross-pair residual inner products average to zero, so the scalar pair costs still add even when different planes overlap.

For an even scalar soft-threshold cost,

$$h_\lambda(t) = \min_{z \geq 0}\{\tfrac{1}{2}(|t| - z)^2 + \lambda z\}, \quad F_\lambda(B, C) = h_\lambda((B+C)/\sqrt{2}) + h_\lambda((B-C)/\sqrt{2}),$$

the mixed derivative is

$$\partial_B\partial_C F_\lambda = \tfrac{1}{2}\left[\mathbf{1}_{B+C<\sqrt{2}\lambda} - \mathbf{1}_{|B-C|<\sqrt{2}\lambda}\right] \leq 0.$$

Continuity of the first derivative extends this inequality across the Huber boundaries. If B, C apply the same monotone quantile function f to correlated normals, Gaussian integration by parts gives $\partial_\rho \mathbb{E}F_\lambda = \mathbb{E}[(\partial_B\partial_C F_\lambda) f'(Z_1) f'(Z_2)] \leq 0$, with smooth approximation at the boundaries. Thus increasing copula correlation lowers this rotated L1 cost at fixed marginals, while singleton and generating costs remain fixed. Exponential and the tested lognormal quantiles satisfy the required integrability. This is a code comparison, not a prediction that training varies monotonically.

If the factor columns are linearly independent, a left inverse makes both codes affine-ReLU realizable. For $\Delta = M_{\text{gen}} - M_{\text{rot}} > 0$ and $\bar{n} = \mathbb{E}|S|$, the activation argument gives

$$\liminf_{\lambda \to 0} \mathcal{E}_{\text{frac}} \geq \Delta^2/\bar{n}$$

for $o(\lambda)$-near-optimal sequences; division by $T = \mathbb{E}\sum_i \alpha_i^2$ gives the normalized bound. Cross-pair nonorthogonality therefore preserves the obstruction. A globally overcomplete union has no affine inverse, so Appendix G.7 instead tests trained encoders directly.

E.8 Changing dependence while preserving exact feasibility

To construct the second population in Theorem 2, pair strengths must be similar without collapsing onto $B_1 = B_2$. That collapse removes a dimension of pair variation and changes which functions reconstruct exactly. We instead perturb a shared quantile and mix in independent draws, preserving continuous two-dimensional variation and each marginal law. For any $c > 2 - \sqrt{2}$, taking η close to one and ϵ small makes $s_B > (2 - \sqrt{2})/c$, so the same argument applies; the paper's numerical examples use c = 3/4.

Let U be uniform on $(0, 1)$ and J uniform on $[-\varepsilon, \varepsilon]$, independently, with $0 < \varepsilon < 1/2$. Put

$$B_1 = -\mu \log(1 - U), \qquad B_2 = -\mu \log\big(1 - ((U + J) \bmod 1)\big). \tag{E.14}$$

A circular shift preserves uniform measure, so both marginal magnitudes remain exponential. Draw from this coupling with probability η and from the independent law with probability $1 - \eta$, independently of supports and signs. For $\eta < 1$ the joint density is strictly positive throughout the positive quadrant. Thus both examples have variable magnitudes on full pair events.

Exact feasibility is unchanged by the coupling. For $\eta < 1$, the independent and mixed laws are mutually absolutely continuous on the positive pair quadrant; singleton laws, the zero atom, and sign supports also agree. Their pushforwards therefore have the same null sets. A fixed measurable f reconstructs exactly under one law precisely when it does under the other, including at every finite affine-ReLU width. Expected coefficient and nonzero approximation costs may still differ because the two densities weight responses differently.

For a fixed positive shift v, split the quantile integral at $1 - v$. Monotonicity of the exponential quantile gives

$$\frac{\mathbb{E}|B_1 - B_2|}{2\mu} = \int_{1-v}^{1} -\log(1 - u)\, du - \int_{0}^{v} -\log(1 - u)\, du = -v \log v - (1 - v) \log(1 - v).$$

Negative shifts give the same value. Averaging over J, and then over the mixture, yields

$$\mathcal{D} = \frac{1 - \eta}{2} + \eta d_\varepsilon, \qquad d_\varepsilon = \frac{1}{2} - \frac{\varepsilon \log \varepsilon}{2} + \frac{(1 - \varepsilon)^2 \log(1 - \varepsilon)}{2\varepsilon}. \tag{E.15}$$

At $\varepsilon = .02$, $d_\varepsilon = .0540532$. With $c = 3/4$, the rotated comparison becomes cheaper for $\eta > .630229$. Evaluating (E.15) at the two coupling strengths used in the paper, with $p = 1/4$ and $\mu = 1.5$, gives

η	s_B	$M_{\mathrm{rot}}/M_{\mathrm{gen}}$	M_{rot}
.75	.834460	.971674	.728756
.95	.923649	.924374	.693281

against $M_{\mathrm{gen}} = .75$. Figure 5 uses the second row. The remaining worked comparisons use $\eta = .75$, whose exact activation saving is .0212443. Because exact reconstruction costs λM_{rot} while $G_\lambda = \lambda M_{\mathrm{gen}} + O(\lambda^2)$, this gap yields a strict order-λ advantage.

The margin need not rely on a visually narrow cloud. At $c = 3/4$ and full coupling, the activation inequality remains strict for quantile width $\varepsilon < .123340\ldots$. To retain the same null sets as the independent law, take $\eta = .99$ and $\varepsilon = .10$: then $\mathcal{D} = .191535\ldots$ and $M_{\mathrm{rot}}/M_{\mathrm{gen}} = .985460\ldots$. Its 1.45% saving is an exact-code comparison; the training measurements in Figure 5 retain their original $\varepsilon = .02$ population.

In both populations, each $a_{i,s}$ has an atom $1 - p/2$ at zero and an exponential positive part with probability $p/2$. Balanced signs keep the input covariance $2p\mu^2 I_2$ and the support law fixed for every η, although coefficient covariance changes: $\mathrm{Cov}(a_{1,s}, a_{2,t}) = [pc\mathbb{E}(B_1 B_2) - p^2\mu^2]/4$. Signs cancel the input cross moment while coupling remains visible in nonnegative coefficients. Thus the laws are statistically different even though the exactly reconstructing functions coincide and, by Proposition C.2, the source Gram is unchanged. If $\mathcal{L} \le C_\lambda + o(\lambda)$, then $\lambda M \le \lambda M_{\mathrm{rot}} + o(\lambda)$. Equation (E.10), with total expected activity $2p$, gives limiting assignment error at least $(M_{\mathrm{gen}} - M_{\mathrm{rot}})^2/(2p)$; its grouped extension gives $\liminf \sum_i \mathbb{E}\|\alpha_i w_i - v_i\| \ge M_{\mathrm{gen}} - M_{\mathrm{rot}} > 0$. These statements concern original-scale contributions and disjoint groups; shared rows may still support coordinated readouts or edits.

Vanishing objectives retain the recovery incentive. Both class infima vanish with λ, so divide by λ to expose the surviving difference. Equation (E.8) places the independent optimum between $.75 - \lambda$ and $1.125(1 - e^{-\lambda/1.5})/\lambda$, both tending to .75; the coupled rotated code gives $C_\lambda/\lambda \le .728756$. Hence the small-penalty optima differ by at least .02124 in the limit, without requiring identification of the coupled optimizer, and the contribution budget turns this saving into a persistent recovery deficit.

The saving at the experimental penalty. The exact rotated code compares leading costs as $\lambda \to 0$. At a finite penalty, soft thresholding makes a stronger feasible competitor. A scalar magnitude $t \ge \lambda$ saves $\lambda^2/2$ relative to its unshrunk response; smaller magnitudes save a nonnegative amount. A singleton supplies two rotated magnitudes $B/\sqrt{2}$, while a pair supplies at least the combined magnitude $(B_1 + B_2)/\sqrt{2}$. Exponential marginal tails therefore give

$$L_{\mathrm{rot,soft}} \le \lambda M_{\mathrm{rot}} - \frac{\lambda^2}{2} p(4 - 3c) e^{-\sqrt{2}\lambda/\mu}. \tag{E.16}$$

The coefficient $p(4 - 3c)$ counts two scalar savings on singleton events of total probability $2p(1 - c)$ and one on pair events of probability pc. The encoder $(\pm v_k^\top X - \lambda)_+$ realizes this comparison with four unit decoders and zero output bias. At the central parameters and $\lambda = .1$, its cost is at most .0708849, against generating inference $G_\lambda = .0725546$, certifying a saving of at least 2.30%.

Checking the fixed-marginal cost comparison. At $p = .25, c = .75, \mu = 1.5, \lambda = .1$, we enumerate supports and signs and integrate exponential quantiles by order-80 product Gauss–Legendre quadrature. The coupled calculation uses $\eta = .75, \varepsilon = .02$ and includes circular wrap; 200,000 independent samples check it. We average paired differences from generating soft-threshold cost and add its exact expectation $G_\lambda = 2p\mu^2(1 - e^{-\lambda/\mu})$; paired standard errors use those differences. This compares specified encoders. Figure 5 instead trains width-eight models with the calibration/test grouping of Appendix G.6.

For the finite-price crossing between the two shrunken reference codes, Sobol quadrature at 2^{18} and 2^{20} points changes costs by at most 1.6×10^{-6} and the crossing by less than 6×10^{-5}. Mixture costs are affine in η; at $\lambda = .1$, their crossing is $\eta = .61268$. The contribution inequality supplies the further step from a feasible saving to the class-wide obstruction. Training across the crossing is reported separately in Appendix G.8.

F Encoder computation and activity pricing

We first prove the width-uniform count bound and individual recovery result. We then give the richer encoder's exact responses and training protocol, followed by the finite-price and TopK comparisons. All objectives use unit decoder columns, nonnegative outputs, and the bias conventions stated locally.

F.1 A count bound that survives increasing width

A finite dictionary cannot reconstruct a continuously varying s-dimensional event with fewer than s active rows. Increasing width could defeat this argument by supplying successively finer reconstruction spaces. Independently gated affine responses impose an additional cost on this refinement.

An affine gate that fires at an interior input must also fire on one side of a hyperplane through that input. We measure the least probability any such active region must contain. For a law P absolutely continuous on an s-dimensional affine space L, define its halfspace depth at $x \in L$ by $\delta_P(x) = \inf_{H \ni x} P(H)$, over open halfspaces in L. Almost every x is interior to the convex support, and every halfspace through such an interior point has positive probability. Absolute continuity makes halfspace probabilities continuous in the unit normal; compactness of the unit sphere gives $\delta_P(x) > 0$ almost surely.

Lemma F.1 (Count bounds the rows that reach an interior region)**.** *Suppose P is an event law of probability $\pi > 0$ and an affine-ReLU or affine-Jump encoder has $\mathbb{E}n_g \le \bar{n}$. At most $\bar{n}/(\pi\delta)$ rows can fire anywhere in $K_{\delta,R} = \{x \in L : \|x\| \le R,\ \delta_P(x) \ge \delta\}$.*

Proof. Every such row fires on a halfspace of conditional probability at least δ, and therefore contributes at least $\pi\delta$ to mean count. Sum over rows. Constant active gates have conditional probability one and satisfy the same bound. □

Now consider a finite mixture of event laws absolutely continuous on spaces of dimensions $s(S)$, with zero-dimensional atoms allowed. For a sequence of affine-Jump SAEs with bounded mean count and $\mathbb{E}\|X - f(X)\|^2 \to 0$,

$$\mathbb{P}(n_g < s(S) \mid S) \longrightarrow 0, \qquad \liminf \mathbb{E}n_g \geq \mathbb{E}s(S). \tag{F.1}$$

Indeed, on $K_{\delta,R}$ the lemma bounds the available rows independently of total width. Reconstructions using fewer than s of them lie in a bounded number of affine spaces $b + \operatorname{span}(D_J)$ of dimension below s. On a bounded ball, an absolutely continuous law assigns uniformly vanishing probability to an ϵ-tube around any such space: truncate its density at a finite height, bound tube volume, then remove the truncation. Projecting an ambient affine space into L only decreases distances. The probability of an undercount is bounded by these tube probabilities plus the conditional squared error divided by ϵ^2. Send the sequence index to infinity, then $\epsilon \to 0$, and finally $\delta \downarrow 0$, $R \uparrow \infty$. This proves (F.1) without bounds on encoder slopes or bias.

Fixed continuous noise changes the limit. For an absolutely continuous d-dimensional input with finite second moment, signed coordinate ReLUs reconstruct exactly with d active outputs almost surely. Jump thresholds β add only $O(\beta^2)$ reconstruction error. Applying (F.1) to that full-dimensional law gives $C_\beta/\beta \to d$ and $\mathbb{E}|n_g - d| \to 0$. Thus a fixed nonzero continuous noise distribution changes the small-price limit from noiseless activity to ambient dimension. This follows from the same count bound; it does not locate a finite-price noise threshold.

F.2 Individual recovery on the raw paired-ray source

The coverage bound determines how many rows reconstruction needs. We now identify which rows carry those responses. Isolated occurrences fix their scale; opposite occurrences test whether a row also reads an unrelated factor. The following assumptions supply precisely these tests.

Write $X = \sum_{i=1}^{r} x_i w_i$, with finitely many linearly independent unit w_i, and define the nonnegative coefficients $\alpha_{i,+} = (x_i)_+$, $\alpha_{i,-} = (-x_i)_+$ on $w_i, -w_i$. Let S index the nonzero signed coefficients and $\bar{n} = \mathbb{E}|S|$. Assume a positive empty atom, positive probability of every signed singleton, and a singleton magnitude density continuous and strictly positive on $(0, h_{i,s})$ for some $h_{i,s} > 0$ on each signed ray. Conditional magnitudes on every occupied signed support are absolutely continuous in all its active coordinates. They can be dependent. Assume $\mathbb{E}\|X\|^2 < \infty$.

Theorem 3 (JumpReLU case; the ReLU extension is proved below)**.** *Let C_β be the infimum of J_β over all finite-width affine-Jump SAEs with free bias and unit decoders. Then $C_\beta/\beta \to \bar{n}$. Every sequence with $J_\beta \leq C_\beta + o(\beta)$ has $b \to 0$, $\mathbb{E}|n_g - |S|| \to 0$, and an injective assignment $\pi(i,s)$ of rows to the $2r$ signed features such that*

$$\max_{i,s} \|d_{\pi(i,s)} - s w_i\| \to 0, \qquad \sum_{i,s} \mathbb{E}\big(g_{\pi(i,s)} - \alpha_{i,s}\big)^2 \to 0. \tag{F.2}$$

Widths and all parameters may vary with β. Orthogonality is unnecessary.

Proof. First match the generating activity. The factor coordinates have linear readings on the input span. Use their positive and negative parts with threshold β. At most r coefficients, each smaller than β, are omitted per input, so squared reconstruction error is at most $r^2\beta^2$. This gives $C_\beta \leq \beta\bar{n} + O(\beta^2)$. A near-optimal sequence consequently has vanishing reconstruction error and bounded mean count. Each signed event has dimension $|S|$, so (F.1) gives vanishing undercounts and $\liminf \mathbb{E}n_g \geq \bar{n}$. The objective upper bound gives $\limsup \mathbb{E}n_g \leq \bar{n}$. Since $\mathbb{E}(n_g - |S|)_+ = \mathbb{E}n_g - \bar{n} + \mathbb{E}(|S| - n_g)_+$, excess counts also vanish. Approximate minimizers establish the same lower bound on C_β/β. At the empty atom, any active row would give a fixed positive excess-count probability. Thus eventually $g(0) = 0$, and vanishing error there forces $b \to 0$.

Then identify a row on each ray. Fix a signed singleton tu, $u = sw_i$. Choose a closed interval $[t_0, t_1] \subset (0, h_{i,s})$. Halfspace depth is bounded below there. Only a bounded number of rows can reach this interval, and exactly one fires on almost all of it. Some row is therefore the sole active row on a set of probability bounded below within $[t_0, t_1]$. Its reconstruction there is $b + d(\gamma t + a_j)$, with $\gamma = e_j^\top u$. The target stays away from zero, so vanishing error and unit decoder norm force $d \to u$. The singleton density is bounded above on $[t_0, t_1]$; these sets consequently have Lebesgue measure and variance bounded below. Affine least squares on them gives $\gamma \to 1$ and $a_j \to 0$.

Weak occurrences remove a positive threshold. The row fires at some points of $[t_0, t_1]$, so its threshold is bounded. If a subsequential limit $\theta_* > 0$ remained, the row would miss a fixed lower interval. The same argument supplies another row there with slope one and intercept zero. The lower row's threshold is below the interval where the first row fires. Both would then fire on a fixed subinterval of the original interval, contradicting vanishing excess count. Hence $\theta_j \to 0$.

Opposite occurrences remove cross-factor readings. For another factor ℓ, any nonvanishing $e_j^\top w_\ell$ has a sign for which the row fires throughout a fixed interval of the corresponding signed singleton. Its decoder tends to u, a different direction. Either it fires alone and leaves positive error, or an additional row causes excess count. Both are impossible, so all cross-factor readings vanish. The affine score thus converges in L^2 to sx_i. The pointwise bound $|q\mathbf{1}_{\{q>\theta\}} - (q)_+| \le \theta$ and the Lipschitz property of ReLU give response convergence. Repeat for finitely many signed features. Their distinct limiting directions make the matching injective. Applying this argument to every subsequence proves (F.2) for the minimum matching error of the full sequence. □

Under the same assumptions, vanishing reconstruction error and $\mathbb{E}n_g \to \bar{n}$ suffice for individual recovery: the coverage bound removes undercounts, the mean removes persistent overcounts, and singleton tests identify the rows. The result permits dependent magnitudes and invertible nonorthogonal geometry, although rare singleton events weaken the tests. It controls matched rows, not every unused response; cancelling rows $d, -d$ can have vanishing expected count while retaining nonvanishing second moments.

Holding the encoder fixed isolates the price. The conclusion also holds for affine-ReLU encoders trained with J_β. Their exact generating code bounds the infimum by $\beta\bar{n}$, and Lemma F.1 again removes undercounts and excess counts. A singleton interval identifies a slope-one, zero-intercept row on each ray; opposite singletons remove cross-readings, and ReLU Lipschitzness gives coefficient convergence. No threshold-limit step is needed. Thus changing the charged resource yields recovery with the encoder class fixed, although optimization must still handle the discrete count.

Occurrence recovery and the role of the count budget. Theorem 3 also recovers the occurrence pattern of the matched rows. For any signed feature and $t > 0$,

$$\mathbb{P}(\alpha_i > 0, g_{\pi(i)} = 0) \le \mathbb{P}(0 < \alpha_i \le t) + \mathbb{E}(g_{\pi(i)} - \alpha_i)^2/t^2.$$

First let the price vanish at fixed t, then let $t \downarrow 0$. The first probability tends to zero by continuity of probability on the shrinking sets, without an additional density assumption. Summing over the fixed finite set of features gives vanishing expected false negatives FN. If FP counts matched false positives, unmatched active rows only increase total activity, so

$$FP \le |n_g - |S|| + FN, \qquad \mathbb{E}|\widehat{S} \triangle S| \le \mathbb{E}|n_g - |S|| + 2\mathbb{E}FN \to 0.$$

Here $\widehat{S}$ is the generating-feature index set declared positive by its matched rows. Rare unmatched responses remain compatible with this conclusion.

L1 coefficient convergence alone does not force raw-positive support recovery. On the protected signed source, begin with exact positive-part readings and replace one response $(x_i)_+$ by $(x_i + \lambda)_+$. The coefficient change is at most λ, so reconstruction cost and the change in its L1 penalty are $O(\lambda^2)$. The certificate bounds the infimum below by $\lambda M_0 - O(\lambda^2)$, making this sequence $o(\lambda)$-near-optimal. Yet the changed response is positive whenever factor i is absent, a fixed-probability event. A count penalty charges that event regardless of its small amplitude. A positive detection threshold asks a different support question and must be specified when reporting support accuracy.

Why the signed singleton assumptions matter. Positive-only bounded magnitudes admit a different allocation even with exact per-input counts. Let $X = (\alpha_1, \alpha_2)$, with each nonzero coefficient in $[1, 1.1]$. Use decoders $e_1, e_2, (2, -1)/\sqrt{5}, (-1, 2)/\sqrt{5}$, scores $\alpha_1 - 1.1\alpha_2$, $\alpha_2 - 1.1\alpha_1$, $\sqrt{5}(2\alpha_1 + \alpha_2)/3$, $\sqrt{5}(\alpha_1 + 2\alpha_2)/3$, and thresholds $.5, .5, 1.95, 1.95$. Singleton rows fire only alone; the mixed rows fire only on pairs and reconstruct them exactly with positive coefficients. Every input has the generating count, while the singleton responses miss every pair. With continuous event magnitudes, (F.1) makes this exact code $o(\beta)$-near-optimal over arbitrary finite widths. An invertible map T preserves the construction by replacing d_j with $Td_j/\|Td_j\|$, and multiplying its transformed score and threshold by $\|Td_j\|$. Thus this boundary concerns the identifying events, including opposite singleton tests, rather than orthogonality.

F.3 A RICHER ENCODER CAN TRANSFER STRENGTH WITHOUT THE SINGLETON COST

Both recovery arguments use affine active regions. We now implement the response from Section 3.2 that transfers the smaller strength to a summed direction, first for arbitrary nonparallel directions and then on the independent exponential source where one layer already computes the true strengths.

Proposition F.2 (Sharing two readable contributions)**.** *Let u, v be nonparallel unit directions with nonnegative response functions $a_u(X), a_v(X)$. Suppose the response class can realize $\min(a_u, a_v)$ and the remaining positive parts. Adding the unit decoder $(u+v)/\|u+v\|$ permits exact reconstruction of $a_u u + a_v v$ with total activation on each input*

$$a_u + a_v - (2 - \|u + v\|) \min(a_u, a_v). \tag{F.3}$$

Its per-input activity does not exceed $\mathbf{1}_{\{a_u>0\}} + \mathbf{1}_{\{a_v>0\}}$. The expected activation saving is strictly positive whenever $\mathbb{P}(a_u > 0, a_v > 0) > 0$ and the original activation is integrable.

Proof. Use (9). The residual coefficients are $(a_u - a_v)_+$ and $(a_v - a_u)_+$, so at most one is positive. A singleton is unchanged; an unequal pair uses one summed-direction row and one residual row; an equal pair uses only the summed-direction row. Summing their magnitudes proves (F.3). Nonparallel unit directions have $0 < \|u + v\| < 2$. □

Other generating responses can be left unchanged. No independence or minimum coactivity is required; $\min(a_u, a_v)$ controls the saving. For these two exact codes, any positive L1 weight favors the transfer even after adding a nonnegative count penalty, because count never increases. This comparison does not identify the optimum of that combined objective. The premise that the coefficients are readable also remains essential in an overcomplete source.

Applying the transfer to an existing code. Proposition F.2 applies equally to learned responses g_j, g_k. Their minimum is observable from the encoder outputs, so this version needs no generating inverse. Replace the two responses by their positive differences and add $\|d_j + d_k\| \min(g_j, g_k)$ on the normalized sum direction, keeping every other response and the bias fixed. Reconstruction is unchanged, including any existing error. If a response class is closed under this operation and adding its output, no finite L1 optimum in that union of widths can have two nonparallel rows coactive with positive probability. The improvement is strict at every positive L1 price. A fixed row budget or a class unable to compute these differences blocks this conclusion.

Proposition F.3 (Expressivity removes the independent-law protection)**.** *On the independent two-factor source, an eight-output multilayer-ReLU encoder with unit decoders and nonnegative responses reconstructs exactly with expected activation*

$$M_{\text{share}} = 2p\mu - (2 - \sqrt{2})pc\mu/2, \qquad \mathbb{E}n_g = 2p. \tag{F.4}$$

At $p = 1/4, c = 3/4, \mu = 1.5$, its expected activation is $.667624\ldots$, below $M_{\text{gen}} = .75$ by $.082376\ldots$. For any enlarged L1 class containing this encoder, every sequence within $o(\lambda)$ of its class infimum has limiting summed disjoint-group contribution error at least $M_{\text{gen}} - M_{\text{share}}$ as $\lambda \to 0$.

Proof. Write $X = x_1 w_1 + x_2 w_2$ with orthonormal w_1, w_2, and use their linear readings $x_i = w_i^\top X$. For $s, t \in \{-1, 1\}$ take the four diagonal rows and four axial rows

$$d^{\text{diag}}_{s,t} = (sw_1 + tw_2)/\sqrt{2}, \qquad g^{\text{diag}}_{s,t} = \sqrt{2}\min\{(sx_1)_+, (tx_2)_+\}, \tag{F.5}$$

$$d^{\text{axis}}_{i,s} = sw_i, \qquad g^{\text{axis}}_{i,s} = (sx_i - |x_{3-i}|)_+, \quad i = 1, 2. \tag{F.6}$$

All responses are nonnegative. With $A = |x_1|$ and $B = |x_2|$, the diagonal supplies $\min(A, B)$ in each occupied signed coordinate; the larger coordinate's axis supplies $|A - B|$. The sum is X, with total activation

$$m_{\text{share}}(X) = \max(A, B) + (\sqrt{2} - 1)\min(A, B). \tag{F.7}$$

The formulas are finite ReLU computations: $|x| = (x)_+ + (-x)_+$ and $\min(a, b) = a - (a - b)_+$ for $a, b \geq 0$. The latter is already nonnegative, so an additional ReLU leaves it unchanged. Hidden intermediate values are not decoder responses and are not charged by the SAE output penalty.

On singletons the code uses one axis; on a pair with unequal positive magnitudes it uses one diagonal and one axis. Equal magnitudes have zero probability under the independent continuous law. Thus its count agrees almost surely with the generating count. The pointwise activation saving is $(2-\sqrt{2})\min(A,B)$, which vanishes off the pair event. That event has probability pc, and independent exponential magnitudes satisfy $\mathbb{E}\min(B_1,B_2)=\mu/2$, proving (F.4). Writing the same saving through $s_B=\mathbb{E}[\min(B_1,B_2)\mid Z_1=Z_2=1]/\mu$ keeps it available for any magnitude law on this occurrence process:

$$\frac{M_{\text{share}}}{M_{\text{gen}}}=1-\frac{(2-\sqrt{2})\,c\,s_B}{2}, \tag{F.8}$$

which is the form used in Section 3.2. Only the sign and occurrence structure enters, so the dependence between magnitudes reaches this comparison through s_B alone, exactly as it reaches (7).

Finally the enlarged class infimum is at most λM_{share}. Near-optimality therefore gives $\mathbb{E}\sum_j g_j \le M_{\text{share}}+o(1)$. Applying the contribution budget (8) to the four nonnegative generating rays gives the asserted deficit for every disjoint grouping. □

Why sharing violates the affine certificate. The diagonal responses vanish on all singletons. The positive axial response $(x_1-|x_2|)_+$ has singleton certificate value $p\mu/2$ but mean response $p(1-c)\mu/2+pc\mu/4$. Their ratio is $1/(1-c/2)=1.6$ at $c=3/4$, violating dual feasibility for the singleton field (E.1). One affine score must use the same cross-slope in both signs of the other coordinate; the new response switches that slope to subtract $|x_2|$. The certificate's cross-context charge is precisely what this additional operation avoids. The total activation (F.7) is itself a convex function of the input; what the extra layer supplies is the individual sharing responses that realize it.

The resulting affine-encoder gap. Fix the eight-column dictionary D_8 containing the four signed axes and four signed diagonals, and set the decoder bias to zero. The singleton certificate in Appendix E.1 applies to every affine-ReLU encoder using this dictionary, so

$$\Psi_{\text{aff}}(D_8,0)\ge\lambda M_{\text{gen}}-\frac{\lambda^2 V_\phi}{2}. \tag{F.9}$$

Exact sparse inference is not restricted to affine-ReLU responses. It may use the summed-direction construction above separately on each input, giving

$$\Phi(D_8,0)\le\lambda M_{\text{share}}. \tag{F.10}$$

Subtracting gives

$$A_{\text{aff}}(D_8,0)\ge\lambda(M_{\text{gen}}-M_{\text{share}})-\lambda^2V_\phi/2. \tag{F.11}$$

For the present source, $V_\phi=2$, so the lower bound is positive whenever $0<\lambda<M_{\text{gen}}-M_{\text{share}}=.082376\ldots$. The generating dictionary has zero encoder gap because optimal soft-thresholded coding is affine-ReLU. Its best objective, allowing a free shared bias, is $G_\lambda=2p\mu^2(1-e^{-\lambda/\mu})$: symmetry and convexity make zero bias optimal. At $\lambda=.1$, $G_\lambda=.07255464$ and the feasible summed-direction construction costs .06676238. This finite-price advantage compares two dictionaries; the positive lower bound on $A_{\text{aff}}(D_8,0)$ applies over its separate small-price range.

What unchanged count establishes. Correct counts need not identify generating responses in the richer class. Fix width $m\ge 8$ and minimize the count objective. The exact summed-direction code costs $2p\beta$, forcing vanishing reconstruction error for any no-costlier sequence. On a continuous pair event, reconstructions with at most one active row lie in at most m affine lines, whose shrinking neighborhoods have uniformly vanishing probability; singletons similarly require one row. Hence $\liminf \mathbb{E}n_g\ge 2p$, the infimum divided by β tends to $2p$, and this code is $o(\beta)$-near-optimal despite missing pair coefficients in its singleton rows. This uniform line-covering argument fails when width grows, so equality of two feasible counts alone gives no arbitrary-width conclusion.

Training the enlarged class.

At $\lambda=.1$, soft-thresholding the true strengths gives contribution NMSE .0021, while the exact summed-direction code scores .0938 after splitting its diagonal latent evenly. These analytic references locate the trained entries in Table 3 without identifying them with either construction.

We train $2\to 16\to 8$ ReLU encoders at $p=.25, c=.75, \lambda=.1$, under independence or coupling $\eta=.95, \varepsilon=.02$; only the eight outputs are penalized. The dictionary is either two frozen copies of

Table 3: **Available generating responses and the representation reached by training.** Median contribution NMSE over 16 starts, with eight output directions in every condition and $\lambda = .1$. These are separate runs on fresh batches; Figure 5 uses a separate 32-start fixed-pool cohort. Appendix F.3 gives the architecture, objectives, and recoding control.

Encoder and dictionary	Independent	Coupled
One-layer, trainable from random start	.00209	.489
Two-layer, generating dictionary frozen	.00209	.00210
Two-layer, trainable from generating start	.0377	.405
Two-layer, trainable from random start	.0498	.388

the generating rays, the same trainable initialization, or random trainable unit rows. Each condition and the width-eight affine baseline use 16 starts, 8,000 fresh-batch updates, 20,000 assignment inputs, and 50,000 test inputs. Direct and quadratic scores agree within 9.2×10^{-9}.

Median independent/coupled objectives are $.07339/.06770$ for trainable affine encoders, $.07339/.07335$ with frozen generating directions, $.07093/.06523$ for two-layer generating starts, and $.06989/.06522$ for two-layer random starts. Frozen two-layer encoders recover under both laws, while trainable dictionaries lower the objective and increase contribution error. Their activation-weighted angles from the nearest generating ray are $10.25°$ independently and $36.08°$ under coupling, tying the objective decrease to changed directions rather than unavailable generating responses.

We also optimize nonnegative L1 codes at each saved dictionary by enumerating supports of at most two rows; dual points certify gaps below 10^{-6} and an independent solver agrees. Whole-row assignments use separate 3,000-example calibration and test samples. Recoding random trainable dictionaries changes median error $.0587 \to .1043$ independently and $.643 \to .548$ under coupling. Better inference can therefore move contribution error either way; these finite runs do not certify global optimization of the enlarged class.

F.4 The energy scale of a finite count price

Under a count penalty, a continuously varying pair costs two active outputs in either frame, while an isolated factor costs one output in the generating frame and two after rotation. The generating code therefore wins as $\beta \to 0$. At finite price, however, a threshold can drop a weak rotated contrast and recover part of the pair saving; this explains the calibrated $\beta = .16$ comparison in Appendix G.5. The calculation below locates that boundary between the two codes, without claiming they exhaust all encoders.

Take two balanced signed factors with uniform $[1, 2]$ strengths, separately from the exponential theorem. Let each factor occur with probability p and jointly with probability $\tau = pc$, then rotate by $45°$ and hard-threshold at $\sqrt{2\beta}$. For $p < 1/2$ and $0 < \beta < 1/8$, the optimized generating cost and feasible rotated cost are

$$G_\beta = 2p\beta, \qquad H_\beta = (4p - 3\tau)\beta + \tau\psi_\beta, \qquad \psi_\beta = \mathbb{E}\min\{(B_1 - B_2)^2/4, \beta\}. \tag{F.12}$$

Subtracting the costs also gives $H_\beta - G_\beta \geq p(2 - 3c)\beta$ in the stated range, since $\psi_\beta \geq 0$. Thus at $c \leq 2/3$ this rotated competitor cannot win however small the contrast becomes. All singleton projections and the common pair coordinate exceed threshold. Their event counts supply the first term in H_β; the contrast supplies ψ_β. For the generating dictionary, each coordinate bias outside $(-\sqrt{2\beta}, \sqrt{2\beta})$ costs at least $(1 - p)\beta$ on its zero atom. Inside this interval, every nonzero magnitude costs β, and bias adds squared error at zero. Thus for $p < 1/2$ its global bias optimum is zero. The rotated comparison is feasible at zero bias, so a strict $H_\beta < G_\beta$ remains valid when all biases are optimized.

To keep marginal laws fixed, with probability η draw U uniform on $[0, 1]$ and V uniform on $[-\varepsilon, \varepsilon]$, then set $(B_1, B_2) = (1 + U, 1 + (U + V) \bmod 1)$. Otherwise draw independent uniforms. In the coupled component, $v = |V|$ is uniform on $[0, \varepsilon]$: with probability $1 - v$ the difference is v, and with probability v wrapping makes it $1 - v$. Consequently both code costs are obtained by one-dimensional

integrals:

$$\psi_{\text{ind}} = \int_0^1 2(1-v)\min(v^2/4, \beta)\, dv,$$
$$\psi_{\text{coup}} = \varepsilon^{-1} \int_0^{\varepsilon} \big[(1-v)\min(v^2/4, \beta) + v\min((1-v)^2/4, \beta)\big] dv. \tag{F.13}$$

Here $\psi_\beta = (1-\eta)\psi_{\text{ind}} + \eta\psi_{\text{coup}}$.

Set $\beta = b\varepsilon^2$ and hold $b > 0$ fixed. Wrapping has probability $O(\varepsilon)$ and the clipped cost divided by β is bounded by one. Removing that vanishing term and rescaling v gives

$$\frac{\psi_{\text{coup}}}{\beta} \longrightarrow \chi(b) = \int_0^1 \min\{v^2/(4b), 1\}\, dv = \begin{cases} 1 - \frac{4}{3}\sqrt{b}, & b \le 1/4, \\ 1/(12b), & b \ge 1/4. \end{cases} \tag{F.14}$$

The independent component tends to one. Thus $(G_\beta - H_\beta)/\beta \to p[c\eta(1-\chi(b)) - 2(1-c)]$. A crossing exists when $c\eta > 2(1-c)$ and solves $\chi(b) = 1 - 2(1-c)/(c\eta)$. At $c = .75$, $\eta = .99$, it is $b = .255154639$. Quadrature split at clipping boundaries confirms convergence across six widths from .1 to .0003: the rescaled crossing moves from .277316 to .255211. Rescaling all magnitudes by μ rescales the price by μ^2. The crossing quantifies how much penalty these two codes require to preserve a contrast of a given strength. Locating the optimum across all encoders would additionally require controlling directions outside these two codes.

Per-example normalization changes which gates are inexpensive.

The count proof relies on interior inputs having positive halfspace depth. Per-example normalization moves nonzero two-factor inputs onto a circle, where narrow affine caps are available. With $Y = X/\|X\|$ and m equally spaced unit decoders d_j, the responses $g_j(Y) = d_j^\top Y \mathbf{1}_{\{d_j^\top Y > \cos(\pi/m)\}}$ activate exactly one row away from boundaries and incur squared error at most $\sin^2(\pi/m)$.

If $P_0 = \mathbb{P}(Y = 0) > 1/2$ and $0 < \beta < 1/8$, the arbitrary-finite-width infimum is $\beta(1 - P_0)$. The pointwise floor $\min\{\|Y - b\|^2/2, \beta\}$ proves the lower bound, and the cap construction attains it as $m^{-2} = o(\beta)$. At $p = .25, c = .75$, its mean count is .3125 rather than the generating .5, so a pair event leaves at least one disjoint target group empty and creates positive contribution error. Any fixed finite dictionary still needs two rows for exact reconstruction of a continuous pair arc; zero halfspace depth makes the cheaper limit possible only as width grows. Restoring the radius produces conic raw-input gates with the same count and error scaled by $\|X\|^2$. Thus per-example normalization changes the response class relevant to the theorem and differs from dataset-scale normalization (Gao et al., 2025; Rajamanoharan et al., 2024b).

F.5 Quota feasibility and occurrence recovery under TopK

A TopK encoder retains the k largest nonnegative affine-ReLU scores and minimizes reconstruction error without a coefficient penalty. If the generating code fits the quota and reconstructs exactly, it attains global loss zero. No alternative can have strictly lower loss. Within a smooth gate region its second variation is $\mathbb{E}\|\dot{f}(X)\|^2 \ge 0$, since the residual vanishes. There is therefore no loss-decreasing instability of this exact solution analogous to the penalized generating reference. A preference caused by insufficient quota and an ambiguity between feasible exact reconstructions require separate comparisons.

What exact counts add. The binary event construction also fits TopK. For linearly independent unit u, v, let $X = Z_1 u + Z_2 v$, $c = \|u + v\|$ and independent $Z_i \sim \text{Bernoulli}(p)$. At width four and quota two, use decoders $u, v, (u+v)/c, (u+v)/c$ and affine-ReLU scores $(Z_1 - Z_2)_+$, $(Z_2 - Z_1)_+$, and two copies of $c(Z_1 + Z_2 - 1)_+/2$. At most two are positive. They reconstruct exactly with the same per-input activity $Z_1 + Z_2$ as the generating code padded by two silent rows. Each singleton row still misses the pair, with recall $1 - p$ and coefficient NMSE p. The objective assigns both representations zero loss.

This construction also permits varying singleton magnitudes in $(0, 1)$ while keeping pair magnitudes one, by replacing Z_i in the scores with the coefficients. When magnitudes instead vary continuously on *every* event, exact counts force each selected affine reconstruction to fit a full-dimensional set:

Proposition F.4 (Continuous magnitudes and exact activity counts)**.** *Let $n \geq 2$ generating unit rays be pairwise nonparallel, with every singleton event of positive probability. On each realized nonempty support S, suppose the generating columns are linearly independent, $|S| \leq k$, and the conditional magnitude law is absolutely continuous on $(0, \infty)^S$. A finite-width TopK SAE with affine scores, unit decoder columns, $k \geq 2$, exact reconstruction and mean activity $\mathbb{E}n_g = \mathbb{E}|S|$ recovers all generating coefficients and directions, up to permutation and inactive rows. The generating dictionary may be overcomplete.*

Proof. First recover the per-input count from its mean. Conditional on S, a reconstruction using fewer than $|S|$ rows lies in a finite union of affine spaces of dimension below $|S|$. Absolute continuity assigns that union probability zero. Exact reconstruction thus forces $n_g \geq |S|$ almost surely, and equality of their means forces $n_g = |S|$. On a singleton event, some fixed row is the sole positive output on a set of magnitudes of positive measure. The affine reconstruction identity there forces its decoder parallel to the generator and b in that span. Two nonparallel generators give $b = 0$. Positivity and unit decoder norm then identify one row j_i with decoder w_i, zero score offset, and score t on the entire ray tw_i. Every other score must be nonpositive on the observed singleton magnitudes: quota $k \geq 2$ would otherwise retain a second positive output.

Now fix a generating event S and a positive-row set J of positive conditional probability. Let $W_S = [w_i]_{i \in S}$ and let D_J contain the selected decoder columns, with scores $E_J X + a_J$. Absolute continuity makes the affine reconstruction identity hold on a set of positive Lebesgue measure in the S coordinates, hence $D_J E_J W_S = W_S$ and $D_J a_J = 0$. Since $|J| = |S| = \operatorname{rank}(W_S)$, D_J has full column rank and $a_J = 0$. Any additional row in J would now have a homogeneous score nonpositive on every generator, so could not fire on their positive combination. A generating row for an absent feature likewise has nonpositive readings on all present generators. Only rows indexed by S can fire; their independence and exact reconstruction give the generating coefficients. □

The proposition also exposes an encoder obstruction. Exact recovery requires $e_{j_i}^\top w_i = 1$ and $e_{j_i}^\top w_j = 0$ for every coactive neighbor. If every pair can co-occur, these rows form a left inverse of the generating dictionary, requiring $n \leq d$; overcomplete sources cannot satisfy exact reconstruction and exact counts through this architecture. On a restricted coactivation graph the necessary condition becomes $w_i \notin \operatorname{span}\{w_j : j \text{ coactivates with } i\}$.

The conclusion relies on exact reconstruction, continuous variation on every event, and per-input counts. Mean quotas alone allow crowded events, and excess activity permits other exact inverse codes. Appendix G tests which feasible allocations finite training reaches.

F.6 Trained recovery across further source laws

Section 3.2 compares three exact codes: rotation becomes cheaper past $\kappa = 2 - \sqrt{2}$, while transferring $\min(a_u, a_v)$ to the summed direction is cheaper for every $\kappa > 0$. Theorem 2, rather than these two-code comparisons, supplies the class-wide guarantee at the independent reference population. We now test how trained models behave across further laws.

Population and protocol. Use (6) with $p = 1/4$, exponential mean 1.5, and nearby-quantile half-width $\varepsilon = .02$. The sampler gives $s_B = \frac{1}{2} + \eta(\frac{1}{2} - .0540532)$. The sweep fixes $\eta = 1$, hence $s_B = .945947$, and sets $c = \kappa / s_B$ on the grid $\{0, .10, .20, .30, .375, .45, .52, 2 - \sqrt{2}, .65, .692737, .78, .85\}$. These maximal-dependence laws differ from both reference populations.

Models have eight unit decoders, one shared bias, $\lambda = .1$, and either $g = (WX + a)_+$ or a $2 \to 32 \to 8$ ReLU encoder. Normalized Gaussian initializations and zero biases are reset across conditions. Sixteen paired starts use 8,000 fresh-batch Adam updates at batch 4,096 and rate 10^{-3} in float64; decoder gradients are projected and columns renormalized. Seeds are 77123/44131 for architectures, 91173 for data, and 51935 for evaluation. Curves report medians and interquartile ranges.

Unlike the primary protocol, this sweep fits and evaluates Ω on the same 65,536 inputs, producing a favorable in-sample error. The largest normalized fit gap is 9.47×10^{-10}. Figure 5 and Table 3 use held-out evaluation, so their values are not directly comparable.

Reimplementation check. This separate implementation reproduces the reference anchors: one-layer NMSE is .002129 (IQR .002122–.002140) independently and .444 (.410–.485) under coupling, within .020 of the original cohorts. The sweep realizes $\kappa = .692737$ through $(c, \eta) = (.732321, 1)$ and gives .437, so it remains a separate measurement. A 500,000-draw sampler audit returns $s_B = .5004 \pm .0016$ and $.9246 \pm .0029$ against .5 and .92365. Recomputed analytic ratios reproduce $M_{\text{share}} = .667624$ and $M_{\text{rot}} = .693281$.

The measured band. The one-layer median stays near the true-strength soft-thresholding reference, 2.1×10^{-3}, through $\kappa = .45$, then reaches .00635 at .52, .129 at $2 - \sqrt{2}$ (IQR .00252–.209), .304 at .65, and .437 at .692737. The two-layer median has already risen to .0109 at $\kappa = .10$, then .0880 at .52, .227 at the crossing, and .354 at .692737. Thus the one-layer class tolerates a range of co-occurrence that the two-layer class does not.

One-layer error begins rising below $2 - \sqrt{2}$ and varies across starts near it. The vertical line marks only an exact-code crossing, not a training transition; finite-price shrinkage moves the crossing, and Appendix G.8 constructs mixtures cheaper than either endpoint.

At $\kappa = .692737$, two layers score .354 against one layer's .437, reproducing Table 3. The budget (8) constrains near-optima; it does not order the finite errors reached by two training procedures. We leave this ordering unexplained.

Equal κ does not determine the trained representation. Distinct (c, η) pairs can share $\kappa = c\, s_B$ and the three reference-code costs. We repeat $\kappa \in \{.375, .52, .65\}$ at lower η and higher c. Four of six interquartile ranges overlap, while the two-layer .65 condition separates at medians .3158 and .2376. The median paired difference is $-.0875$, with a 20,000-resample percentile interval $[-.1208, -.0167]$. Matching κ preserves the reference calculation but leaves finite-price inference and the response-class infimum free to vary, so it does not determine the trained representation.

Achieved activity does not make gates comparable. To place a magnitude price, a learned threshold, and a fixed quota on one empirical scale, we report the activity each gate achieves. Figure 3 therefore sweeps $\lambda \in \{0, .003, .01, .03, .1, .3\}$ for ReLU and $\beta \in \{5\times10^{-7}, 5\times10^{-5}, 5\times10^{-4}, 5\times10^{-3}, 5\times10^{-2}, .2\}$ for JumpReLU; Top4 has no price. All use one-layer encoders and trained decoders. Median activity spans .46–3.76, .23–3.31, and 3.22–3.84, respectively. At $\kappa = .375$ and activity near 3.3, their errors are .0745, .0736, and .1198; by $\kappa = .85$ they converge near .114. Matching activity therefore does not determine recovery across gates and source laws.

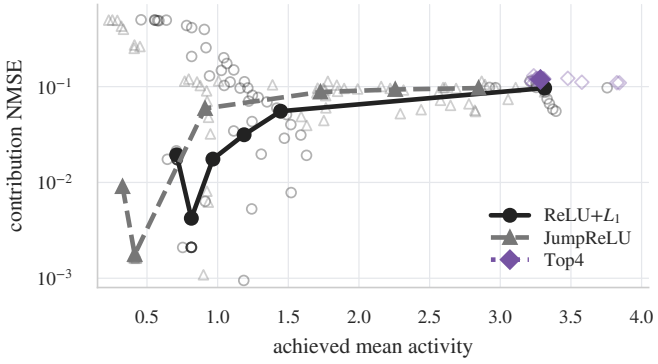


Figure 3: **Achieved activity does not collapse the gates onto one recovery curve.** Hollow marks are condition medians over eight starts; heavier curves aggregate each penalty level across the κ sweep. The activity ranges overlap.

The record contains 1,760 start-level entries: 32 anchors, 384 for the band, 96 for the invariance test and 1,248 for the gate comparison.

G TRAINING PROTOCOLS, CONFIRMATION, AND CONTROLS

The primary calibrated study, the two-factor illustration, and the architecture study use different cohorts. We specify the shared implementation first and then identify the departures before each

result. All scheduled starts enter their reported aggregates; evaluation data are never used for model selection.

G.1 Shared implementation and cohort record

Reconstruction errors sum coordinates and average examples. Decoders have unit columns, free encoder and decoder biases, and untied encoder weights. Gradients of decoder columns are projected onto their sphere tangent before Adam, then columns are renormalized after the update. Adam uses moments $(.9, .999)$, offset 10^{-8}, and no weight decay. Random decoder columns are normalized Gaussian vectors; the encoder is initialized from their transpose and both biases start at zero.

The common fixed-pool recipe uses scores $E(X - b) + a$, float32, 32,768 training inputs, batches of 256 sampled with replacement, and 6,000 updates. Learning rates are .003 through step 4,000, .001 thereafter, and .0003 after step 10,000 for continuations to 18,000. Jump thresholds start at .1 and use a rectangular straight-through derivative of bandwidth .05; hard gates and actual counts determine the forward objective and evaluation. Gradients reach b through both encoding and reconstruction.

The separate fresh-batch recipe uses $EX + a$, float64, fresh batches of 256, and 8,000 updates at .005, reduced to .001 at 5,200 and .0002 at 7,200. Jump thresholds start at $\sqrt{2\beta}$ with bandwidth $h = \min(.025, \sqrt{2\beta}/4)$. Threshold derivatives for output magnitude and count are respectively $-\theta h^{-1}\mathbf{1}_{|q-\theta|<h/2}$ and $-h^{-1}\mathbf{1}_{|q-\theta|<h/2}$. Selected scores receive ordinary reconstruction gradients. Structured initializations use their stated frame and inactive surplus outputs. Centering preserves the affine response class but changes parameter gradients.

Study	Source and procedure	Replication	Calibration / test
Figure 5	two-factor L_1, fixed pool	32 starts per law	8,000 / 24,000
Figure 4	calibrated complete/overcomplete, three gates	8 dictionaries, 3 paired starts	4,096 / 8,192
Table 4	complete L_1 coding	4 dictionaries, 1 start	8,192 common inputs per law; replicated
Table 5	complete L_1 continuation	8 dictionaries, 3 paired starts	common destination-law inputs
Table 3	two-factor one/two-layer L_1, fresh batches	16 starts per law	20,000 / 50,000

G.2 Calibration, confirmation and reference geometry

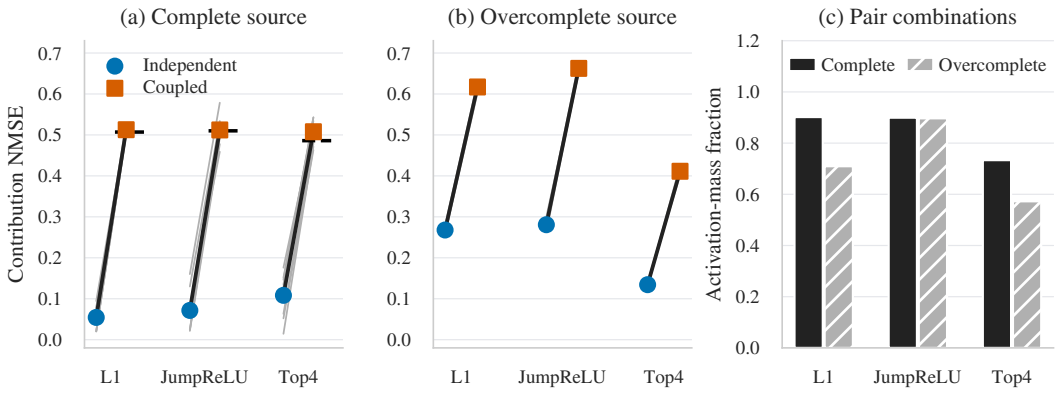


Figure 4: **The matched change persists with interacting and overcomplete factors.** Thin segments pair laws within a dictionary; thick segments give cohort means, and dashed marks are rotated-code references. Coupling increases contribution error and pair-combination activation, but only complete-source L_1 begins from declared reliable recovery. Other cells measure deterioration from their own baseline.

Calibration uses independent strengths on dictionary seeds 100 and 101, with initialization seeds 1000 and 1001. At width $2F$, screen L1 prices .05, .1, .2, .4, .8, and 1.6, and JumpReLU prices .005, .02, .08, .16, .32, and .64. Train for 6,000 updates, then continue the best two settings per gate to 18,000. Top4 crosses widths $F, 2F, 4F$ with auxiliary weights zero and $.03125$. Independent validation contribution error selects the setting, with objective as a tiebreaker; coupled results never enter selection. The frozen settings are:

Source	Gate	Width	Price/quota	Auxiliary weight
32 rays / 16D	L1	64	.2	0
	JumpReLU	64	.16	0
	Top4	64	4	0
128 rays / 32D	L1	256	.4	0
	JumpReLU	256	.16	0
	Top4	512	4	.03125

Confirmation uses dictionaries 200–207 and starts 2000–2002 under both laws, producing 288 models. The prespecified reliable-recovery control requires 18/24 starts and 6/8 dictionary medians below .10 contribution NMSE; only complete-source L1 passes (23/24, 8/8). Continuous paired effects are reported for every cell.

The primary study uses the common implementation above. Its calibrated JumpReLU schedule changes rates from .005 to .001 at step 5,200 and .0002 at 7,200, with log thresholds clamped to $[-12, 4]$. The auxiliary Top4 condition fits a detached residual with latents inactive for 200 updates; evaluation uses ordinary Top4. Fractional assignments use 4,096 calibration and 8,192 test inputs with normalized gap target 10^{-6}. Bootstrap resampling averages the three starts within each dictionary, and quadrupling calibration data changes wide-Top4 errors by at most $.0025$. Reference codes use the same support law, directions and prices; overcomplete references claim only code availability.

Decoder geometry uses each latent's maximum cosine to a signed true ray or within-pair sum/difference, weighted by mean activation. This weighting is invariant to proportional splitting. Complete coupled activation shares within 15° of combinations are $.902, .899, .733$ for L1, JumpReLU and Top4; overcomplete shares are $.710, .897, .573$. Independent complete L1 instead has true-ray alignment .991 and rotated alignment .727, close to $\cos 45^\circ$, so its small coding advantage does not indicate pair rotation.

Count and activation at the trained prices. For complete coupled L1, faithful and rotated references have activities 2.617 and 2.335, and expected activations 2.613 and 2.327. Trained activity is 4.817 at expected activation 2.333, consistent with splitting around the rotated code. JumpReLU trained/reference rotated activities are $1.553/1.526$ and expected activations $2.529/2.502$. The unthresholded exact-code identity, activity 3.2 versus four, cannot describe these finite-price responses.

G.3 Comparing dictionaries after optimizing their codes

Table 4: **The coupled encoder already beats optimal coding on the generating directions.** Mean L_1 objectives at $\lambda = .2$, on four preselected complete-source dictionaries with one paired start each and 8,192 common evaluation inputs per law. Recoding optimizes nonnegative activations and one shared bias, and its generating-dictionary bracket closes to 4.42×10^{-8}. Read down the coupled row: the trained encoder is already below the best code on the generating directions, so no improvement to inference at those directions can account for its allocation. Read down the independent row: the trained encoder is above that reference, which is the ordering an inference shortfall would produce.

Strengths	Trained encoder	Recoded learned D	Recoded true D
Independent	.593	.567	.575
Coupled	.535	.518	.576

For dictionaries 200–203 and start 2000, we minimize the empirical L1 objective over nonnegative codes and one shared bias on 8,192 common inputs, using price .2 for the complete source and $.4$ for the overcomplete source. Appendix D.2 supplies the primal and dual; centering and globally scaling

each residual witness enforce $\sum_k y_k = 0$ and $D^\top y_k \le \lambda \mathbf{1}$. An independent 8,192-example block replicates the comparison. Maximum true-dictionary gap is 4.42×10^{-8}; learned-dictionary gaps are 1.06×10^{-4} complete and 2.75×10^{-6} overcomplete. Duplicating generating rays leaves their optimum unchanged, so the reference is the best code on those directions at any width.

The mean true-minus-learned advantages are $.00718/.05803$ for independent/coupled complete sources and $.01647/.08601$ overcomplete; every certified lower advantage is positive on both blocks. Replication gives $.00716/.05810$ and $.01709/.09064$. The overcomplete trained/learned-recoded/generating-recoded triplets are $1.3468/1.0316/1.0481$ independently and $1.2917/.9648/1.0508$ under coupling. Every complete coupled encoder beats its true-dictionary lower bound by .0293–.0529, while every independent encoder remains above its true-dictionary optimum.

These empirical brackets have a different scope from Theorem 2. The confirmation dictionaries are nonorthogonal across pair blocks, the price is finite, and the recoder optimizes each input separately rather than using an affine-ReLU response. Each difference permits a small independent-law recoding advantage outside the theorem's certificate. We therefore keep attained encoder objectives separate from recoded objectives.

G.4 Continuation, fixed directions and training access

Continue every complete-source confirmation checkpoint for 12,000 updates either on the original or the other strength law. Retained-Adam and reset-Adam arms begin at identical weights and preserve paired sampling and schedule. Each gate has 192 arms: 24 starts per law, two starting laws, two continuation laws, and two optimizer-state choices. Across three gates, all 576 endpoints are evaluated on both fixed test laws. Compare switched minus same-law endpoints on the destination law, averaging starts within dictionaries before bootstrapping. This avoids confounding changes in evaluation difficulty with changes in the model.

Table 5: **Changing the training law redirects the same model.** Switched minus same-law L_1 continuation on the complete source, evaluated on matched destination-law inputs. Parentheses give optimizer-reset results. Every start has the displayed sign of objective change.

Change	Δ contribution NMSE	Δ objective
Add coupling	$+.1358$ $(+.1346)$	$-.02147$ $(-.02139)$
Remove coupling	$-.4218$ $(-.4219)$	$-.07266$ $(-.07264)$

L1 independent-to-coupled effects are $+.1358\,[.0995, .1763]$ with retained Adam and $+.1346$ after reset; all 24 starts and eight dictionaries deteriorate. True-ray alignment falls $.9934 \to .9158$ as combination alignment rises $.7217 \to .8350$. Reversing the switch gives $-.4218\,[-.4389, -.4026]$ and $-.4219$ after reset, with 18/24 starts ending below .10. Initially low-error JumpReLU and Top4 starts change by $+.0064\,[-.0002, .0173]$ and $-.00043\,[-.00084, -.00001]$, respectively. Separate L1 transfers give $+.0688$ at copula correlation .9 and $+.2064$ overcomplete, with the same conclusions after resetting Adam.

On identical destination-law samples, every L1 start also moves to a lower objective: the mean changes are $-.021468$ when adding coupling and $-.072658$ when removing it, matching Table 5 and persisting after optimizer reset. The recovery movement therefore follows the destination objective.

Freezing the generating directions rules out pair rotation and tests inference in that geometry. Complete L1 errors remain near $.032/.035$, and JumpReLU near $.052/.050$, under independent/coupled laws. On four overcomplete dictionaries, random-start L1 changes $.2633 \to .3030$, compared with $.2688 \to .6191$ under joint training; Top4 changes $.1254 \to .1240$, compared with $.1350 \to .4173$. Fixing directions therefore removes most of the coupling deficit, while the remaining inference error leaves expressivity and optimization unresolved.

Complete-source Top4 retains a random-start failure even with generating directions fixed: at width 32, random-start errors are $.4001/.3776$, whereas aligned starts give $.0119/.0129$. Random encoders miss $27.5\%/24.6\%$ of true energy through nonpositive scores, compared with $1.7\%/2.6\%$ displaced by the quota, so increasing the quota cannot recover the missing responses. Exact-inverse starts

perturbed at scale 10^{-3} finish below 7.2×10^{-7}. Same-sample finite differences agree with the implemented gradients in smooth regions, leaving the optimization route unresolved.

G.5 Finite count prices and access to faithful solutions

On the complete source, faithful and rotated hard-threshold codes cross near $\beta = .0079$ for $\rho = .99$ and .1095 for $\rho = .9$; no optimization over other encoders enters these values. At $\beta = .16$, rotation becomes cheaper by dropping weak contrasts. Random-start independent/coupled JumpReLU errors are $.615/.664$ at .002, $.365/.568$ at .01, and $.071/.511$ at .16, while faithful inverse starts remain near .00005, .0005, and .020. At the smaller prices every random model stays above an available faithful objective, establishing an optimization shortfall rather than a recovery boundary. The count theorem assumes near-optimality and gives no convergence guarantee.

A two-factor factorial attributes most of this training difference to schedule and data reuse: at 8,000 updates, the original schedule gives independent errors .00030–.00049, while the common schedule gives .0237 with a fixed pool and .0019–.0021 with fresh batches. Precision has a much smaller effect. These controls set the calibration schedule without selecting on coupled recovery.

G.6 Robustness of the training and recovery comparison

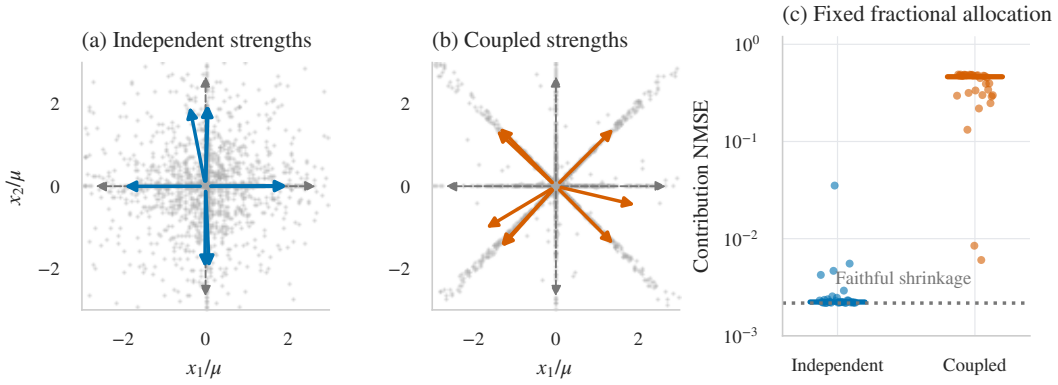


Figure 5: **Changing paired strengths changes learned contributions.** The laws share directions, occurrences, marginal magnitudes and covariance, and differ only in how paired magnitudes move together. Gray arrows mark generating directions and colored arrows trained decoders: coupling moves the decoders toward sums and differences and raises median contribution NMSE from .00223 to .464, against .00213 for faithful soft-threshold shrinkage. Width eight can hold both four-direction codes, so a shortage of outputs does not force the rotation.

Fractional evaluation of the fresh-batch cohort. For the same 32 starts per law, fractional assignments fitted on 8,000 inputs and tested on 24,000 give medians .002182 independently and .490007 under coupling, compared with $.002182/.745343$ for saved whole groups. The relaxation reduces the coupled deficit without approaching the true-strength soft-thresholding reference ($.002177/.002178$). All 68 convex fits have normalized gaps below 9.03×10^{-9} and agree with direct vector evaluation within 10^{-9}. The plot reports finite-model squared error; the theorem proves mean-absolute convergence.

Changing the training recipe.

On the same two-factor population, 32 paired L1 starts give independent/coupled errors $.002183/.489540$ without centering and $.002184/.489431$ with centering.

For JumpReLU at $\beta = .01$, 16 paired starts cross centering with two threshold settings: initial threshold $\sqrt{2\beta}$ and bandwidth .025, or fixed threshold .1 and bandwidth .05. Reconstruction FVU stays between .000346 and .000386, while contribution errors vary:

Encoder input	Threshold setting	Independent NMSE	Coupled NMSE
Raw	Price-based	.000455	.0679
Raw	Fixed	.000439	.0436
Centered	Price-based	.000467	.1037
Centered	Fixed	.000440	.0950

Thus optimization parameterization changes the finite-price outcome within the same encoder class. Allocation uses 8,000 calibration and 24,000 test inputs; every normalized gap is below 10^{-8}.

The common recipe then changes initialization, data reuse, precision and schedule together. All 32 L1 starts enter Figure 5; after 6,000 updates, 31/32 independent starts and 2/32 coupled starts have error below .02. Continuing the same checkpoints gives:

Gate	Updates	Independent NMSE	Coupled NMSE
L1	6,000	.00223	.4636
L1	18,000	.00220	.4897
JumpReLU	6,000	.1244	.1171
JumpReLU	18,000	.1373	.2652

The L1 separation survives the recipe change. After 18,000 updates, only 15/32 independent JumpReLU starts fall below .02 even though their median objective decreases from .01008 to .00612; the explicit generating encoder attains .00462. These finite-price runs therefore do not meet the count theorem's near-optimality premise.

G.7 Additional width, activity, data, and magnitude controls

To vary joint strengths without changing geometry or feature frequency, construct $r = F/2$ unit factors in independently drawn orthogonal pairs, include both signs as F nonnegative rays, and set SAE width to F. The primary sizes are $(d, F) = (16, 64), (32, 128), (32, 512)$, with two dictionaries and three starts at the first two sizes and one dictionary with three starts at the last.

Each input selects two pair blocks. A selected block activates both factors with probability .6 and either singleton with probability .2, giving mean activity 3.2, maximum four and conditional partner probability .75. Signs are independent and balanced. Correlated standard normals (H_1, H_2) produce unit-mean exponential strengths $B_i = -\log \Phi(H_i)$; the primary comparison uses $\rho = 0, .99$, with $.5, .9$ as intermediate controls. This Gaussian copula differs from the circular-quantile law in Figure 5.

For the matrix U of underlying columns,

$$\mathbb{E}X = 0, \qquad \mathrm{Cov}(X) = (6.4/r)UU^\top,$$

independently of ρ. Almost surely, every valid union of up to eight underlying columns is linearly independent, so coefficients are unique given U. This does not identify an unknown dictionary or supply an affine inverse.

Training and evaluation. These controls use the common fixed-pool recipe of Appendix G.1. Penalties are $\lambda \in \{.01, .05, .2\}$, $\beta \in \{.005, .02, .08\}$ and TopK quotas $k \in \{4, 8\}$. Prices remain fixed within each dependence pair rather than being matched across architectures.

Fractional assignment uses 4,096 calibration and 4,096 test inputs. The normalized gap target is 10^{-6}; twelve severe JumpReLU cases reach at most 3.75×10^{-5}, negligible beside errors near one. Reconstruction FVU uses centered input energy and contribution NMSE uses coefficient energy. All scheduled runs enter the medians.

At $F = 64, d = 16$, L1, JumpReLU and Top4 reconstruction FVU changes .0596 $\to$ .0525, .0480 $\to$ .0439, and .0960 $\to$ .0867, while activity changes 10.88 $\to$ 9.83, 4.27 $\to$ 3.72, and 3.97 $\to$ 3.98. Better reconstruction therefore cannot explain the larger contribution error. Every tested price has the same sign at the first two scales, although near-ceiling baseline errors obscure effect size.

Preference on identical inputs. For baseline models at $\lambda = .2$, $\beta = .08$, or $k = 4$, evaluate each independently and coupled-trained pair on both laws. Allocations use 4,096 calibration inputs and

Table 6: Matched dependence at every primary penalty. Entries give median fractional contribution NMSE, independent → coupled strengths. Six paired runs span two dictionaries at each of the first two sizes; the last uses three starts of one dictionary. Prices are not matched for activity across architectures; dashes mark untested settings.

Gate	Price / quota	$F = 64, d = 16$	$F = 128, d = 32$	$F = 512, d = 32$
L1	0.01	$0.778 \to 0.820$	$0.836 \to 0.874$	—
L1	0.05	$0.716 \to 0.780$	$0.758 \to 0.828$	$0.979 \to 0.983$
L1	0.2	$0.631 \to 0.763$	$0.479 \to 0.751$	$0.962 \to 0.976$
JumpReLU	0.005	$0.874 \to 0.884$	$0.924 \to 0.929$	—
JumpReLU	0.02	$0.845 \to 0.859$	$0.895 \to 0.907$	$0.991 \to 0.992$
JumpReLU	0.08	$0.793 \to 0.847$	$0.492 \to 0.770$	$0.990 \to 0.993$
TopK	4	$0.392 \to 0.687$	$0.209 \to 0.597$	$0.448 \to 0.737$
TopK	8	$0.821 \to 0.845$	$0.407 \to 0.706$	$0.671 \to 0.844$

two 8,192-input test blocks. On coupled inputs, errors change .642 → .769 (L1), .799 → .849 (JumpReLU), and .397 → .688 (Top4), while objectives fall $4.77\%, 5.90\%, 16.43\%$. All six pairs per gate have lower objective and larger error. On independent inputs, objectives instead rise $5.16\%, 4.90\%, 24.12\%$. Both blocks preserve every sign; fixing rather than refitting assignments changes errors by at most .00408. This establishes an attained preference reversal without identifying the global optimum.

Available width and permitted activity. Hold the 64-ray, 16-dimensional population fixed and compare 64, 128 and 256 SAE latents under the common schedule, with two dictionaries and three paired starts. Cross width with Top4 and Top8, and retain L1 at .2 and JumpReLU at .08. Four responses suffice for every generating support; Top8 separately tests a looser activity constraint. Fractional assignments share calibration and test draws, with normalized gaps below 10^{-6}.

Gate	64 latents	128 latents	256 latents
L1, .2	.631 → .763	.610 → .744	.613 → .748
JumpReLU, .08	.793 → .847	.771 → .846	.776 → .843
Top4	.392 → .687	.269 → .517	.248 → .453
Top8	.821 → .845	.773 → .796	.717 → .746

Entries are independent → coupled median contribution NMSE. More rows improve Top4, whereas a larger quota lowers reconstruction error and worsens assignment. At width 256, continuing to 18,000 updates leaves Top4 near $.242/.454$; Top8 FVU improves $.0258/.0247 \to .0175/.0152$ as contribution error worsens $.717/.746 \to .804/.837$. An inactive-latent loss gives Top8 errors $.809/.837$, and Top4 uses about 100 rows despite only 64 generating rays. Available width, used width and faithful assignment are therefore distinct.

Which alternatives survive the controls? The following paired Top4 controls use fresh dictionaries. The auxiliary term, reported separately because it changes the objective, has weight .03125 and fits a detached residual with inactive latents. Lognormal strengths match the exponential marginals' first two moments; the geometric control sets within-pair cosine to .3.

Top4 control	Independent NMSE	Coupled NMSE
18,000 training updates	.392	.684
Eightfold training pool	.418	.696
Auxiliary loss, baseline size	.383	.688
Lognormal strengths	.427	.665
Within-pair cosine .3	.417	.647
Auxiliary loss, $F = 512, d = 32$	.415	.711

The continuation uses six paired runs; other controls use three starts of one dictionary. They test persistence under longer training, more data, an auxiliary loss, another strength law and nonorthogonal geometry. Independent errors are retained to measure the additional coupling effect.

At copula correlation .99, the rotated/generating L_1 activation ratio is .911, versus 1.149 independently. Under a count penalty, rotation is 11.28% more expensive at $\beta = .001$ and 9.37% cheaper at .05. For fixed $\rho < 1$, continuous pairs differ almost surely, so vanishing-price rotation uses two outputs per selected block versus 1.6 for generation; exact ties use 1.4. The tie and small-price limits therefore do not commute.

G.8 Reference-price crossings and low-price training

A fixed-law price sweep. Use the two-factor exponential population with $p = .25, \mu = 1.5, c = .63, \eta = .95, \varepsilon = .02$ and vary only λ. Enumeration and quadrature place the generating–rotated crossing at .0269; doubling quadrature order changes the difference there by 3.1×10^{-8}. At six prices, train 16 paired random starts and both structured controls for 24,000 updates, repeat with independent magnitudes, and continue coupled runs at .01, .03, .1 to 56,000 updates.

The independent controls retain low group error at all six prices. The coupled outcomes are:

L1 penalty λ	.01	.03	.06	.1	.2	.3
NMSE $<$ .02 after 24,000 updates (of 16)	9	9	7	7	7	2
After 56,000 updates, where tested	9	9	—	7	—	—

At $\lambda = .3$, generating shrinkage error is already .0176, close to the cutoff. The .0269 crossing does not mark a training transition: seven high-error runs at .01 remain at .067–.862 after continuation, with objectives on both sides of the generating reference.

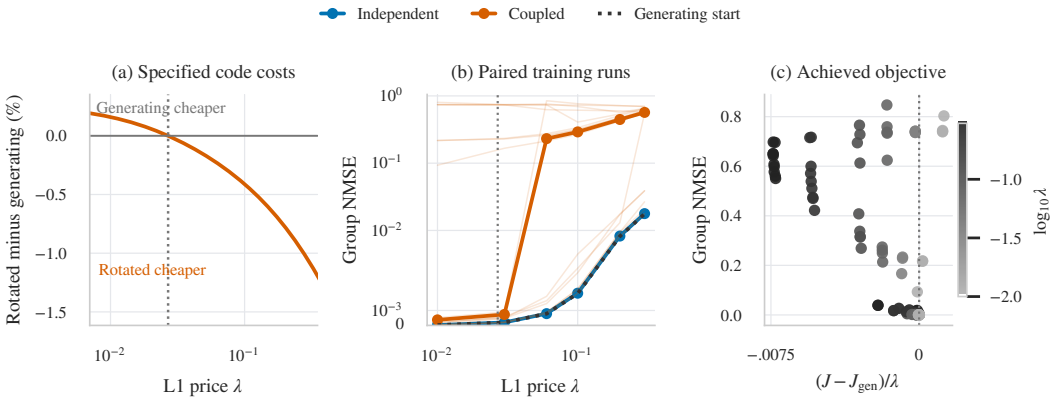


Figure 6: **The cost crossing does not determine which training runs recover.** (a) Conditional quadrature compares the generating and rotated soft-thresholded codes on the same coupled population. (b) Thin lines follow 16 paired random starts per law; thick lines show median group error after 24,000 updates, and the dotted curve follows the generating start. (c) Each coupled run is paired with its population-integrated objective difference from generating reconstruction. Large recovery errors occur at objectives both above and below this reference. The generating code is feasible; its cost does not identify the class optimum. All prices share data laws and paired random streams.

Why comparing two codes misses a finite-price allocation. An eight-row encoder can retain both sets of directions. Let $S_\theta(x)$ denote coordinatewise signed soft thresholding at θ, and let Q_{rot} be the orthogonal 45° change of frame. Consider

$$\widehat{X} = (1-t)S_{\theta_a}(X) + tQ_{\text{rot}}S_{\theta_r}(Q_{\text{rot}}^\top X), \qquad 0 \le t \le 1.$$

The signed notation abbreviates eight nonnegative affine-ReLU responses on four generating and four diagonal rays. Their slopes are scaled by $1-t$ and t, with separate thresholds θ_a, θ_r. No negative feature amplitudes are introduced.

The construction allocates reconstruction between the two direction sets before choosing their shrinkage. More generally, combining feasible codes with weights $1-t, t$ gives

$$J_{\text{mix}} = (1-t)J_a + tJ_b - \frac{t(1-t)}{2}\mathbb{E}\|\widehat{X}_a - \widehat{X}_b\|^2.$$

The penalty is linear, while averaging distinct reconstructions lowers squared error. Separate thresholds enlarge the choice, so endpoint costs can miss a better mixture.

Changing coactivity at a fixed price. A separate sweep varies the pair probability at $\lambda = .1$, holding $p = .25$, balanced signs, exponential mean 1.5, and all other population choices fixed. For $\eta = .95$, the generating–rotated crossings occur at $c = .6235$ and $.7224$ for quantile half-widths .02 and .10. Width-eight models use 16 paired and structured starts, 24,000 updates, and held-out whole-group assignments. Both structured starts can persist beyond a crossing; fractional evaluation removes changes caused only by proportional splitting. Appendix I.2 uses this cohort to compare fixed allocation with reduction from the available decoded vectors.

H MATCHED-WIDTH AFFINE-RELU SAE

This auxiliary population isolates a question that the larger training study cannot solve analytically: can the generating representation remain a stable trained solution even when the same-width objective admits a better event-based representation? We give explicit parameters for both solutions and calculate their losses and local stability. Write $A = 1$ during the calculation; multiplying activations, biases, and codes by A scales every reported loss and curvature by A^2 when λ/A is fixed.

H.1 GENERATING SOLUTION AND LOCAL STABILITY

Let $X = Z_1 u + Z_2 v$, where u, v are orthonormal and Z_1, Z_2 are independent $\mathrm{Bernoulli}(p)$ variables. In the active regime $t := \lambda/A < 1 - p$, the generating dictionary is

$$D_T = [u, v], \qquad b_T = \frac{pt}{1-p}(u + v),$$

with exact codes

$$z_i^* = \left(1 - \frac{t}{1-p}\right) Z_i.$$

Each code is an affine ReLU of $x - b_T$.

To test local stability of the exact-code, bias-profiled decoder objective, rotate the two decoder columns on the sphere and reoptimize the codes and decoder bias. Put $q = 1 - p$, $\Delta = q - t$, and

$$h = p\Delta - \frac{p^2 t^2}{q}, \qquad k = -\frac{2p^2 t\Delta}{q}, \qquad \gamma = p\Delta.$$

In orthonormal symmetric and antisymmetric coordinates, the Hessian separates into

$$H_{\mathrm{rot}} = \begin{pmatrix} h+k & -\gamma \\ -\gamma & q \end{pmatrix}, \qquad H_{\mathrm{merge}} = \begin{pmatrix} h-k & \gamma \\ \gamma & q \end{pmatrix}.$$

The first determinant is

$$\det H_{\mathrm{rot}} = pq(q^2 - t).$$

The merge block becomes singular only at

$$t_{\mathrm{merge}} = \frac{q\left(4p + \sqrt{1+8p} - 1\right)}{8p}, \qquad q^2 < t_{\mathrm{merge}} < q,$$

so it remains positive when the rotation block crosses zero. Thus the generating solution is locally stable for

$$t < (1-p)^2$$

and becomes a saddle above this value. At $p = 1/4$ and $t = 1/2$, the two blocks are

$$H_{\mathrm{rot}} = \begin{pmatrix} 1/48 & -1/16 \\ -1/16 & 3/4 \end{pmatrix}, \qquad H_{\mathrm{merge}} = \begin{pmatrix} 1/16 & 1/16 \\ 1/16 & 3/4 \end{pmatrix}.$$

Profiling the bias takes the Schur complement of the lower-right entry, giving decoder curvatures $1/64$ and $11/192$. In particular, the smallest bias-profiled decoder curvature is

$$p\big((1-p)^2 - t\big) = \frac{1}{64} > 0.$$

Strict active and inactive KKT margins keep the active-set calculation valid in a neighborhood. This curvature statement concerns decoder directions and bias after profiling the codes; it does not assert positive curvature in every redundant encoder parameterization. The perturbation check at the end tests return in the redundant encoder parameterization separately.

H.2 Explicit lower event representation

At $p = 1/4$ and $t = 1/2$, the four event probabilities are

$$P(00) = \frac{9}{16}, \quad P(10) = P(01) = \frac{3}{16}, \quad P(11) = \frac{1}{16}.$$

Define

$$a_{11} = \frac{\sqrt{514} - 10}{20}, \qquad a_{01} = \frac{\sqrt{314} - 10}{20},$$
$$b_E = \frac{1}{4}u + \frac{3}{20}v, \qquad d_{11} = \frac{15u + 17v}{\sqrt{514}}, \qquad d_{01} = \frac{-5u + 17v}{\sqrt{314}}.$$

Both decoder columns have unit norm. The affine-ReLU gates

$$g_{11}(x) = 2a_{11}\left(\langle u + v, x\rangle - \frac{3}{2}\right)_+,$$
$$g_{01}(x) = 2a_{01}\left(\langle v - u, x\rangle - \frac{1}{2}\right)_+$$

produce the eventwise codes

x	0	u	v	$u+v$
g_{11}	0	0	0	a_{11}
g_{01}	0	0	a_{01}	0.

On the two active events,

$$d_{11}^\top(u + v - b_E - a_{11}d_{11}) = \frac{1}{2},$$
$$d_{01}^\top(v - b_E - a_{01}d_{01}) = \frac{1}{2},$$

so the active KKT equalities hold. Every inactive correlation is strictly below $1/2$; the smallest slack is

$$\frac{1}{2} - \frac{87}{10\sqrt{514}} = 0.11625945\ldots > 0.$$

The smallest positive code and smallest inactive gate margin are both

$$a_{01} = 0.38600226\ldots > 0.$$

Thus the codes are exact nonnegative-Lasso solutions and are realized exactly by the displayed gates.

Direct evaluation on the four events gives

$$L_E = \frac{19}{400} + \frac{3\sqrt{314} + \sqrt{514}}{640} = 0.1659870368\ldots,$$

whereas $L_T = 1/6$. The strict gap is

$$L_T - L_E = 0.0006796299\ldots > 0.$$

The displayed event bias is feasible and has exact gates, although its mean residual is slightly nonzero. Joint optimization of all SAE parameters reaches the nearby stationary event solution of loss 0.1659670862.

H.3 Initialization and robustness checks

All optimization experiments use exact population expectations over the four events. In 3,000-step Adam runs, all 12 small perturbations of the generating solution returned within 0.035° of it. Four of 32 random starts reached the lower stationary event basin, four reached the generating basin, and the remainder reached higher dead-row, one-row, or asymmetric basins. Symmetry supplies a mirror event basin.

The strict objective gap and positive profiled curvature resolve the local-versus-global question on the same width-two population. The experiment checks this explicit construction; it does not estimate the prevalence of its basin in other sources.

I READING, CONTRIBUTION ALLOCATION, AND SELECTIVE REDUCTION

Section 5 compares two operations on identical inputs. We give its calibration and certificate, then show that the binary event code used for intuition is selected by a population objective. The geometric bound explains why a small activation saving can accompany a much larger contribution error.

I.1 READING AND REMOVAL ON IDENTICAL CONFIRMATION INPUTS

Table 7: **Reading succeeds more accurately than selective reduction on the same coupled inputs.** Both columns use the generating coefficient-energy denominator. The complete-source reader applies the known inverse to the reconstruction; the overcomplete reader is an affine map fitted on separate calibration data. Reduction values are certified lower bounds over the reduction set $\mathcal{C}(X)$ of Section 5. An affine reader on raw overcomplete inputs scores .74. Appendix I.1 includes independent-law results, where the ordering reverses for overcomplete JumpReLU and Top4.

	Complete		Overcomplete	
Gate	Reader	Reduction	Reader	Reduction
L_1	.0343	.352	.197	.414
JumpReLU	.0929	.351	.314	.475
Top4	.0126	.314	.130	.284

Use the eight overcomplete confirmation dictionaries, three gates, and start 2000 fixed in advance under both laws (48 models). Fit an affine reader of signed-ray strengths on 16,384 examples and select ridge regularization on an independent 8,192-example block. The reported reader uses the same 512 test inputs as the removal certificate; the full 8,192-example test is a robustness check.

For each target $\alpha_i w_i$ and input, minimize $\|\alpha_i w_i - Dz\|^2$ over $0 \le z \le g(X)$. Feasible codes and dual lower bounds certify the distance, with maximum normalized gap below 6.7×10^{-16}. Targets may independently reuse the available activations, so the test relaxes both fixed assignment and the shared per-latent budget. Reader squared coefficient errors and removal squared vector errors use the same coefficient-energy denominator because the generating directions are unit length.

	Independent		Coupled	
Gate	Reader	Removal	Reader	Removal
L1	.1708	.2006	.1969	.4141
JumpReLU	.2509	.1904	.3141	.4752
Top4	.0946	.0702	.1296	.2841

Entries are dictionary means on matched inputs; every dictionary and the full reader test preserve the within-law ordering. Under coupling, readers outperform reduction for every gate; under independence, JumpReLU and Top4 reverse this order. Reading establishes accessible information, whereas the certificate excludes reductions in its specified class. Raw-input affine readers score near .74, so the nonlinear code makes the targets more accessible to affine reading without creating information.

A constructive reader on the central source. In our orthogonal signed source, reconstruct the input from all SAE activations, project onto u_i, and take the positive or negative part: $\widehat{\alpha}_{i,\pm} = (\pm u_i^\top f(X))_+$. The signed positive-part map is one-Lipschitz, giving

$$\sum_{i,\pm} \mathbb{E}(\widehat{\alpha}_{i,\pm} - \alpha_{i,\pm})^2 \le \mathbb{E}\|f(X) - X\|^2. \tag{I.1}$$

This readout estimates all generating strengths accurately as reconstruction error vanishes in either population. It uses the known generating axes and can combine the same latent into several estimates. The contribution test instead asks whether the existing decoded vectors supply the separate features one would want to remove, even after splitting latents into fixed fractions.

Why contribution error is not an information measure. On the fixed overcomplete Top4 models, increasing the inference quota to six retains the entire Top4 code: applying Top4 again recovers it

exactly. It also enlarges the set of possible reductions. Yet contribution error changes $.134 \to .171$ independently and $.415 \to .436$ under coupling, while removal error decreases approximately $.070 \to .059$ and $.281 \to .262$. A fixed assignment of decoded contributions and a reader using the whole code answer different questions. The main contribution metric is therefore interpreted as allocation error throughout.

For the complete source, the known factor inverse applied to the SAE reconstruction supplies the reader in Table 7. It measures reconstruction in coefficient coordinates; the reduction column tests a different operation on the same code.

I.2 Which contributions can existing activations supply?

A small activation deficit can hide a much larger geometric error. Even with an input-dependent assignment, every contribution obtainable by reducing existing activations lies in

$$\mathcal{C}(X) = \left\{ \sum_j z_j d_j : 0 \le z_j \le g_j(X) \right\}.$$

Allowing each target to use this entire set independently relaxes the column budgets of fractional allocation. Consequently

$$\mathcal{E}_{\mathrm{frac}} \ge \mathbb{E} \sum_i \operatorname{dist}(\alpha_i w_i, \mathcal{C}(X))^2. \tag{I.2}$$

For any nonzero vector v and target t, the support function of this set gives

$$\operatorname{dist}(t, \mathcal{C}(X)) \ge \frac{\left(v^\top t - \sum_j g_j (v^\top d_j)_+\right)_+}{\|v\|}.$$

This follows from $v^\top(t - z) \le \|v\|\|t - z\|$ and $\sup_{z \in \mathcal{C}(X)} v^\top z = \sum_j g_j (v^\top d_j)_+$, so any candidate v certifies a lower bound. A diagonal contribution $B(u + v)$ can reconstruct a pair exactly, yet reductions cannot separately supply Bu and Bv; even independent best projections leave normalized squared error $1/2$. Divide (I.2) by T for the normalized metric.

We apply these bounds to the continued width-eight L_1 coactivity sweep of Appendix G.8: sixteen paired starts per condition at $\lambda = .1$, $c \in \{.75, .90\}$, coupled $\eta = .95, \varepsilon = .02$, and independent $\eta = 0$, trained for 24,000 updates. This cohort is separate from Figure 5. Fixed assignments use 8,000 calibration and 24,000 test inputs; a second projected-gradient solver verifies normalized gaps below 10^{-8}. The table reports median normalized errors and bounds on the same test inputs:

Conditional partner probability c	.75	.90
Independent fractional error	.00213	.00216
Coupled whole-group error	.812	.689
Coupled fractional error	.489	.488
True-strength soft thresholding	.00213	.00211
Global activation bound	.0092	.0241
Inputwise activation bound	.0552	.0666
Available-vector lower certificate	.269	.326

The inputwise bound averages $(\sum_i \alpha_i - \sum_j g_j)_+^2 / |S|$, with zero on empty inputs; the geometric certificate uses the actual decoders. Box-constrained least squares and the diagonal calculation check it. Table entries estimate expectations on held-out data, while each feasible witness certifies its individual input.

I.3 An optimal event code with inaccurate individual detectors

The main text's binary code assigns separate latents to the two singleton inputs and their pair. To justify studying that representation, we show that a small-penalty version attains the global objective minimum. The key lower bound asks how much activation any unit dictionary needs to reconstruct an input relative to its bias. Equality then identifies which latents can fire.

Proposition I.1 (High fidelity with distributed occurrences)**.** *Let $X = Z_1 u + Z_2 v$, with independent $Z_i \sim \mathrm{Bernoulli}(1/4)$ and linearly independent unit u, v. For all sufficiently small positive λ, a width-three affine-ReLU SAE is globally optimal over all widths, biases, unit dictionaries and nonnegative responses. At width three its bias, dictionary and responses on the four inputs are unique up to permutation. Reconstruction MSE and singleton decoder cosine deficits are $O(\lambda^2)$, yet every individual latent of every global optimum has occurrence F1 at most $6/7$ for either generator. Overlapping readouts of the responses recover each occurrence exactly. For orthogonal u, v, all statements hold for $0 < \lambda \le 1/2$.*

Appendix D.3 proves the radial lower bound used here, following the principle of Dorrell (2026). For a centered input $\nu = x - b$, it gives the samplewise floor $\psi_\lambda(\|\nu\|)$. Above threshold, equality requires total code activation $\|\nu\| - \lambda$ and every active decoder to align with ν; below threshold, the unique optimal code is zero. The additional work for Proposition I.1 is therefore to find the population-optimal bias and to show that three affine-ReLU event gates attain these equality conditions simultaneously.

Nonorthogonal directions and uniqueness. Let u, v be any fixed linearly independent unit directions and $\delta = \min(1, \|u+v\|) > 0$. The zero atom has mass $P_0 = 9/16$. On the ball $\|b\| \le \lambda$, it contributes a quadratic of curvature P_0 to $F_\lambda(b) = \mathbb{E}\psi_\lambda(\|X - b\|)$; all other gradient terms have total norm at most $7\lambda/16$. The gradient points outward on the boundary. The unique minimizer is therefore interior and globally optimal, with $\|b_\lambda\| \le 7\lambda/9$. If $\lambda < 9\delta/16$, every nonzero input remains outside the threshold ball. Stationarity expresses b_λ as a convex combination of all four vertices with positive weights, so it is in the parallelogram's interior. Distinct vertices then have distinct positive directions from the bias.

One detector per nonzero vertex attains radial equality: use the linear dual coordinates x_1, x_2 satisfying $X = x_1 u + x_2 v$ in the strict event gates below, with positive coefficients $\|x - b_\lambda\| - \lambda$. Every global optimum must have this bias and attain equality on each input. Each used latent serves only one vertex. There are three such vertices, so at width three the directions and positive coefficients are unique up to permutation. Encoder parameters that agree on these four inputs can differ elsewhere. Reconstruction error is $O(\lambda^2)$, and each singleton decoder has cosine deficit $O(\|b_\lambda\|^2) = O(\lambda^2)$. The individual F1 ceiling and exact occurrence readouts follow from event probabilities, independently of the angle. The orthogonal calculation below gives explicit constants and covers $\lambda \le 1/2$.

For orthogonal u, v, symmetry reduces the bias to $b = s(u + v)$. In the cell where only zero lies within distance λ, the radial objective is

$$F_\lambda(s(u+v)) = \frac{9}{16}s^2 + \frac{3\lambda}{8}\sqrt{(1-s)^2 + s^2} + \frac{\sqrt{2}\lambda}{16}(1-s) - \frac{7\lambda^2}{32}. \tag{I.3}$$

Its strictly increasing derivative vanishes when $\frac{9}{8}s + \frac{3\lambda}{8}(2s-1)/\sqrt{(1-s)^2+s^2} - \frac{\sqrt{2}\lambda}{16} = 0$. With $a = (6 + \sqrt{2})/18$, it is negative at zero and positive at $a\lambda$, so $0 < s_\lambda < a\lambda$. This validates the cell through $\lambda = 1/2$: the zero distance is below λ, singleton distances exceed $1/\sqrt{2}$, and the pair distance exceeds $\sqrt{2}(1 - a/2) > 1/2$. Symmetry and convexity give global optimality; the zero atom supplies local strict convexity and hence uniqueness. The implicit-function theorem gives $s_\lambda = a\lambda + O(\lambda^2)$.

Put $\rho_\lambda = \sqrt{(1-s_\lambda)^2 + s_\lambda^2}$ and choose

$$d_{10} = \frac{(1-s_\lambda)u - s_\lambda v}{\rho_\lambda}, \qquad d_{01} = \frac{-s_\lambda u + (1-s_\lambda)v}{\rho_\lambda}, \qquad d_{11} = \frac{u+v}{\sqrt{2}}. \tag{I.4}$$

Let $c_\lambda = \rho_\lambda - \lambda$ and $a_{11} = \sqrt{2}(1-s_\lambda) - \lambda$, both positive. In raw coordinates $x_1 = u^\top X, x_2 = v^\top X$, the gates $c_\lambda(2x_1 - 2x_2 - 1)_+$, $c_\lambda(2x_2 - 2x_1 - 1)_+$, and $a_{11}(2x_1 + 2x_2 - 3)_+$ activate precisely on $10, 01, 11$. They attain the radial bound with strict margins; wider SAEs can append silent rows. The bound holds for every width and nonnegative response class, proving global optimality of the construction.

The squared error is $2s_\lambda^2$ at zero and λ^2 elsewhere, giving $\mathrm{MSE} = 9s_\lambda^2/8 + 7\lambda^2/16$ and singleton cosine $(1 - s_\lambda)/\rho_\lambda$. Radial equality confines each used row of every global optimum to one event, even with parallel duplicates. Singleton precision is one and recall $3/4$, giving F1 $6/7$; a pair row has F1 $2/5$. Combining singleton and pair responses detects each occurrence exactly, but the two

detectors reuse the pair response. They therefore do not form disjoint generating-contribution groups. At $\lambda = 0$, equality no longer restricts expected activation and the generating reconstruction is also optimal.

Replacing d_{11}, g_{11} by two identical unit columns with responses $g_{11}/2$ preserves reconstruction and expected activation. Each half-response is still an affine ReLU. The resulting width-four global optimum has $\|g(X)\|_0 = Z_1 + Z_2$ on every input, while retaining the individual F1 ceiling. Duplicating the pair row costs no extra L_1 activation but adds one active row on every pair. L0 pricing charges for that copy, while a TopK quota may exclude it. Near-perfect reconstruction is a limit: at positive λ, shrinking all codes of a nonconstant exact reconstruction reduces coefficient cost to first order and adds squared error only to second order.

The limiting code and the edit. As $\lambda \downarrow 0$, this optimum approaches the exact binary code (3). The main text gives its reader and rerouting edit. Deleting the singleton and pair outputs that read Z_1 also removes $Z_1 Z_2 v$, so the excess squared removal error has expectation $1/16$. Raising the companion output restores the selective operation; reductions alone cannot do so.

Self-consistency of event responses. Zhu et al. (2026) penalize the decode–encode cycle discrepancy $C = \mathbb{E}\|g(f(X)) - g(X)\|^2$. Exact reconstruction makes $C = 0$ for any deterministic encoder. Use the exact code (3), with $b = 0$ and $D = [u, v, (u + v)/\sqrt{2}]$; the affine-ReLU vertex gates realize them. They reconstruct every input and have $C = 0$, while the singleton responses still miss every pair occurrence. Augment (5) by ηC, with any $\eta \geq 0$. This SAE has loss $\lambda(3/8 + \sqrt{2}/16)$. For any dictionary composed only of the generating rays, including duplicates, the unaugmented loss is at least $\lambda/2 - \lambda^2/3$. With both directions available, optimal nonnegative codes and shared bias attain this bound: each coordinate has bias $\lambda/3$ and code $(1 - 4\lambda/3)Z_i$. Consequently the event SAE has strictly lower augmented loss whenever $0 < \lambda < 3(2 - \sqrt{2})/16$, regardless of the cycle weight. Independent copies in orthogonal planes add these strict gaps, giving the same comparison for any even number of generators. Self-consistency can improve inference within a dictionary while leaving this joint allocation incentive intact.

J RESIDUAL ENERGY AND OCCURRENCE COVERAGE

We first test whether residual diagnostics register the trained allocation change, then derive what they determine about occurrence coverage under exact retained strengths.

J.1 RESIDUAL DIAGNOSTICS ON THE CONFIRMATION MODELS

Evaluate all 288 models from Section 4, without recovery-based selection. Fit an affine map from input to residual on 4,096 fresh examples and test on 8,192. Normalize residual energies by input energy and contribution error by coefficient energy; report each coupled-minus-independent cohort change relative to its independent mean.

Source	Gate	Contribution	Total residual	Nonlinear residual
Complete	L1	+843.3%	-5.1%	+5.5%
Complete	JumpReLU	+616.9%	+3.0%	+4.0%
Complete	Top4	+368.9%	+3.5%	+2.3%
Overcomplete	L1	+130.4%	-1.6%	+3.4%
Overcomplete	JumpReLU	+136.0%	+2.3%	+3.4%
Overcomplete	Top4	+206.2%	-8.6%	-7.6%

These are changes of cohort means. For overcomplete JumpReLU, the nonlinear-energy difference is .00690 in input-energy units, with dictionary-bootstrap interval $[.00269, .01050]$ and individual effects up to .01403, while contribution allocation changes much more relative to baseline.

The trained models need not satisfy the common-retention assumptions used in the exact calculation below. Moreover, observing a scalar parabolic relation would not by itself identify omitted features. As a control, rescale a code by $\gamma > 0$, preserving its information and active entries. After refitting the affine residual predictor, $N' = \gamma^2 N$ because the input and added constant project out. Calibration-

selected gains near .95 reduce N by about 10% in selected JumpReLU controls. The derivations below isolate what omission itself implies.

J.2 THE RESIDUAL MEASUREMENT

Let $Q_X h$ be what remains after subtracting the best affine prediction of a square-integrable vector h from X. For $f = b + Dg$,

$$Q_X(X - f) = -DQ_X g, \qquad N = \mathbb{E}\|DQ_X g\|^2. \tag{J.1}$$

Error outside the decoder span is affine in $X - b$; nonlinear error comes from decoded response functions, which may cancel. Thus nonlinear residual energy measures predictability from the original input; it does not test whether the SAE reconstruction can be inverted.

Whole-feature omission depends on geometry. Fixed-subset retention is affine for a full-column-rank source, but can require nonlinear reading when the source is overcomplete. For $X = Z_1 e_1 + Z_2 e_2 + Z_3(e_1 + e_2)/\sqrt{2}$ with independent Bernoulli$(1/2)$ coefficients, a piecewise-linear reader recovers Z_1, while retaining only $Z_1 e_1$ has nonlinear residual energy $1/16$. Thus the zero-nonlinearity statement below relies on orthogonal geometry.

J.3 AN EXACT FINITE LAW FOR THE PAIRED-RAY SOURCE

Balanced signs yield an exact calculation without a population limit while allowing dependent magnitudes. Let $X = Wx$, with $x_i = \epsilon_i A_i$, $A_i \geq 0$, and balanced signs independent of the amplitude vector. Let I_i depend on amplitudes and supports but not signs, set $f = W(x_i I_i)_i$, and define $t_i = \mathbb{E}A_i^2 > 0$, $q_i = \mathbb{E}[A_i^2 I_i]/t_i$, and

$$V = W \operatorname{diag}(t_i) W^\top, \qquad B = W \operatorname{diag}(t_i(1 - q_i)) W^\top. \tag{J.2}$$

Sign averaging gives zero means and $\operatorname{Cov}(r) = \operatorname{Cov}(r, X) = B$ for $r = X - f$. The affine regression identity therefore yields

$$\mathcal{E} = \operatorname{tr} B, \qquad N = \operatorname{tr} B - \operatorname{tr}(BV^\dagger B). \tag{J.3}$$

These identities hold for arbitrary W. To realize f as an SAE function of X, independent columns suffice: dual readings implement symmetric Jump thresholds, even in oblique geometry. For unit independent columns, $W^\top V^\dagger W = \operatorname{diag}(t_i^{-1})$, hence $N = \sum_i t_i q_i(1 - q_i)$. For arbitrary W and common q, $B = (1 - q)V$ and $N/\mathcal{E} = q$ exactly.

Held-out verification with overlapping geometry. On the matched pair sources of Appendix G.7, retain exact strengths above $\tau \in \{.5, 1, 2\}$. Compare orthogonal $(d, F) = (16, 32)$ with overlapping $(16, 64)$ and $(32, 128)$ sources at copula correlations $0, .99$. Unit-mean exponentials give

$$q(\tau) = e^{-\tau}(\tau^2 + 2\tau + 2)/2, \qquad \mathbb{P}(B > \tau) = e^{-\tau}.$$

With $T = 6.4$, the prediction is $\mathcal{E} = (1 - q)T$ and $N = q(1 - q)T$. Across 18 conditions, fitting and testing on separate 65,536-sample sets gives a largest relative discrepancy of .65%. This validates the selection calculation under exact magnitudes and sign-independent selection.

J.4 THE ENERGY LAW FOR OCCURRENCE SELECTION

Without sign symmetry, the intercept and other coefficients explain part of a selected response. For independent nonnegative coefficients with means μ_i, second moments $t_i > 0$, variances $v_i > 0$, and orthonormal generators, let $h_i = \alpha_i \mathbf{1}_{A_i}$, where A_i may depend on all coefficients. Put $q_i = \mathbb{E}h_i^2/t_i$, $c_i = \mathbb{E}h_i$, and $C_{ij} = \operatorname{Cov}(h_i, \alpha_j)$. Then

$$N = \sum_i t_i q_i(1 - q_i) - \sum_i \frac{t_i}{v_i}(c_i - \mu_i q_i)^2 - \sum_{i \neq j} \frac{C_{ij}^2}{v_j}. \tag{J.4}$$

Orthogonality separates output coordinates and independence diagonalizes input covariance, giving error $t_i q_i - c_i^2 - \sum_j C_{ij}^2/v_j$. Substitute $C_{ii} = t_i q_i - \mu_i c_i$ and complete the square. The two corrections are affine predictability from the mean and from other coefficients; the latter vanishes for independent coordinate thresholds but can remain under TopK.

Put $p_i = \mathbb{P}(\alpha_i > 0)$, $p_{\max} = \max_i p_i$, $U_i = t_i q_i (1 - q_i)$, and $U = \sum_i U_i$. The total correction obeys

$$0 \le U - N \le \sum_i p_i U_i \le p_{\max} U. \tag{J.5}$$

Let P project onto affine functions of the coefficients and set $u_i = h_i - q_i \alpha_i$. Then $\langle u_i, \alpha_i \rangle = 0$, $\|u_i\|_{L^2}^2 = U_i$, $u_i = Z_i u_i$ for $Z_i = \mathbf{1}_{\{\alpha_i > 0\}}$, and $N_i = U_i - \|P u_i\|_{L^2}^2$.

Write $r_i = \mu_i^2 / t_i$. Every affine function orthogonal to α_i has the form $\ell = a(1 - \mu_i \alpha_i / t_i) + b^\top (\alpha_{-i} - \mu_{-i})$. Independence, with $V_b = \mathbb{E}[(b^\top (\alpha_{-i} - \mu_{-i}))^2]$, gives

$$\|\ell\|_{L^2}^2 = a^2 (1 - r_i) + V_b,$$
$$\|Z_i \ell\|_{L^2}^2 = a^2 (p_i - r_i) + p_i V_b \le p_i \|\ell\|_{L^2}^2.$$

Thus $|\langle u_i, \ell \rangle| \le \sqrt{p_i U_i} \|\ell\|_{L^2}$. Taking $\ell = P u_i$, which is orthogonal to α_i, gives $\|P u_i\|_{L^2}^2 \le p_i U_i$ and proves (J.5). Hence discrepancies confined to rare occurrences have little additional affine predictability.

The constant is sharp: for independent Bernoulli(p) coefficients and $h_1 = h_2 = Z_1 Z_2$, $U = 2p^2(1 - p)$ and $N = (1 - p)U$. Orthogonality gives $\mathcal{E} = \sum_i t_i (1 - q_i)$, while $\sum_i \mu_i^2 \le p_{\max} T$ implies $V_X / T \to 1$ as $p_{\max} \to 0$, without fourth moments or identical laws.

Optimizing a constant bias leaves N unchanged and subtracts at most $p_{\max} T$ from total error, so the zero-bias identity remains asymptotically valid.

Writing $\omega_i = t_i / T$, $\bar{q} = \sum_i \omega_i q_i$, and $e = \mathcal{E} / T$, the identity becomes $N/T = e(1 - e) - \mathrm{Var}_\omega(q_i) - \Delta$, where $\Delta = (U - N)/T \ge 0$. At common $0 < q < 1$, $\mathcal{E} = (1 - q)T$ and

$$(1 - p_{\max}) q \le N / \mathcal{E} \le q.$$

The Bernoulli example makes $1 - p_{\max}$ sharp. Ratios use total energies before division.

For iid coefficients, exact TopK energy retention is $q_i = \mathbb{E}[B^2 \mathbb{P}(\mathrm{Binomial}(n - 1, p\mathbb{P}(B' > B)) \le k - 1 \mid B)] / \mathbb{E}B^2$; a threshold θ gives $q_i = \mathbb{E}[B^2 \mathbf{1}\{B > \theta\}] / \mathbb{E}B^2$; replacing B^2 by one gives recall. Figure 7 evaluates these formulas for $n = 128$, $p = 4/128$, and TopK quotas $1, 2, 4, 8$.

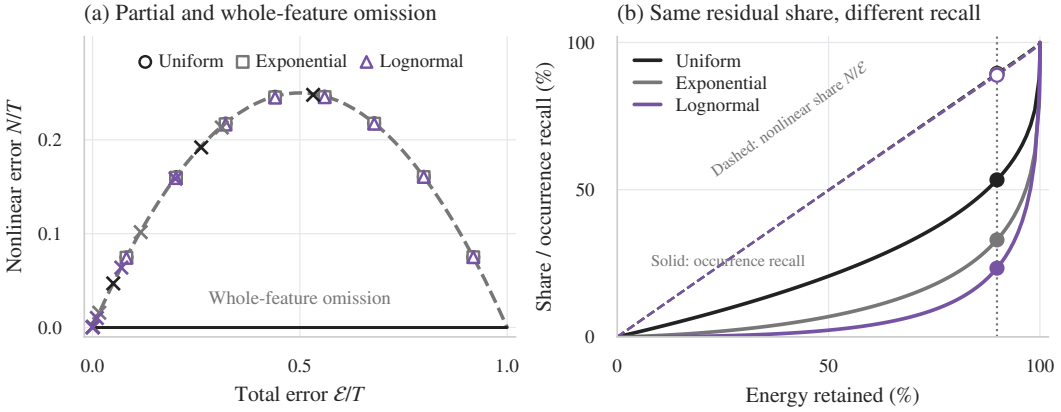


Figure 7: **Residual energy does not determine occurrence coverage.** Independent nonnegative coefficients on 128 orthogonal generators have expected activity four. Threshold and TopK selection retain exact magnitudes. At 89.84% retained energy, occurrence recall ranges from 23.3% to 53.3% across magnitude laws, while $N/\mathcal{E}$ remains between 88.9% and 89.5%. Whole-feature omission has zero nonlinear error in this geometry.

For exponential magnitudes at $q = .9$, $N/\mathcal{E}$ is $.7644, .8419, .8935$ at occurrence probabilities $.5, .25, 1/32$: the diffuse approximation requires $p_{\max} \to 0$.

Heavy tails separate retained energy from occurrence coverage. For Pareto-I magnitudes with minimum b_0 and $\nu > 2$, retaining amplitudes above $t \ge b_0$ gives $r_{\mathrm{occ}} = (b_0/t)^\nu$ and energy retention $q = (b_0/t)^{\nu - 2}$, hence

$$r_{\mathrm{occ}} = q^{\nu/(\nu-2)}.$$

At $q = .9$, coverage is .10942 for $\nu = 2.1$ and 6.35×10^{-10} for $\nu = 2.01$. Near the second-moment boundary, energy and residual diagnostics therefore cannot certify coverage; for $2 < \nu \leq 4$, squared magnitudes also have infinite variance.

J.5 The same curve can arise through calibration

Let $r_c = r - \mathbb{E}r$, $X_c = X - \mathbb{E}X$, $E_c = \mathbb{E}\|r_c\|^2$, $V_X = \mathbb{E}\|X_c\|^2 > 0$, and $C = \mathbb{E}[r_c^\top X_c]$. The scalar predictor $(C/V_X)X_c$ gives

$$N \leq E_c - \frac{C^2}{V_X}. \tag{J.6}$$

The full affine predictor can only improve this bound.

Bias and gain calibration.

The same parabolic shape follows from calibration alone. For $f = b + h$ and $r = X - f$, stationarity under $b \mapsto b + v$ and $h \mapsto sh$ gives $\mathbb{E}r = 0$ and $\mathbb{E}r^\top h = 0$, hence $E_c = C = \mathcal{E}$. Substitution in (J.6) gives

$$N \leq \mathcal{E} - \mathcal{E}^2/V_X, \qquad V_X = \mathbb{E}\|X - \mathbb{E}X\|^2 > 0. \tag{J.7}$$

This applies to TopK under common positive score scaling and to JumpReLU when scores and thresholds scale together. Bias changes can be offset in encoder intercepts. Auxiliary objectives may change stationarity; for L1, the gain derivative instead gives $\mathbb{E}r^\top h = \lambda \mathbb{E}\sum_j g_j$ and the stronger bound $N \leq \mathcal{E} - (\mathcal{E} + \lambda \mathbb{E}\sum_j g_j)^2/V_X$.

Least-squares bias and gain fitting can place any square-integrable nonlinear response under (J.7). A parabolic curve alone therefore identifies neither a gate nor exact inference; a violation excludes a justified calibration premise.

K Public-checkpoint coding audit

The audit measures how much a checkpoint's own coding objective can improve at its existing decoder. It does not infer a generating dictionary. This appendix supplies the sampling rule, the preserved loss, the certificates, and the inference/refitting controls behind the checkpoint audit in Section 5.

K.1 Checkpoints, data, and the preserved objective

The checkpoints are `gpt2-small-res-jb` SAEs at layers 1, 4, 8, and 11. Inputs are the `hook_resid_pre` activations, without normalization. Each has 24,576 affine-ReLU latents and a 768-dimensional decoder. One streaming activation store is consumed in order, in batches of 32 OpenWebText sequences of 128 tokens. The first 16 batches supply 65,536 selection tokens, the next 16 supply 65,536 calibration tokens, and 128 later batches supply 524,288 evaluation tokens. These occupy disjoint stream positions.

The coding subset contains 512 positions sampled uniformly without replacement from the 4,096 positions of the first evaluation batch, under a seed chosen before the run, using the same positions at every layer. Thus the coding brackets use 32 contexts; explained variance and rowwise statistics use the larger evaluation sets. The original position array was not retained and the upstream corpus revision was not pinned. The sampling rule can be repeated, but those exact historical inputs cannot be reconstructed from the surviving record.

The checkpoint loss is

$$L_\theta = \mathbb{E}_i \left[\frac{\|x_i - b - Dg_\theta(x_i)\|^2}{d\|x_i - \bar{x}_{\text{batch}}\|} + 8 \times 10^{-5} \|g_\theta(x_i)\|_1 \right]. \tag{K.1}$$

We preserve these batch-based target weights. If original decoder column j has norm c_j, replacing it by d_j/c_j and its activation by $c_j g_j$ preserves reconstruction and gives the transformed penalty $p_j = 8 \times 10^{-5}/c_j$. The reconstruction weight is $\omega_i = 2/[d\|x_i - \bar{x}_{\text{batch}}\|]$. The unit-column problem is therefore $\frac{1}{2}\omega_i \|x_i - b - Dz_i\|^2 + p^\top z_i$. This change of coordinates preserves the checkpoint objective rather than replacing it with a common unweighted penalty.

K.2 Numerical bounds and local opportunities

At fixed decoder and bias, the dual in Appendix D.2 becomes

$$\max_{D^\top y_i \le p} \left[y_i^\top (x_i - b) - \frac{\|y_i\|^2}{2\omega_i} \right].$$

A nonnegative code supplies a primal upper value and any feasible dual vector supplies a lower value. Subtracting both from the checkpoint loss brackets the available improvement. These are numerical optimization brackets on the evaluated inputs, rather than confidence intervals for a text population.

Table 8: **Available coding improvement in the four public checkpoints.** Gap and gap-closed columns use the fixed 512-position coding subset. EV and silent-response columns use the larger evaluation set. The gap denominator is checkpoint loss. Silent share divides $\sum_j Q_j^{\text{silent}}$ by $\sum_j Q_j$; silent value divides it by $\sum_j U_j$. The two denominators answer different questions.

Layer	EV	Gap	Silent share	Silent value/U	Gap closed
1	.9568	5.21–9.10%	88.7%	12.4%	10.1–17.6%
4	.9964	6.57–8.36%	80.9%	20.4%	19.5–24.8%
8	.9892	9.79–11.07%	76.9%	32.2%	27.5–31.1%
11	.9614	13.26–15.01%	87.4%	60.4%	29.5–33.4%

The row calculation in Appendix D.1 applies pointwise after multiplying by ω_i and replacing λ by p_j/ω_i. Write U_j for the improvement from the best response relative to an inactive latent, V_j for the improvement its current response delivers, and $Q_j = U_j - V_j$ for the improvement left by that response. Restricting Q_j to inputs where the current response is zero gives Q_j^{silent}. Silent responses account for 77–89% of the improvements available by optimizing one latent at a time. These opportunities overlap: $\sum_j Q_j$ exceeds the coordinated coding gap by factors of 7–15. The per-latent quantities locate the missed improvement; the primal–dual bracket measures the attainable total.

Uniformly rescaling inputs, reconstructions, and codes under (K.1) scales both loss terms and U, V, Q together, leaving Q/U unchanged. Increasing the effective thresholds by 25% reduces Q/U within every checkpoint. Layer 4 at multiplier 1.25 and layer 8 at .75 have near-matched median thresholds 1.93 and 1.81 yet give $Q/U = .180$ and .551. These controls rule out those scale/threshold explanations for the measured difference between checkpoints; no depth effect is inferred from them.

K.3 Iterative coding and encoder refits

Starting from the checkpoint codes and holding D, b fixed, we take 32 nonnegative soft-thresholding steps. Equivalently, for the sample objective divided by ω_i, the update is

$$z_i^+ = \left[z_i + \tau D^\top (x_i - b - D z_i) - \tau p/\omega_i \right]_+, \qquad \tau = \frac{.99}{1.01 \lambda_{\max}(D^\top D)}.$$

The iteration count is fixed in advance. The objective decreases monotonically and closes the fractions in Table 8.

For affine refitting, freeze the decoder and its bias and train the encoder weights and bias for 1,024 Adam steps on calibration tokens with batch size 256. One run begins from the checkpoint at learning rate 10^{-4}; the other uses rate 3×10^{-4} after Gaussian perturbations of 10^{-3} times each tensor's standard deviation. On 65,536 held-out evaluation tokens, both runs at every layer finish worse than the checkpoint by 3×10^{-5} to 4×10^{-3} of objective. These finite searches find no held-out improvement but do not optimize the nonconvex encoder class. The certified total gap therefore remains divided between class restrictions and imperfect fitting.