SUPPORT-CONDITIONED RECOVERY AND IDENTIFIABILITY IN NEURAL FEATURE SUPERPOSITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Superposition creates a deceptively simple question: when does a sparse feature dictionary support claims about features rather than merely provide a useful reconstruction? We separate three logically different problems—population identification of feature directions, finite-sample estimation, and coefficient recovery on the supports a representation actually visits. For recovery, we introduce a support-conditioned exact-recovery rate and a computable certificate that exposes both within-support interference and competition from inactive atoms; familiar coherence and Welch criteria emerge only as worst-case corollaries. For identification, fourth-order cumulants determine the dictionary under deterministic Kruskal and generic symmetric-tensor conditions, while Gaussian source subspaces give a concrete non-identifiability mechanism. A global-to-local analysis yields parametric finite-sample accuracy once basin separation and local conditioning hold; departures from independent sources appear as measurable cross-cumulant bias and admit a goodness-of-fit diagnostic. Experiments are organized by what each dataset can actually identify. Controlled support-law sweeps, overcomplete recovery studies, and bottleneck audits show that worst-pair geometry, statistical identifiability, optimizer success, and operational recoverability can move independently. The result is an audit framework whose claims are deliberately no stronger than its observables.

1 INTRODUCTION

The central difficulty in neural feature superposition is not simply that features overlap. It is that distinct inferential tasks are often compressed into one notion of feature recovery. A decoder can have favorable geometry while its directions remain unidentifiable from the activation law; an identifiable dictionary can be poorly conditioned or difficult to estimate from finite data; and a dictionary that violates a global coherence bound can still recover every support that a network actually visits. These possibilities are mathematically compatible. Treating them as one question obscures both the evidence required for a claim and the mechanism responsible when that claim fails.

This paper develops a claim hierarchy that keeps those questions separate. *Identifiability* is a population statement: does the law of the observed activations determine the feature directions, up to unavoidable symmetries? *Estimation* asks whether that identifiable object can be recovered stably from finitely many samples. *Operational recovery* conditions on a fixed dictionary and asks whether coefficients are uniquely recoverable on the support law realized in use. Decoder-only geometry is weaker still, while downstream predictiveness is a separate behavioral property. The hierarchy matters because no implication runs automatically in the direction that interpretability arguments often assume.

We study the overcomplete sparse linear representation model in which observed activations are generated by a unit-column dictionary, sparse latent coordinates, and isotropic noise. The decisive change of viewpoint is to treat the support distribution as part of the object being audited. A worst-case frame summary asks what happens on the most adversarial support whether or not that support is ever realized. Our support-conditioned quantities instead ask what happens on the supports the representation actually visits, while retaining classical worst-case coherence as a conservative endpoint.

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.

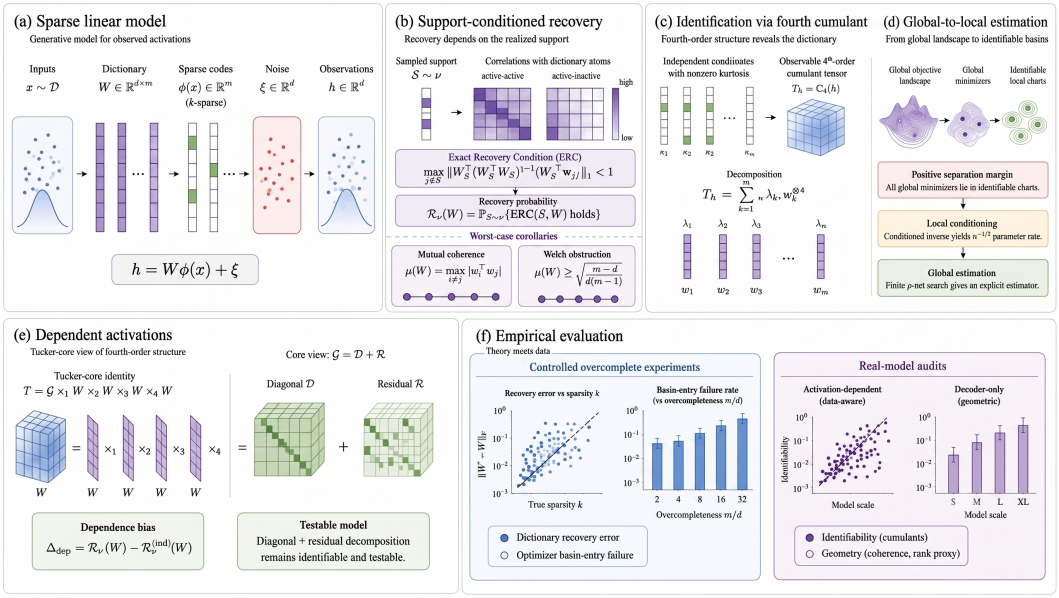


Figure 1: **Claim-aligned audit of superposed features.** The pipeline separates (a) the sparse representation model, (b) support-conditioned coefficient recovery, (c) population identification from fourth-order statistics, (d) global-to-local finite-sample estimation, (e) dependence as a testable cumulant-core deviation, and (f) experiments whose conclusions are matched to the available observables.

Our contributions are fourfold. First, we introduce an exact support-conditioned recovery rate and a computable coherence certificate that separates interference among active atoms from competition by inactive atoms. Second, we identify the dictionary from fourth-order activation statistics under deterministic Kruskal or generic symmetric-tensor conditions, and isolate Gaussian source subspaces as a rotational obstruction. Third, we separate global basin capture from local inverse conditioning in finite-sample recovery and make source dependence visible as a measurable cross-cumulant bias rather than an implicit assumption violation. Fourth, we organize the experiments by the strength of the observable: controlled data test ground-truth recovery, matched activations support support-conditioned diagnostics, and decoder-only checkpoints are restricted to geometry-only conclusions. Figure 1 summarizes this logic.

Appendix map. The supplement is deliberately navigable rather than archival. Notation and key identities are collected in section A; the logical dependencies among the appendix arguments are summarized in section B; admissible source families in section C; auxiliary geometry and concentration tools in section D; support-recovery proofs in sections E.1 to E.3; and cumulant-identification proofs in sections F.1 to F.3. Finite-sample feature discovery is split into robust recovery, the information lower bound, and global-to-local upper bounds in sections G.1 to G.3. Geometry beyond the worst-case endpoint and the downstream-estimation consequence appear in sections H to J. Experimental construction, reporting conventions, synthetic recovery, public-SAE audit, benchmark activations, architectures, metrics, and protocol summary are indexed in sections K.1 to K.8. Decoder-only, activation-conditioned, controlled-recovery, coherence-training, statistical-identification, method-comparison, and sample-grid results are in sections L.1 to L.7; reproducibility details and practical pseudocode close the appendix in sections M and N.

2 Related Work

Feature superposition and sparse autoencoders. Superposition models learned representations as overcomplete feature collections (Elhage et al., 2022); scalable sparse autoencoders recover candidate directions from language-model activations (Gao et al., 2025; Rajamanoharan et al., 2024). Recent evaluations distinguish reconstruction, interpretability, and control (Makelov et al., 2025; Karvonen

et al., 2025), while work on feature geometry, sparsity misspecification, cross-run consistency, and identifiability shows why decoder inspection alone is insufficient (Hindupur et al., 2026; Chanin and Garriga-Alonso, 2025; Song et al., 2026; Nelson et al., 2026; Klindt et al., 2026). Our focus is narrower and more operational: when the directions are observationally identified, and when coefficients are recoverable under the support law the representation actually realizes.

Sparse recovery and dictionary learning. Classical sparse approximation separates uniqueness, exact recovery, coherence, and algorithmic guarantees (Donoho and Elad, 2003; Tropp, 2004); average-case analyses study typical subdictionaries (Tropp, 2008). Welch bounds and Grassmannian packings characterize worst-case frame geometry (Welch, 1974; Strohmer and Heath Jr, 2003; Conway et al., 1996), while dictionary-learning theory establishes recovery under structural and sparsity assumptions (Gribonval and Schnass, 2010; Arora et al., 2014; Anandkumar et al., 2015). We retain those worst-case guarantees but place them after a support-conditioned quantity, where they function as distribution-free corollaries rather than proxies for the operational question.

Cumulants and tensor identifiability. Higher-order cumulants break rotational ambiguity for independent non-Gaussian sources (Comon, 1994; Hyvärinen and Oja, 2000). We combine CP uniqueness and tensor-learning results (Kruskal, 1977; Sidiropoulos and Bro, 2000; Anandkumar et al., 2014; Schramm and Steurer, 2017) with generic symmetric identifiability (Chiantini et al., 2017), local conditioning (Vannieuwenhoven, 2017), and robust estimation (Minsker, 2015; Joly and Lugosi, 2016; Lugosi and Mendelson, 2019). The distinction here is architectural: population uniqueness, basin entry, local conditioning, and source dependence are stated as separate assumptions or diagnostics, so a successful numerical factorization is never used as retrospective evidence that the population model was identifiable.

3 Support-Conditioned Recovery and Feature Identifiability

3.1 Model and Operational Recovery

Let $W = [w_1|\cdots|w_K] \in \mathbb{R}^{d\times K}$ have unit-norm columns and $K > d$. We observe

$$h = W\phi(x) + \xi, \qquad \xi \sim \mathcal{N}(0, \sigma^2 I_d), \qquad \xi \perp x, \tag{1}$$

where $\phi(x)$ is sparse. Let $S(x) = \mathrm{supp}(\phi(x))$ and let ν denote the law of $S(x)$. The empirical support law is the uniform measure over supports observed in a finite sample. Sparse decoding assumes W known and asks whether coefficients are recoverable on a realized support; dictionary identifiability, treated in section 3.3, asks whether the law of h determines W in the first place.

For a support S, Tropp's exact-recovery condition (ERC) is positive when every coefficient vector supported on S is the unique basis-pursuit solution of its noiseless observation (Tropp, 2004). We use it to define the population quantity

$$\mathcal{R}_\nu(W) = \Pr_{S\sim\nu}\{\mathrm{ERC}(S; W) > 0\}. \tag{2}$$

Unlike a coefficient-distribution success rate, $\mathcal{R}_\nu$ is uniform over coefficient values once the support is fixed. It therefore answers a clean operational question: how often does the realized support admit a support-wise guarantee?

To obtain a transparent sufficient condition, write $\mu_{\mathrm{in}}(S; W)$ for the largest inner product among active atoms and $\mu_{\mathrm{out}}(S; W)$ for the largest active–inactive inner product. For $r = |S|$, define the margin

$$\Gamma(S; W) = 1 - \frac{r\mu_{\mathrm{out}}(S; W)}{1 - (r-1)\mu_{\mathrm{in}}(S; W)} \tag{3}$$

when the denominator is positive, and $-\infty$ otherwise; let $\mathcal{C}_\nu(W) = \Pr_{S\sim\nu}\{\Gamma(S; W) > 0\}$.

Proposition 3.1 (Support-specific coherence certificate)**.** *For every unit-column dictionary and support law, $\mathcal{R}_\nu(W) \geq \mathcal{C}_\nu(W)$. If ν is supported on sets of size at most s and the global mutual coherence satisfies $\mu_W < (2s-1)^{-1}$, then both rates equal one.*

The first statement is support dependent and directly accounts for inactive competitors; the second is the familiar distribution-free endpoint. For comparison, if ν_s is uniform over all size-s supports,

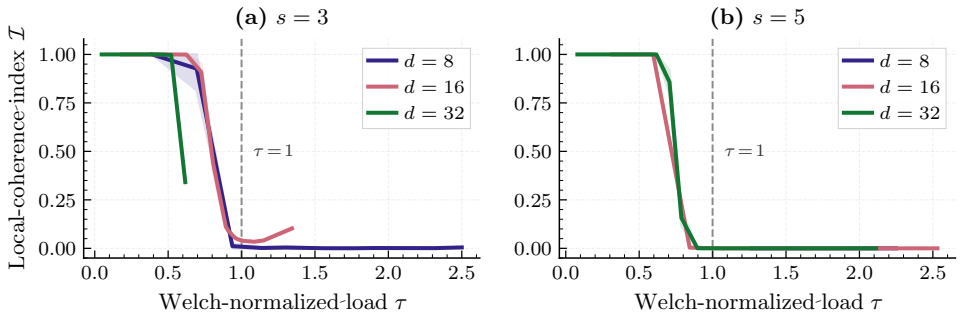


Figure 2: **Worst-case coherence phase transition.** The horizontal axis is the Welch-normalized load $\tau(d, K, s)$ and the vertical axis is the local-coherence index $\mathcal{I}(W, s)$. Random spherical dictionaries move sharply from universal certification to failure near $\tau = 1$; partial support-conditioned recovery can remain high beyond this distribution-free endpoint.

define the local-coherence index $\mathcal{I}(W, s)$ as the fraction of supports whose active subdictionary has coherence below $(2s-1)^{-1}$. Its endpoint has the exact worst-case meaning

$$\mathcal{I}(W, s) = 1 \iff \mu_W < (2s-1)^{-1} \implies \mathcal{R}_\nu(W) = 1 \quad \text{for every } \nu \text{ with } |S| \le s. \tag{4}$$

Partial values of $\mathcal{I}$ are geometry-only summaries: they ignore inactive atoms and must not be read as $\mathcal{R}_\nu$ or $\mathcal{C}_\nu$. Complete definitions and the null-space proof are in sections E and E.1.

3.2 Worst-Case Capacity as a Corollary

The support-conditioned rate may remain high even when a small set of unused supports is pathological. Global capacity is therefore deliberately stronger: it asks whether one dictionary certifies every support law with sparsity at most s. Let $a_s = (2s-1)^2$ and $\tau(d, K, s) = ((K-d)a_s)/(d(K-1))$, the squared Welch lower bound normalized by the squared recovery threshold.

Lemma 3.2 (Uniform coherence certificate)**.** *If $\mu_W < (2s-1)^{-1}$, every vector with at most s nonzero coordinates is the unique sparsest representation of its image and is recovered by basis pursuit.*

Theorem 3.3 (Worst-case coherence capacity)**.** *For $s \ge 2$, define*

$$K_{\mathrm{W}}(d, s) = \begin{cases} \infty, & d \ge a_s, \\ \left\lceil \dfrac{d(a_s - 1)}{a_s - d} \right\rceil - 1, & d < a_s. \end{cases} \tag{5}$$

If $K > K_{\mathrm{W}}(d, s)$, the Welch bound rules out the global coherence certificate for every unit-column dictionary. Conversely, whenever a real equiangular tight frame exists below the threshold it attains the certificate; random spherical dictionaries attain it with high probability in the corresponding logarithmic high-dimensional regime. Thus the obstruction is a worst-case design limit, not a statement that the realized support law is unrecoverable.

The full Welch argument, finite-dimensional edge cases, and the uniform-support envelope beyond the endpoint are given in sections D, E.3, H and I.

3.3 Identifiability from Activation Distributions

Recovery presumes that the dictionary being decoded is itself a meaningful population object. For the exact identification model we assume mutually independent latent coordinates with positive variances, nonzero fourth cumulants, and finite eighth moment; sparse activation may make individual variances small, and no K-uniform variance lower bound is imposed.

Assumption 3.4 (Nondegenerate moments)**.** Each latent coordinate has strictly positive variance. No lower bound uniform in the overcomplete width is required.

Assumption 3.5 (Finite eighth moment)**.** The observation satisfies $(\mathbb{E}\|h\|_2^8)^{1/8} < \infty$.

Assumption 3.6 (Independent latent coordinates)**.** The centered latent coordinates are mutually independent. The support-size bound used for sparse decoding is a separate condition, not an identification assumption.

Assumption 3.7 (Fourth-cumulant signal)**.** Every centered latent coordinate has a nonzero fourth cumulant.

Concrete admissible source families and the Bernoulli–Exponential calculation are in section C.

Gaussian observation noise contributes no fourth cumulant, so independence gives a symmetric rank-one decomposition

$$\mathcal{T}_h = \sum_{k=1}^{K} \lambda_k w_k^{\otimes 4}. \tag{6}$$

Lemma 3.8 (Cumulant decomposition)**.** *Under Assumptions 3.6 and 3.7, the population fourth-cumulant tensor has the form in* (6)*.*

Theorem 3.9 (Fourth-cumulant identifiability)**.** *The weighted directions in* (6) *are unique up to signed permutation under either of two standard regimes: a deterministic Kruskal condition on the dictionary, or generic symmetric identifiability below the subgeneric quartic rank, excluding the known exceptional pair* $(d, K) = (4, 8)$*. Pairwise distinct cumulant weights and mutual-coherence assumptions are not required.*

The deterministic clause covers $K \leq 2d - 2$ when the Kruskal rank is d; the generic clause extends much further in the overcomplete regime. Sections F.1 and F.2 gives the exact inequalities and proof reductions.

Proposition 3.10 (Gaussian-subspace non-identifiability)**.** *An isotropic Gaussian latent subspace of dimension at least two admits a continuum of inequivalent unit-column dictionaries with the same observation law when the corresponding latent variances are not fixed in advance. Its directions are therefore not identifiable.*

This negative control is conceptually important: non-Gaussianity breaks rotational symmetry, whereas equal nonzero kurtosis weights do not by themselves create the same ambiguity.

Global basin capture and local conditioning. Uniqueness is not yet an estimator. Write $\Phi_4(\lambda, W) = \sum_k \lambda_k w_k^{\otimes 4}$ and search over the compact class

$$\Theta_B = \{(\lambda, W) : 0 < \lambda_{\min} \leq |\lambda_k| \leq \lambda_{\max},\ \|w_k\|_2 = 1\}. \tag{7}$$

Around a regular identifiable parameter, the local inverse has condition parameter $\gamma_* > 0$, while compactness and global uniqueness yield a strictly positive separation margin from every inequivalent basin. The global minimum-distance estimator therefore enters the correct aligned basin once the robust cumulant error is below half that margin and then achieves the usual parametric $n^{-1/2}$ rate, scaled by the fourth-order ambient dimension, the eighth-moment radius, and γ_*^{-1}. This theorem assumes global optimization; the practical cumulant–candidate–refinement routine in section N is deliberately treated as a local algorithm whose basin-entry failures remain visible. The information lower bound and the global-to-local upper bounds are separated in sections G.2 and G.3.

Dependent activations are a model deviation, not a footnote. For arbitrary centered latents, multilinearity expresses the observed fourth cumulant through the full latent cumulant core. Without structure on that core, invertible latent reparameterizations destroy dictionary identifiability. If the core is approximately diagonal, the off-diagonal cross-cumulants induce an additive population bias bounded by the dictionary operator norm and the off-diagonal core mass. The same perturbation analysis that controls sampling noise then controls this bias, and the distance from the observed cumulant tensor to the exact diagonal model provides a finite-sample goodness-of-fit check. Hence source dependence degrades the guarantee quantitatively and can be rejected empirically; it is not silently absorbed into the theorem. Full identities, misspecification certificates, and the context-conditioned extension are in section G.3.

3.4 Scope and Diagnostic Order

These distinctions impose an order of inference. The diagonal or near-diagonal cumulant model must first survive an empirical adequacy check. Population uniqueness is then a property of the

admitted model; global separation determines whether inequivalent basins remain distinguishable; and local conditioning governs how tensor error is amplified after the correct basin has been entered. Support-conditioned recovery addresses a later and different question: whether the realized code can be decoded on the supports that actually occur. Model misspecification, non-identifiability, basin collision, ill conditioning, sampling error, and adverse active–inactive correlations are therefore separate diagnoses rather than variants of a single aggregate score. The uniform-support geometry beyond the endpoint is analyzed in section H; the probe consequence in section J concerns estimation after a representation has been fixed and establishes neither identification nor causal sufficiency (Hewitt and Liang, 2019; Geiger et al., 2021; Belinkov, 2022).

Finite-sample resolution. Population uniqueness becomes statistically useful only after three quantities are controlled. Robust fourth-cumulant estimation bounds the error in the population object that carries the identifying information. Compactness and uniqueness then induce positive separation from inequivalent factorizations, which determines the scale at which basin capture is guaranteed. Inside the aligned basin, the Jacobian of the tensor map converts tensor error into dictionary error, with the inverse conditioning parameter governing amplification. Concentration, separation, and conditioning therefore address different stages of the argument. A small tensor error cannot compensate for a nearly singular local inverse, and a well-conditioned local chart cannot exclude a distant competing basin. The resulting rate is meaningful only when these quantities are reported together; the full construction appears in sections G.1 and G.3.

The lower and upper arguments deliberately live at different resolutions. The information bound in section G.2 shows that even an oracle revealing the latent coefficients must resolve the local degrees of freedom of the dictionary modulo signs and permutations. The global upper bound instead estimates a symmetric fourth-order tensor before fitting a factorization, so its ambient tensor dimension reflects a conservative proof route rather than an intrinsic fourth-order parameter count for feature discovery. After a pilot estimate enters an identifiable chart, the sample-split tangent refinement in section G.3 works directly in local dictionary coordinates. This distinction prevents an important overreading of the theory: the theorem certifies one statistically valid path from samples to the population object, while the practical routine sacrifices global optimization guarantees for computational tractability.

Optimization is kept outside the identifiability claim. The practical optimizer is not part of the identifiability theorem. Experiments use a cumulant-based candidate stage followed by projected refinement rather than an oracle global minimizer. A failed restart can therefore occur even when the population dictionary is unique, while a successful numerical fit cannot establish uniqueness of the population model. We report aligned recovery error, restart-level success, conditioning diagnostics, and Gaussian controls so these events remain distinguishable. The Gaussian negative control in table 4 preserves rotational ambiguity in the observation law; an optimization failure instead occurs after the target directions are already determined by that law. The practical routine and its relation to the analyzed estimator are given in section N.

Table 1: **Operational recovery changes continuously with the support law while the dictionary is held fixed.** The localized-defect dictionary distribution and all geometric quantities are unchanged across rows; only the mixture weight q on hazardous supports varies. Values are means over four dictionary seeds. The construction, replication unit, and basis-pursuit protocol are specified in section K.3.

Hazard q	$\mathcal{R}_{\widehat{\nu}_n}$	$\mathcal{C}_{\widehat{\nu}_n}$	Certificate gap	BP success
0.00	1.000	0.878	0.122	1.000
0.05	0.952	0.835	0.117	0.975
0.10	0.914	0.804	0.111	0.950
0.25	0.734	0.653	0.081	0.850
0.50	0.511	0.448	0.064	0.850
1.00	0.000	0.000	0.000	0.488

Table 2: **A sharp geometry transition at fixed support complements the support-law sweep.** The paired-decoy construction uses $s = 8$ and normalized separation $\widetilde{\tau} = \tau/\sqrt{s(s-1)}$. The exact support certificate and active–inactive margin coincide. At the transition the two ℓ_1 representations tie; strict positive margin restores the active solution. The construction and threshold derivation are given in section L.3.

$\widetilde{\tau}$	$\mathbf{ERC} = \boldsymbol{\Gamma}$	Active ℓ_1	Decoy ℓ_1	BP active?
0.80	−0.208	8.000	6.621	No
0.95	−0.046	8.000	7.651	No
1.00	0.000	8.000	8.000	Non-unique
1.05	0.042	8.000	8.351	Yes
1.20	0.150	8.000	9.415	Yes

Two orthogonal stress tests localize operational failure. Table 1 holds the dictionary fixed and moves probability mass onto hazardous supports; recovery and certification fall even though every decoder-only geometric statistic is unchanged. Table 2 holds the active support fixed and instead moves the inactive decoys across the exact active–inactive threshold. Below the threshold the decoy representation has smaller ℓ_1 norm, equality produces non-uniqueness, and a strictly positive margin restores the active solution. The two tables therefore vary different objects and meet at the same conclusion: operational failure is determined by the interaction between the realized support and the surrounding dictionary, not by a single global coherence number. The residual gap between the certificate and sampled basis-pursuit success in table 1 is also informative, because the certificate is uniform over coefficients on a support whereas an individual draw may still be recovered when that sufficient condition fails.

The same discipline applies one layer earlier at the population model. Cross-cumulants form an off-diagonal latent cumulant core whose induced perturbation can be compared with sampling error. A small residual yields a bias-aware neighborhood around the independent-source solution; a large residual rejects the diagonal cumulant model before any identifiability claim is made. Thus support-law failure, active–inactive geometric failure, and cumulant-model misspecification are separated by the observable that changes in each intervention rather than collapsed into one failure label.

4 Empirical Evidence

The experiments follow the same hierarchy. Controlled dictionaries expose the generating features and support law; learned bottlenecks provide matched activations, sparse codes, and decoders; public SAE checkpoints without a shared activation protocol support only decoder-geometry claims. Every reported claim is therefore paired with the strongest observable that can justify it. Protocols, replication units, optimizer budgets, uncertainty summaries, and complete result tables are in sections K and L.

Table 3: **Operational recovery can change while the dictionary geometry is fixed.** All rows use $(d, K, s) = (64, 256, 2)$; values are means over four independently generated dictionaries. The two localized-defect rows use the same dictionary construction and differ only in the support law. Construction and replication details are in section K.3.

Geometry	Support	μ_W	$\mathcal{R}$	$\mathcal{C}$	BP success
Random spherical	Uniform	0.506	1.000	0.829	1.000
Four orthobases	Uniform	0.504	1.000	0.898	1.000
Localized defect	Safe	0.999	1.000	0.878	1.000
Localized defect	Hazardous	0.999	0.000	0.000	0.488

Support laws dominate the operational conclusion. The localized-defect construction in table 3 has near-maximal mutual coherence. Yet safe supports are uniformly recovered in the reported trials, while concentrating the support law on the defect collapses both the exact support-conditioned rate and its certificate. The decoder geometry has not changed; only the support distribution has. This is the concrete failure mode that a worst-pair summary cannot represent.

Identifiability is distinct from basin entry. For $d = 8$, $s = 3$, $K \in \{10, 12, 14\}$, and Gaussian observation noise $\sigma = 0.05$, Bernoulli–Exponential sources show decreasing aligned column error with sample size. At $n = 25{,}000$, the practical routine succeeds in $5/5$, $4/5$, and $4/5$ trials, respectively, while an isotropic Gaussian source block succeeds in $0/5$. The Gaussian control realizes theorem 3.10; the remaining failures in identifiable settings are algorithmic basin failures rather than evidence against population uniqueness.

Table 4: **Finite-sample recovery behaves differently in identifiable and rotationally ambiguous source models.** Entries are median maximum aligned column error [IQR] over five trials. The success count uses the prespecified threshold $\mathcal{E}(\widehat{W}, W) < 0.3$ at $n = 25{,}000$; the two lower-sample regimes and conditioning diagnostics are in section L.5.

Source	K	n = 10k	n = 25k	Success
Bernoulli–Exp.	10	0.167 [0.147, 0.220]	0.116 [0.111, 0.123]	5/5
Bernoulli–Exp.	12	0.150 [0.132, 0.218]	0.111 [0.104, 0.124]	4/5
Bernoulli–Exp.	14	0.194 [0.176, 0.221]	0.149 [0.144, 0.197]	4/5
Gaussian control	12	1.045 [0.938, 1.066]	1.004 [0.958, 1.070]	0/5

Table 5: **Overcomplete recovery under fixed update budgets.** All methods use $(d, K, s) = (8, 12, 3)$. Entries are median maximum aligned column error [IQR] over five trials; success uses $\mathcal{E}(\widehat{W}, W) < 0.3$ at $n = 10{,}000$. The budgets are algorithm-specific; see section L.6.

Method	n = 2.5k	n = 10k	Success
CBFR	0.242 [0.192, 0.280]	0.121 [0.117, 0.121]	5/5
ℓ_1 SAE	1.078 [1.049, 1.102]	1.085 [0.988, 1.088]	0/5
TopK SAE	1.068 [1.014, 1.190]	1.114 [1.005, 1.192]	0/5
Gated SAE	0.991 [0.990, 1.046]	1.112 [1.069, 1.117]	0/5
JumpReLU SAE	1.024 [1.008, 1.051]	1.104 [1.088, 1.145]	0/5

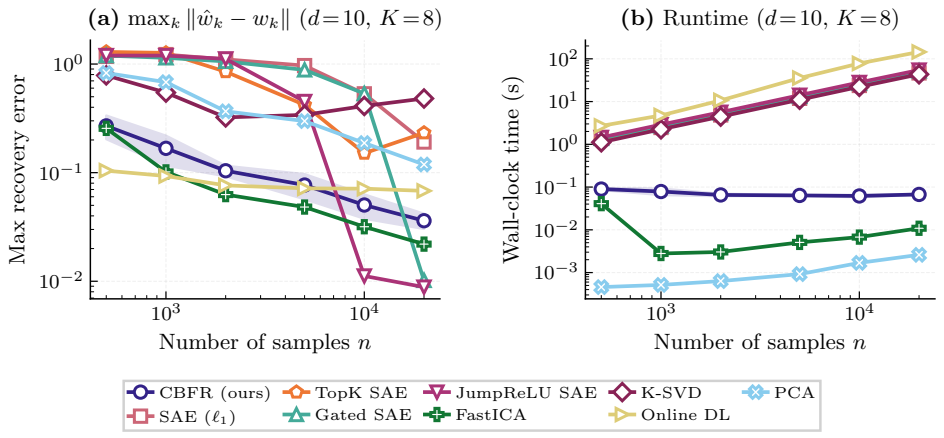


Figure 3: **Determined-regime recovery and runtime calibration.** Nine methods are compared at $(d, K) = (10, 8)$. Recovery error and wall-clock time are shown separately; the overcomplete experiment is reported in table 5.

Method comparison separates calibration from overcomplete recovery. Figure 3 checks a broad method set where $K < d$ and every method can return all requested components. Table 5 then fixes $K > d$ and the declared update budgets. CBFR meets the recovery criterion in all five $n = 10{,}000$ trials, while none of the four SAE variants does so in these trials. The comparison is specific to the stated objectives, data model, and budgets; it is not normalized by wall-clock time, floating-point cost, or hyperparameter-search effort.

Learned bottlenecks separate signal from intervention. Matched activations permit a two-stage test. Table 6 first asks whether fourth-order structure rises above a covariance-matched Gaussian baseline. Table 7 then intervenes on decoder geometry and measures the accompanying reconstruction change. The first is a distributional diagnostic; the second is a paired geometric intervention.

Table 6: **Fourth-order signal on matched activations.** Mean $\pm$ standard deviation over three split seeds. Values above one exceed the covariance-matched Gaussian baseline; they do not establish a diagonal cumulant core.

Data	Samples	Base acc.	$\mathbf{S_4}/\mathbf{S_4^{Gauss}}$
OptDigits	1,797	$97.1 \pm 0.7\%$	4.88 ± 0.57
WDBC	569	$97.0 \pm 1.4\%$	17.62 ± 1.08
Wine	178	$94.1 \pm 2.1\%$	2.22 ± 0.03
Iris	150	$87.7 \pm 8.1\%$	2.92 ± 0.53

Table 7: **Geometry–utility tradeoff under coherence regularization.** Arrows compare TopK with its regularized counterpart. Δ acc. reports the reconstructed-activation task-accuracy change in percentage points; certificates count fits with $\mu_W < 1/3$. Normalized reconstruction-error diagnostics and the split-level protocol are in table L.1 and section K.5.

Data	μ_W base→reg.	$\mathcal{I}(W,2)$ base→reg.	Δ acc. (pp)	Cert. base→reg.
OptDigits	$0.785 \to 0.334$	$0.745 \to 0.997$	-1.7	$0/3 \to 1/3$
WDBC	$0.894 \to 0.255$	$0.748 \to 1.000$	-0.7	$0/3 \to 3/3$
Wine	$0.759 \to 0.344$	$0.589 \to 0.992$	-2.2	$0/3 \to 1/3$
Iris	$0.813 \to 0.326$	$0.622 \to 0.997$	$+3.6$	$0/3 \to 2/3$

Taken together, the two tables separate a model diagnostic from a geometry intervention. The fourth-order statistic exceeds its Gaussian control on every benchmark, supplying non-Gaussian signal but not a diagonal cumulant model. Regularization then lowers mean coherence by 0.415–0.639, raises $\mathcal{I}(W,2)$ to 0.992–1.000, and moves seven of twelve fits across the strict certificate. Across datasets, the observed change in reconstructed-activation accuracy lies in $[-2.2, +3.6]$ percentage points. The geometric improvement is therefore not explained by a uniform collapse in reconstruction. It remains a statement about the learned decoder and matched activations, not source independence or semantic uniqueness.

Controls close the inferential chain. The localized-defect construction changes only the support law and isolates operational failure. The Gaussian control preserves rotational ambiguity, so more data cannot identify its latent directions. Failed non-Gaussian restarts occur after population uniqueness is available and instead expose optimization or basin entry. Public checkpoints without matched activations support geometry-only screening, not support-conditioned recovery or cumulant-model validation. These controls make similar numerical failures diagnostically distinct and keep each empirical conclusion tied to the observable that licenses it.

5 Conclusion

Superposition creates three distinct inferential tasks: identifying feature directions, estimating them from finite data, and recovering coefficients on the supports a representation actually visits. Our results show that these tasks require different observables and fail for different reasons. Fourth-order structure can identify directions when the cumulant model is adequate; finite-sample accuracy then depends on basin separation and local conditioning; operational recovery depends on the realized support law rather than worst-pair geometry alone.

This separation yields a practical audit principle: test model adequacy before invoking identifiability, separate statistical instability from optimization failure, and report recovery on the supports that occur. The framework is diagnostic by construction. Its value is not a single score, but a disciplined path from observable structure to claims whose assumptions can be checked and whose failure modes can be localized.

Reproducibility Statement

The paper separates population assumptions, finite-sample guarantees, practical optimization, and operational recovery so each claim can be checked at the appropriate level. Complete proofs and assumptions are in sections C to G; data generation, split rules, independent replicate units, optimizer budgets, metrics, and reporting conventions are in sections K and M; and the practical cumulant-based routine is specified in section N.

AI Use Statement

Generative AI tools were used to assist with language editing, manuscript restructuring, drafting and revising portions of the exposition (including the abstract), format migration, and LaTeX and table preparation. They were not treated as an evidentiary source or as an authority for mathematical correctness. The authors reviewed the AI-assisted edits against the underlying derivations and experimental outputs and take responsibility for the final text, claims, citations, figures, tables, and submitted artifacts.

References

Alekh Agarwal, Animashree Anandkumar, Prateek Jain, Praneeth Netrapalli, and Rashish Tandon. Learning Sparsely Used Overcomplete Dictionaries. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 123–137. PMLR, 2014.

Guillaume Alain and Yoshua Bengio. Understanding Intermediate Layers Using Linear Classifier Probes. *CoRR*, abs/1610.01644, 2016.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor Decompositions for Learning Latent Variable Models. *Journal of Machine Learning Research*, 15 (1):2773–2832, 2014.

Animashree Anandkumar, Rong Ge, and Majid Janzamin. Learning Overcomplete Latent Variable Models through Tensor Methods. In *Proceedings of the 28th Conference on Learning Theory*, volume 40 of *Proceedings of Machine Learning Research*, pages 36–112. PMLR, 2015.

Sanjeev Arora, Rong Ge, and Ankur Moitra. New Algorithms for Learning Incoherent and Overcomplete Dictionaries. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 779–806. PMLR, 2014.

Yonatan Belinkov. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, 48(1):207–219, 2022.

Aditya Bhaskara, Moses Charikar, and Aravindan Vijayaraghavan. Uniqueness of Tensor Decompositions with Applications to Polynomial Identifiability. In *Proceedings of the 27th Conference on Learning Theory*, volume 35 of *Proceedings of Machine Learning Research*, pages 742–778. PMLR, 2014.

Paul Breiding and Nick Vannieuwenhoven. The Condition Number of Join Decompositions. *SIAM Journal on Matrix Analysis and Applications*, 39(1):287–309, 2018.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, et al. Towards Monosemanticity: Decomposing Language Models with Dictionary Learning. *Transformer Circuits*, 2(5):6, 2023.

Bart Bussmann, Noa Nabeshima, Adam Karvonen, and Neel Nanda. Learning Multi-Level Features with Matryoshka Sparse Autoencoders. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 6077–6101. PMLR, 2025.

Emmanuel J Candès and Michael B Wakin. An Introduction to Compressive Sampling. *IEEE Signal Processing Magazine*, 25(2):21–30, 2008.

David Chanin and Adrià Garriga-Alonso. Sparse but Wrong: Incorrect L_0 Leads to Incorrect Features in Sparse Autoencoders. *CoRR*, abs/2508.16560, 2025.

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, pages 82318–82355. Curran Associates, Inc., 2026.

Luca Chiantini, Giorgio Ottaviani, and Nick Vannieuwenhoven. On Generic Identifiability of Symmetric Tensors of Subgeneric Rank. *Transactions of the American Mathematical Society*, 369 (6):4021–4042, 2017.

Pierre Comon. Independent Component Analysis, a New Concept? *Signal Processing*, 36(3): 287–314, 1994.

John H Conway, Ronald H Hardin, and Neil JA Sloane. Packing Lines, Planes, etc.: Packings in Grassmannian Spaces. *Experimental Mathematics*, 5(2):139–159, 1996.

Jingyi Cui, Qi Zhang, Yifei Wang, and Yisen Wang. On the Theoretical Understanding of Identifiable Sparse Autoencoders and Beyond. *CoRR*, abs/2506.15963, 2025.

David L Donoho and Michael Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5): 2197–2202, 2003.

Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders Find Interpretable LLM Feature Circuits. In *Advances in Neural Information Processing Systems*, volume 37, pages 24375–24410. Curran Associates, Inc., 2024.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic Probing: Behavioral Explanation with Amnesic Counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, et al. Toy Models of Superposition. *CoRR*, abs/2209.10652, 2022.

Leo Gao, Tom Dupre 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*. OpenReview.net, 2025.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal Abstractions of Neural Networks. In *Advances in Neural Information Processing Systems*, volume 34, pages 9574–9586. Curran Associates, Inc., 2021.

Rémi Gribonval and Karin Schnass. Dictionary identification—sparse matrix factorization via ℓ_1-minimization. *IEEE Transactions on Information Theory*, 56(7):3523–3539, 2010.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pages 2733–2743, 2019.

Sai Sumedh R Hindupur, Ekdeep S Lubana, Thomas Fel, and Demba Ba. Projecting Assumptions: The Duality Between Sparse Autoencoders and Concept Geometry. In *Advances in Neural Information Processing Systems*, volume 38, pages 11649–11699. Curran Associates, Inc., 2026.

Aapo Hyvärinen and Erkki Oja. Independent Component Analysis: Algorithms and Applications. *Neural Networks*, 13(4–5):411–430, 2000.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ICA Using Auxiliary Variables and Generalized Contrastive Learning. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 859–868. PMLR, 2019.

Emilien Joly and Gábor Lugosi. Robust Estimation of U-Statistics. *Stochastic Processes and their Applications*, 126(12):3760–3773, 2016.

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 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 29018–29049. PMLR, 2025.

Adam Karvonen, Can Rager, Johnny Lin, Curt Tigges, Joseph Bloom, David Chanin, Yeu-Tong Lau, Eoin Farrell, Callum McDougall, Kola Ayonrinde, et al. SAEBench: A Comprehensive Benchmark for Sparse Autoencoders in Language Model Interpretability. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 29223–29264. PMLR, 2025.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR, 2020.

David Klindt, Charles O'Neill, Patrik Reizinger, Harald Maurer, and Nina Miolane. A Unifying Framework from Neural Superposition to Sparse Interpretable Codes. *Nature Machine Intelligence*, pages 1–13, 2026.

Tamara G Kolda and Brett W Bader. Tensor Decompositions and Applications. *SIAM Review*, 51(3): 455–500, 2009.

Joseph B Kruskal. Three-Way Arrays: Rank and Uniqueness of Trilinear Decompositions, with Application to Arithmetic Complexity and Statistics. *Linear Algebra and its Applications*, 18(2): 95–138, 1977.

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*, pages 278–300, 2024.

Gábor Lugosi and Shahar Mendelson. Sub-Gaussian Estimators of the Mean of a Random Vector. *The Annals of Statistics*, 47(2):783–794, 2019. doi:10.1214/17-AOS1639.

Qi Lyu and Xiao Fu. On Finite-Sample Identifiability of Contrastive Learning-Based Nonlinear Independent Component Analysis. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 14582–14600. PMLR, 2022.

George Ma, Zhongyuan Liang, Irene Y Chen, and Somayeh Sojoudi. Do Sparse Autoencoders Identify Reasoning Features in Language Models? *CoRR*, abs/2601.05679, 2026a.

George Ma, Samuel Pfrommer, and Somayeh Sojoudi. Revising and Falsifying Sparse Autoencoder Feature Explanations. In *Advances in Neural Information Processing Systems*, volume 38, pages 63175–63217. Curran Associates, Inc., 2026b.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Towards Principled Evaluations of Sparse Autoencoders for Interpretability and Control. In *International Conference on Learning Representations*. OpenReview.net, 2025.

Samuel Marks, Can Rager, Eric 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*. OpenReview.net, 2025.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781, 2013.

Stanislav Minsker. Geometric Median and Robust Estimation in Banach Spaces. *Bernoulli*, 21(4): 2308–2335, 2015. doi:10.3150/14-BEJ645.

Anish Mudide, Josh Engels, Eric Michaud, Max Tegmark, and Christian Schroeder de Witt. Efficient Dictionary Learning with Switch Sparse Autoencoders. In *International Conference on Learning Representations*. OpenReview.net, 2025.

Walter Nelson, Theofanis Karaletsos, and Francesco Locatello. Toward Identifiable Sparse Autoencoders. *CoRR*, abs/2605.31245, 2026.

Kiho Park, Yo Joong Choe, and Victor Veitch. The Linear Representation Hypothesis and the Geometry of Large Language Models. *CoRR*, abs/2311.03658, 2023.

Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Tom Lieberum, Vikrant Varma, Janos Kramar, 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, pages 775–818. Curran Associates, Inc., 2024.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the Probing Paradigm: Does Probing Accuracy Entail Task Relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, 2021.

Tselil Schramm and David Steurer. Fast and Robust Tensor Decomposition with Applications to Dictionary Learning. In *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1760–1793. PMLR, 2017.

Nicholas D Sidiropoulos and Rasmus Bro. On the Uniqueness of Multilinear Decomposition of N-Way Arrays. *Journal of Chemometrics*, 14(3):229–239, 2000.

Xiangchen Song, Aashiq Muhamed, Yujia Zheng, Lingjing Kong, Zeyu Tang, Mona Diab, Virginia Smith, and Kun Zhang. Mechanistic Interpretability Should Prioritize Feature Consistency in Sparse Autoencoders. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2172–2210, 2026.

Daniel A Spielman, Huan Wang, and John Wright. Exact Recovery of Sparsely-Used Dictionaries. In *Proceedings of the 25th Annual Conference on Learning Theory*, volume 23 of *Proceedings of Machine Learning Research*, pages 37.1–37.18. PMLR, 2012.

Thomas Strohmer and Robert W Heath Jr. Grassmannian Frames with Applications to Coding and Communication. *Applied and Computational Harmonic Analysis*, 14(3):257–275, 2003.

Adly Templeton, Tom Conerly, Jonathan Marcus, Jack Lindsey, Trenton Bricken, Brian Chen, Adam Pearce, Craig Citro, Emmanuel Ameisen, Andy Jones, et al. Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet. *CoRR*, abs/2605.29358, 2026.

Harrish Thasarathan, Julian Forsyth, Thomas Fel, Matthew Kowal, and Konstantinos G Derpanis. Universal Sparse Autoencoders: Interpretable Cross-Model Concept Alignment. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 59304–59325. PMLR, 2025.

Joel A Tropp. Greed Is Good: Algorithmic Results for Sparse Approximation. *IEEE Transactions on Information Theory*, 50(10):2231–2242, 2004.

Joel A Tropp. On the Conditioning of Random Subdictionaries. *Applied and Computational Harmonic Analysis*, 25(1):1–24, 2008.

Nick Vannieuwenhoven. Condition Numbers for the Tensor Rank Decomposition. *Linear Algebra and its Applications*, 535:35–86, 2017.

Lloyd Welch. Lower Bounds on the Maximum Cross Correlation of Signals. *IEEE Transactions on Information Theory*, 20(3):397–399, 1974.

Xudong Zhu, Mohammad Mahdi Khalili, and Zhihui Zhu. AbsTopK: Rethinking Sparse Autoencoders for Bidirectional Features. *CoRR*, abs/2510.00404, 2025.

APPENDIX ROADMAP

Foundations and proofs. sections A to F. **Finite-sample theory.** sections G.1 to G.3. **Further geometry.** sections H to J. **Protocols.** sections K.1 to K.8. **Empirics.** sections L.1 to L.7. **Reproducibility and algorithm.** sections M and N.

A NOTATION AND KEY RELATIONS

Let $a_s = (2s-1)^2$, $t_s = a_s^{-1/2}$, $D_4 = \binom{d+3}{4}$, and let $\mathrm{Sym}^4(\mathbb{R}^d)$ carry the Frobenius inner product. The Kruskal rank $\mathrm{krank}(W)$ is the largest integer r for which every r columns of W are linearly independent. For dictionaries $W, \widehat{W}$ with the same column count, $d_{\pm,\mathrm{perm}}\left(\widehat{W}, W\right)$ denotes the minimum, over permutations and independent sign flips, of the maximum columnwise Euclidean error. For tensor parameters $\theta = (\lambda, W)$, define the aligned product metric

$$d_\Theta(\widehat{\theta}, \theta) = \min_{\pi,\, z_k \in \{\pm 1\}} \left(\left\| \widehat{\lambda}_\pi - \lambda \right\|_2^2 + \sum_{k=1}^{K} \left\| z_k \widehat{w}_{\pi(k)} - w_k \right\|_2^2 \right)^{1/2}. \tag{A.1}$$

In particular, $d_{\pm,\mathrm{perm}}\left(\widehat{W}, W\right) \le d_\Theta(\widehat{\theta}, \theta)$.

Three statements used throughout the paper should not be conflated:

$$\mathrm{Spark}(W) > 2s \Longrightarrow \text{uniform } s\text{-sparse uniqueness}, \tag{A.2}$$

$$\mathrm{ERC}(S; W) > 0 \Longrightarrow \text{BP recovery on support } S, \tag{A.3}$$

$$\mu_W < (2s-1)^{-1} \Longrightarrow \mathrm{ERC}(S; W) > 0 \quad (|S| \le s). \tag{A.4}$$

The last condition is a uniform sufficient certificate. The partial local index $\mathcal{I}(W, s)$ is a graph statistic; only its endpoint $\mathcal{I} = 1$ is equated with the global certificate. These distinctions follow the classical sparse-approximation and compressed-sensing separation between uniqueness, exact recovery, and coherence-based sufficient conditions (Donoho and Elad, 2003; Tropp, 2004; Candès and Wakin, 2008; Tropp, 2008). Related dictionary-learning results study when a sparsely used generating dictionary itself can be identified or recovered (Gribonval and Schnass, 2010; Spielman et al., 2012; Arora et al., 2014; Agarwal et al., 2014).

Example A.1 (Why a partial local score is not an exact-recovery score)**.** Let

$$W = \begin{bmatrix} 1 & 0 & 2^{-1/2} \\ 0 & 1 & 2^{-1/2} \end{bmatrix}, \qquad S = \{1, 2\}. \tag{A.5}$$

Then

$$\mu(W_S) = 0 < \frac{1}{3}, \tag{A.6}$$

$$w_3 = \frac{1}{\sqrt{2}} w_1 + \frac{1}{\sqrt{2}} w_2. \tag{A.7}$$

Thus the local pair (w_1, w_2) passes the $s = 2$ coherence test, yet the observation w_3 has both a one-sparse representation and a representation supported on S. Equivalently,

$$\left\| W_S^\dagger w_3 \right\|_1 = \left\| \begin{bmatrix} 2^{-1/2} & 2^{-1/2} \end{bmatrix}^\top \right\|_1 = \sqrt{2} > 1, \qquad \mathrm{ERC}(S; W) = 1 - \sqrt{2} < 0. \tag{A.8}$$

The external atoms of the full dictionary therefore cannot be omitted from an exact support-recovery criterion.

B LOGICAL DEPENDENCIES OF THE APPENDIX

The appendix follows the same claim hierarchy as the main paper, but supplies the arguments in the order in which they are needed. The recovery branch begins with support-wise exact recovery: the

ERC is the target statement, the support-conditioned coherence margin is a sufficient certificate, and global coherence appears only after the support law has been removed from the question. The Welch argument then limits that global certificate; it does not turn a failed worst-case certificate into a claim of operational failure.

The identification branch begins from the fourth joint cumulant of the observation. Under independent non-Gaussian coordinates, the latent cumulant core is diagonal and the observed tensor becomes a weighted symmetric rank-one sum. Tensor uniqueness is therefore a population statement. Sampling enters only after that step: robust cumulant estimation controls perturbation of the tensor, global separation places a minimum-distance fit in the correct quotient basin, and local conditioning converts tensor error into parameter error. The lower bound and upper bounds are stated separately because they answer different questions and use different information.

Dependence is treated as a model perturbation rather than absorbed into the identification theorem. A general latent cumulant core produces a Tucker transform through the dictionary; its off-diagonal part becomes a population bias in the diagonal CP model. The resulting bias appears in the same perturbation inequalities as sampling error and motivates the model-adequacy diagnostic used later in the empirical protocol. With these dependencies fixed, each subsequent proof can be read locally without changing the meaning of any claim in the main text.

C Assumptions and Admissible Source Families

Sparse geometry versus independent sources. Worst-case sparse recovery conditions concern realizations with $\|\phi\|_0 \le s$, whereas independent Bernoulli supports give $\mathbb{E}\,\|\phi\|_0 = \sum_k p_k \le s$ and a binomial tail. We use the deterministic condition for capacity and probing and the independent-coordinate model for cumulant analysis.

Variance scaling. If $\|\phi\|_\infty \le B$ and $\mathbb{E}\,\|\phi\|_0 \le s$, then

$$\operatorname{tr}\mathbb{E}[\phi\phi^\top] = \mathbb{E}\,\|\phi\|_2^2 \le B^2\mathbb{E}\,\|\phi\|_0 \le B^2 s. \tag{C.1}$$

$$\lambda_{\min}(\mathbb{E}[\phi\phi^\top]) \le K^{-1}\operatorname{tr}\mathbb{E}[\phi\phi^\top] \le \frac{B^2 s}{K}. \tag{C.2}$$

Accordingly, Assumption 3.4 requires only coordinatewise positive variance and never a fixed positive eigenvalue lower bound independent of K.

The Bernoulli–Exponential source used in the recovery experiments. Let $Z_k \sim \mathrm{Bern}(p_k)$ and $E_k \sim \mathrm{Exp}(\beta_k)$ be independent, and set $\phi_k = Z_k E_k$. Since

$$\mathbb{E}[\phi_k^r] = p_k\mathbb{E}[E_k^r] = \frac{p_k r!}{\beta_k^r}, \qquad r = 1,2,3,4, \tag{C.3}$$

the fourth cumulant is

$$\mathrm{cum}_4(\phi_k) = \mathbb{E}\phi_k^4 - 4\mathbb{E}\phi_k^3\mathbb{E}\phi_k - 3(\mathbb{E}\phi_k^2)^2 + 12\mathbb{E}\phi_k^2(\mathbb{E}\phi_k)^2 - 6(\mathbb{E}\phi_k)^4. \tag{C.4}$$

$$\mathrm{cum}_4(\phi_k) = \frac{24p_k - 36p_k^2 + 24p_k^3 - 6p_k^4}{\beta_k^4} = \frac{6p_k}{\beta_k^4}\left(4 - 6p_k + 4p_k^2 - p_k^3\right). \tag{C.5}$$

Let $q(p) = 4 - 6p + 4p^2 - p^3$. Its derivative is $q'(p) = -(3p^2 - 8p + 6) < 0$ on $[0,1]$ because the quadratic $3p^2 - 8p + 6$ has negative discriminant and positive leading coefficient. Thus q decreases from $q(0) = 4$ to $q(1) = 1$ and is strictly positive on $[0,1]$. Hence every active coordinate has nonzero fourth cumulant, all eighth moments are finite, and $\mathbb{E}\,\|\phi\|_0 = \sum_k p_k$.

A bounded source for the probe illustration. If $Z_k \sim \mathrm{Bern}(p_k)$ and $R_k \sim \mathrm{Rad}(\pm 1)$ are independent and $\phi_k = A_k Z_k R_k$, then

$$\mathbb{E}\phi_k = 0, \qquad \mathrm{Var}(\phi_k) = A_k^2 p_k, \qquad \mathrm{cum}_4(\phi_k) = A_k^4 p_k(1 - 3p_k). \tag{C.6}$$

while $|\phi_k| \le |A_k|$ and $\mathbb{E}\,\|\phi\|_0 = \sum_k p_k$. This family is convenient when an almost-sure amplitude bound is required.

Correlated activations. For dependent centered coordinates, define $(\mathcal{K}_\phi)_{ijkl} = \mathrm{cum}(\widetilde{\phi}_i, \widetilde{\phi}_j, \widetilde{\phi}_k, \widetilde{\phi}_\ell)$. Proposition F.1 gives the exact Tucker-core form

$$\mathcal{T}_h = \mathcal{K}_\phi \times_1 W \times_2 W \times_3 W \times_4 W. \tag{C.7}$$

The diagonal CP model is the special case in which $\mathcal{K}_\phi = \sum_k \lambda_k e_k^{\otimes 4}$. This diagonalization is the linear independent-component specialization underlying classical ICA (Comon, 1994; Hyvärinen and Oja, 2000); nonlinear ICA requires additional structure or auxiliary information to restore identifiability (Hyvärinen et al., 2019; Khemakhem et al., 2020; Lyu and Fu, 2022). Proposition F.2 shows that an unrestricted core cannot identify W: invertible latent changes of coordinates preserve the observation while changing the dictionary. Theorem G.8 therefore adopts a quantitative rather than cosmetic relaxation. If $\mathcal{K}_\phi$ is diagonal plus a residual $\mathcal{E}_\phi$, the induced observed perturbation is bounded by $\|W\|_{\mathrm{op}}^4 \|\mathcal{E}_\phi\|_F$ and enters the recovery error additively. Proposition G.9 makes the distance to the diagonal model observable up to the same finite-sample cumulant radius.

D Auxiliary Geometric and Statistical Lemmas

The coherence obstruction and its equality cases are classical frame-theoretic quantities, connecting the Welch bound, Grassmannian frames, and projective packing (Welch, 1974; Strohmer and Heath Jr, 2003; Conway et al., 1996).

Lemma D.1 (Welch energy identity)**.** *For every unit-column dictionary $W = [w_1|\cdots|w_K] \in \mathbb{R}^{d\times K}$,*

$$\sum_{j<k} |\langle w_j,\, w_k\rangle|^2 = \frac{1}{2}\left(\left\|W^\top W\right\|_F^2 - K\right) \tag{D.1}$$

$$= \frac{1}{2}\left(\left\|WW^\top\right\|_F^2 - K\right) \tag{D.2}$$

$$\geq \frac{K(K-d)}{2d}. \tag{D.3}$$

Consequently,

$$\mu_W^2 \geq \frac{K-d}{d(K-1)}. \tag{D.4}$$

Equality in (D.3) *holds exactly when W is a unit-norm tight frame. Equality in* (D.4) *holds exactly when, in addition, all off-diagonal Gram entries have the same absolute value, i.e. when W is an equiangular tight frame.*

Proof. Because the Gram matrix $G = W^\top W$ has diagonal entries $G_{kk} = 1$,

$$\left\|W^\top W\right\|_F^2 = \sum_{j,k=1}^{K} |\langle w_j,\, w_k\rangle|^2 \tag{D.5}$$

$$= K + 2\sum_{j<k} |\langle w_j,\, w_k\rangle|^2, \tag{D.6}$$

which proves (D.1). Cyclicity of the trace gives

$$\left\|W^\top W\right\|_F^2 = \mathrm{tr}(W^\top W W^\top W) \tag{D.7}$$

$$= \mathrm{tr}(WW^\top WW^\top) = \left\|WW^\top\right\|_F^2, \tag{D.8}$$

which proves (D.2). Let $\lambda_1, \ldots, \lambda_d \geq 0$ be the eigenvalues of $WW^\top$. Since every column has unit norm,

$$\sum_{i=1}^{d} \lambda_i = \mathrm{tr}(WW^\top) = \|W\|_F^2 = K. \tag{D.9}$$

Cauchy–Schwarz therefore yields

$$\left\|WW^\top\right\|_F^2 = \sum_{i=1}^{d} \lambda_i^2 \geq \frac{1}{d}\left(\sum_{i=1}^{d} \lambda_i\right)^2 = \frac{K^2}{d}. \tag{D.10}$$

Substitution into (D.2) proves (D.3). Since there are $\binom{K}{2}$ unordered pairs,

$$\binom{K}{2}\mu_W^2 \geq \sum_{j<k} |\langle w_j,\, w_k\rangle|^2 \geq \frac{K(K-d)}{2d}, \tag{D.11}$$

which is (D.4). Equality in Cauchy–Schwarz holds if and only if all eigenvalues are equal, equivalently $WW^\top = (K/d)I_d$. Equality in the passage from the average squared correlation to its maximum additionally requires all off-diagonal absolute correlations to be equal. These are precisely the tightness and equiangularity conditions. □

Lemma D.2 (Spherical inner-product tail)**.** *If u, v are independent and uniform on $\mathbb{S}^{d-1}$, then, for every $0 < t < 1$,*

$$\mathbb{P}\{|\langle u,\, v\rangle| \geq t\} \leq 2(1-t^2)^{(d-1)/2} \leq 2\exp\left[-\frac{(d-1)t^2}{2}\right]. \tag{D.12}$$

Proof. Rotational invariance permits us to condition on v and take $v = e_1$. Let $g_1, \ldots, g_d$ be independent standard Gaussian variables and set $u = g/\|g\|_2$. Writing $S = \sum_{i=2}^d g_i^2$, which is independent of g_1 and has the χ^2_{d-1} distribution, we have

$$|u_1| \geq t \iff g_1^2 \geq t^2(g_1^2 + S) \tag{D.13}$$

$$\iff |g_1| \geq \sqrt{\frac{t^2}{1-t^2}}\,\sqrt{S}. \tag{D.14}$$

The elementary Gaussian tail bound $\mathbb{P}\{|g_1| \geq a\} \leq 2e^{-a^2/2}$ gives, conditionally on S,

$$\mathbb{P}\{|u_1| \geq t \mid S\} \leq 2\exp\left[-\frac{t^2}{2(1-t^2)}S\right]. \tag{D.15}$$

Taking expectations and using $\mathbb{E}e^{-qS/2} = (1+q)^{-(d-1)/2}$ for $q \geq 0$ yields

$$\mathbb{P}\{|u_1| \geq t\} \leq 2\left(1 + \frac{t^2}{1-t^2}\right)^{-(d-1)/2} \tag{D.16}$$

$$= 2(1-t^2)^{(d-1)/2} \tag{D.17}$$

$$\leq 2\exp\left[-\frac{(d-1)t^2}{2}\right], \tag{D.18}$$

where the final step uses $\log(1-x) \leq -x$ for $x \in (0,1)$. □

Lemma D.3 (Uniform mutual-coherence recovery)**.** *Let $S \subset [K]$ have cardinality $r \leq s$. If $\mu_W < (2s-1)^{-1}$, then W_S has full column rank and*

$$\max_{j \notin S} \left\| W_S^\dagger w_j \right\|_1 \leq \frac{r\mu_W}{1-(r-1)\mu_W} \leq \frac{s\mu_W}{1-(s-1)\mu_W} < 1. \tag{D.19}$$

Consequently $\mathrm{ERC}(S; W) > 0$ and every vector supported on S is the unique basis-pursuit solution of its noiseless observation.

Proof. Let $G_S = W_S^\top W_S$. Write $G_S = I_r + E_S$. Each column of E_S has at most $r-1$ nonzero entries, all of absolute value at most μ_W; hence

$$\|E_S\|_1 \leq (r-1)\mu_W \leq (s-1)\mu_W < 1. \tag{D.20}$$

The Neumann series therefore gives

$$\left\|G_S^{-1}\right\|_1 \leq \sum_{m=0}^{\infty} \|E_S\|_1^m \leq \frac{1}{1-(r-1)\mu_W}. \tag{D.21}$$

In particular, G_S is invertible and W_S has full column rank. For $j \notin S$,

$$\left\|W_S^\top w_j\right\|_1 \leq r\mu_W. \tag{D.22}$$

Since $W_S^\dagger = G_S^{-1} W_S^\top$, submultiplicativity and (D.21) imply the first inequality in (D.19). The function $r \mapsto r\mu_W/[1-(r-1)\mu_W]$ is increasing wherever its denominator is positive, which gives the second inequality. Finally,

$$\frac{s\mu_W}{1-(s-1)\mu_W} < 1 \iff (2s-1)\mu_W < 1. \tag{D.23}$$

Thus $\mathrm{ERC}(S;W) > 0$.

For completeness, let x be supported on S and let $0 \neq v \in \ker W$. Since $W_S v_S = -W_{S^c} v_{S^c}$ and W_S has full column rank,

$$v_S = -W_S^\dagger W_{S^c} v_{S^c}. \tag{D.24}$$

Therefore

$$\|v_S\|_1 \leq \sum_{j \notin S} |v_j| \left\| W_S^\dagger w_j \right\|_1 \tag{D.25}$$

$$< \sum_{j \notin S} |v_j| = \|v_{S^c}\|_1 . \tag{D.26}$$

It follows that

$$\|x+v\|_1 = \|x_S + v_S\|_1 + \|v_{S^c}\|_1 \tag{D.27}$$

$$\geq \|x\|_1 - \|v_S\|_1 + \|v_{S^c}\|_1 > \|x\|_1 . \tag{D.28}$$

Hence x is the unique minimizer of basis pursuit. □

Condition-number analyses for tensor-rank and join decompositions motivate the local-conditioning quantities used below (Vannieuwenhoven, 2017; Breiding and Vannieuwenhoven, 2018).

Lemma D.4 (Quantitative local inverse for the quartic map)**.** *Let $\theta_* = (\lambda_*, W_*)$ have nonzero weights and unit-norm columns. For every k, use centered exponential coordinates $z_k \in T_{w_k^*}\mathbb{S}^{d-1}$. Write $r_k = \|z_k\|_2$ and $\mathrm{sinc}(r) = \sin(r)/r$ for $r \neq 0$, with $\mathrm{sinc}(0) = 1$. Then*

$$w_k(z_k) = \mathrm{Exp}_{w_k^*}(z_k) = \cos(r_k) w_k^* + \mathrm{sinc}(r_k) z_k. \tag{D.29}$$

Let

$$\vartheta = (a, z_1, \ldots, z_K), \tag{D.30}$$

$$\theta(\vartheta) = \big(\lambda_* + a,\ w_1(z_1), \ldots, w_K(z_K)\big). \tag{D.31}$$

and define the global quartic tensor map and its local coordinate representation by

$$\Phi_4(\theta) = \Phi_4(\lambda, W) := \sum_{k=1}^K \lambda_k w_k^{\otimes 4}, \qquad \Psi(\vartheta) := \Phi_4(\theta(\vartheta)). \tag{D.32}$$

Choose $r_0 > 0$ small enough that this chart fixes the local signs and no weight crosses zero. Suppose

$$\gamma_* := \sigma_{\min}(D\Psi(0)) > 0, \tag{D.33}$$

$$\sup_{\|\vartheta\|_2 \leq r_0} \left\| D^2\Psi(\vartheta) \right\|_{\mathrm{op}} \leq L_* < \infty. \tag{D.34}$$

With the convention $\gamma_*/L_* = +\infty$ when $L_* = 0$, set

$$r_* = \min\left\{ r_0, \frac{\gamma_*}{L_*} \right\}. \tag{D.35}$$

Then every $\|\vartheta\|_2 \leq r_*$ satisfies

$$\|\Psi(\vartheta) - \Psi(0)\|_F \geq \gamma_* \|\vartheta\|_2 - \frac{L_*}{2}\|\vartheta\|_2^2 \tag{D.36}$$

$$\geq \frac{\gamma_*}{2}\|\vartheta\|_2 , \tag{D.37}$$

and hence

$$d_\Theta(\theta(\vartheta), \theta_*) \leq \frac{2}{\gamma_*}\|\Psi(\vartheta) - \Psi(0)\|_F . \tag{D.38}$$

Proof. Taylor's theorem with integral remainder gives

$$\Psi(\vartheta) - \Psi(0) = D\Psi(0)\vartheta + \int_0^1 (1-t)D^2\Psi(t\vartheta)[\vartheta,\vartheta]dt. \tag{D.39}$$

The second-derivative bound implies

$$\left\| \int_0^1 (1-t)D^2\Psi(t\vartheta)[\vartheta,\vartheta]dt \right\|_F \le \frac{L_*}{2} \|\vartheta\|_2^2 . \tag{D.40}$$

By the definition of the smallest singular value,

$$\|D\Psi(0)\vartheta\|_F \ge \gamma_* \|\vartheta\|_2 . \tag{D.41}$$

The reverse triangle inequality applied to (D.39) proves (D.36). If $\|\vartheta\|_2 \le \gamma_*/L_*$, then $(L_*/2)\|\vartheta\|_2^2 \le (\gamma_*/2)\|\vartheta\|_2$, which proves (D.37).

It remains to compare the chart norm with the quotient parameter metric. Using the identity permutation and positive local signs in the minimum that defines d_Θ,

$$d_\Theta(\theta(\vartheta),\theta_*)^2 \le \|a\|_2^2 + \sum_{k=1}^K \|w_k(z_k) - w_k^*\|_2^2 . \tag{D.42}$$

For the spherical exponential map,

$$\|w_k(z_k) - w_k^*\|_2 = 2\sin\left(\frac{\|z_k\|_2}{2}\right) \le \|z_k\|_2 . \tag{D.43}$$

Therefore

$$d_\Theta(\theta(\vartheta),\theta_*) \le \|\vartheta\|_2 . \tag{D.44}$$

Combining this inequality with (D.37) proves (D.38). □

E Proofs for Support-Conditioned and Worst-Case Recovery

The support-specific proposition needs only the ERC bound; the basis-pursuit implication has already been established in Lemma D.3. We therefore begin with the support-dependent Gram estimate and then pass to the worst-case corollary.

E.1 Proof of Proposition 3.1

Proof. Fix a nonempty support S with $r = |S|$ and suppose $(r-1)\mu_{\mathrm{in}}(S;W) < 1$. Write $G_S = W_S^\top W_S = I + E$. Every column of E has absolute sum at most $(r-1)\mu_{\mathrm{in}}(S;W)$, so the Neumann-series bound gives

$$\left\|G_S^{-1}\right\|_1 \le \frac{1}{1-(r-1)\mu_{\mathrm{in}}(S;W)}. \tag{E.1}$$

In particular, G_S is invertible. For every $j \notin S$,

$$\left\|W_S^\dagger w_j\right\|_1 = \left\|G_S^{-1}W_S^\top w_j\right\|_1 \tag{E.2}$$

$$\le \left\|G_S^{-1}\right\|_1 \left\|W_S^\top w_j\right\|_1 \tag{E.3}$$

$$\le \frac{r\mu_{\mathrm{out}}(S;W)}{1-(r-1)\mu_{\mathrm{in}}(S;W)}. \tag{E.4}$$

Thus $\Gamma(S;W) > 0$ implies $\max_{j\notin S}\left\|W_S^\dagger w_j\right\|_1 < 1$, equivalently $\mathrm{ERC}(S;W) > 0$. Taking probabilities over $S \sim \nu$ proves $\mathcal{R}_\nu(W) \ge \mathcal{C}_\nu(W)$.

If $|S| = r \le s$ and $\mu_W < (2s-1)^{-1}$, then both support-specific coherences are at most μ_W. Hence

$$\frac{r\mu_{\mathrm{out}}(S;W)}{1-(r-1)\mu_{\mathrm{in}}(S;W)} \le \frac{r\mu_W}{1-(r-1)\mu_W} < 1, \tag{E.5}$$

where the final inequality is equivalent to $(2r-1)\mu_W < 1$ and follows from $r \le s$. Hence every support in the support of ν has positive Γ and positive ERC, proving $\mathcal{C}_\nu(W) = \mathcal{R}_\nu(W) = 1$. □

E.2 Proof of Lemma 3.2

Proof. Assume $K \geq s \geq 2$ and $\mu_W < t_s = (2s-1)^{-1}$. First, every set of at most $2s$ columns is linearly independent. Indeed, for $T \subset [K]$ with $|T| = q \leq 2s$, the Gram matrix $G_T = W_T^\top W_T$ has diagonal entries one and, in each row, total off-diagonal absolute mass at most

$$(q-1)\mu_W \leq (2s-1)\mu_W < 1. \tag{E.6}$$

Gershgorin's theorem therefore places every eigenvalue of G_T in the positive interval $[1-(q-1)\mu_W,\, 1+(q-1)\mu_W]$. Thus G_T is positive definite and W_T has full column rank. In particular, if $K \geq 2s$ then

$$\operatorname{Spark}(W) > 2s; \tag{E.7}$$

if $K < 2s$, the same argument applied to $T = [K]$ shows that W itself has full column rank. In either case, if two vectors x, x' with at most s nonzero entries satisfied $Wx = Wx'$, then the columns indexed by $\operatorname{supp}(x) \cup \operatorname{supp}(x')$, a set of cardinality at most $\min\{2s, K\}$, would be linearly dependent. This contradicts the positive-definiteness argument above. Hence every s-sparse vector is the unique sparsest representation of its image.

Lemma D.3 additionally gives $\operatorname{ERC}(S; W) > 0$ for every $|S| \leq s$ and proves directly that every vector supported on S is the unique basis-pursuit minimizer. Finally, every unordered pair $\{j, k\}$ is contained in an s-subset, so

$$\mathcal{I}(W, s) = 1 \iff |\langle w_j,\, w_k \rangle| < t_s \quad \forall j \neq k \tag{E.8}$$

$$\iff \mu_W < t_s. \tag{E.9}$$

This proves both formulations of the lemma. □

E.3 Proof of Theorem 3.3

Proof. Write $a_s = (2s-1)^2$ and $t_s = a_s^{-1/2}$.

Welch obstruction. Lemma D.1 gives

$$\mu_W^2 \geq \frac{K-d}{d(K-1)} = \frac{\tau(d, K, s)}{a_s}. \tag{E.10}$$

Hence $\tau(d, K, s) \geq 1$ forces $\mu_W \geq t_s$, so the global coherence certificate fails. When $d < a_s$,

$$\tau(d, K, s) < 1 \iff K(a_s - d) < d(a_s - 1) \iff K < \frac{d(a_s-1)}{a_s - d}. \tag{E.11}$$

Because K is integral, the largest admissible value is $\lceil d(a_s-1)/(a_s-d) \rceil - 1$, which is precisely $K_{\mathrm{W}}(d, s)$. If $d \geq a_s$, the Welch lower bound remains below the recovery threshold for every finite K, and the theorem therefore sets $K_{\mathrm{W}}(d, s) = \infty$.

Attainment by equiangular tight frames. A real ETF attains equality in the Welch bound. Whenever it exists and $\tau(d, K, s) < 1$,

$$\mu_W^2 = \frac{K-d}{d(K-1)} = \frac{\tau(d, K, s)}{a_s} < \frac{1}{a_s} = t_s^2. \tag{E.12}$$

Lemma 3.2 then gives the uniform s-sparse recovery certificate.

Random spherical dictionaries. For independent uniform columns on $\mathbb{S}^{d-1}$, the failure event is $\{\mu_W \geq t_s\}$. Lemma D.2 and a union bound over unordered pairs yield

$$\mathbb{P}\{\mu_W \geq t_s\} \leq 2\binom{K}{2} \exp\left[-\frac{d-1}{2a_s}\right] \leq K^2 \exp\left[-\frac{d-1}{2a_s}\right]. \tag{E.13}$$

Consequently, for any $0 < \delta < 1$, the global certificate holds with probability at least $1-\delta$ whenever

$$d - 1 \geq 2a_s \left(2 \log K + \log \frac{1}{\delta}\right). \tag{E.14}$$

This is the logarithmic high-dimensional regime asserted in the main theorem. The argument concerns the distribution-free coherence endpoint only; no conclusion about a particular support law is needed for the proof. □

F Cumulant Decomposition and Identifiability Proofs

The fourth-order construction belongs to the broader tensor-decomposition approach to latent-variable and overcomplete dictionary learning (Kolda and Bader, 2009; Anandkumar et al., 2014; 2015; Schramm and Steurer, 2017).

F.1 Proof of Lemma 3.8

Proof. Let $y = h - \mathbb{E}h = W\widetilde{\phi} + \xi$. For arbitrary $u_r \in \mathbb{R}^d$, $r \in [4]$, write $c_{i_1 i_2 i_3 i_4} = \operatorname{cum}(\widetilde{\phi}_{i_1}, \widetilde{\phi}_{i_2}, \widetilde{\phi}_{i_3}, \widetilde{\phi}_{i_4})$ and $g_r = \langle u_r, \xi \rangle$. Multilinearity gives

$$\operatorname{cum}(\langle u_1, y\rangle, \dots, \langle u_4, y\rangle) = \sum_{i_1,\dots,i_4=1}^{K} c_{i_1 i_2 i_3 i_4} \prod_{r=1}^{4} \langle u_r, w_{i_r}\rangle + \operatorname{cum}(g_1, g_2, g_3, g_4). \tag{F.1}$$

Mixed terms involving both $W\widetilde{\phi}$ and ξ vanish because ξ is independent of $\widetilde{\phi}$. The final cumulant in (F.1) is zero because every cumulant of a Gaussian vector of order greater than two vanishes. Mutual independence of the coordinates $\widetilde{\phi}_k$ implies that

$$\operatorname{cum}(\widetilde{\phi}_{i_1}, \widetilde{\phi}_{i_2}, \widetilde{\phi}_{i_3}, \widetilde{\phi}_{i_4}) = 0 \tag{F.2}$$

unless $i_1 = i_2 = i_3 = i_4$. Set $U := \bigotimes_{r=1}^{4} u_r$. Therefore

$$\operatorname{cum}(\langle u_1, y\rangle, \dots, \langle u_4, y\rangle) = \sum_{k=1}^{K} \lambda_k \prod_{r=1}^{4} \langle u_r, w_k\rangle \tag{F.3}$$

$$= \left\langle \sum_{k=1}^{K} \lambda_k w_k^{\otimes 4}, U \right\rangle. \tag{F.4}$$

The left-hand side is the multilinear form associated with $\mathcal{T}_h$. Equality for every quadruple (u_1, u_2, u_3, u_4) proves

$$\mathcal{T}_h = \sum_{k=1}^{K} \lambda_k w_k^{\otimes 4}. \tag{F.5}$$

□

Proposition F.1 (Exact cumulant-core identity)**.** *Assume only that ϕ has finite fourth moments and that the Gaussian residual ξ is independent of ϕ. Let $\mathcal{K}_\phi \in \operatorname{Sym}^4(\mathbb{R}^K)$ be the fourth joint-cumulant tensor of $\widetilde{\phi} = \phi - \mathbb{E}\phi$. Then*

$$\mathcal{T}_h = \mathcal{K}_\phi \times_1 W \times_2 W \times_3 W \times_4 W. \tag{F.6}$$

If

$$\mathcal{K}_\phi = \sum_{k=1}^{K} \lambda_k e_k^{\otimes 4} + \mathcal{E}_\phi, \tag{F.7}$$

then

$$\mathcal{T}_h = \Phi_4(\theta_*) + \mathcal{B}_{\mathrm{dep}}, \qquad \mathcal{B}_{\mathrm{dep}} = \mathcal{E}_\phi \times_1 W \times_2 W \times_3 W \times_4 W, \tag{F.8}$$

and

$$\|\mathcal{B}_{\mathrm{dep}}\|_F \le \|W\|_{\mathrm{op}}^4 \, \|\mathcal{E}_\phi\|_F. \tag{F.9}$$

Proof. For arbitrary $u_1, \dots, u_4 \in \mathbb{R}^d$, the calculation in (F.1) remains valid without independence among the coordinates of $\widetilde{\phi}$. The Gaussian cumulant vanishes and the remaining multilinear form is

$$\sum_{i_1,i_2,i_3,i_4=1}^{K} \operatorname{cum}(\widetilde{\phi}_{i_1}, \widetilde{\phi}_{i_2}, \widetilde{\phi}_{i_3}, \widetilde{\phi}_{i_4}) \prod_{r=1}^{4} \langle u_r, w_{i_r}\rangle \tag{F.10}$$

$$= \langle \mathcal{K}_\phi \times_1 W \times_2 W \times_3 W \times_4 W, \, u_1 \otimes u_2 \otimes u_3 \otimes u_4 \rangle. \tag{F.11}$$

Equality for every quadruple proves (F.6). Substituting (F.7) and using multilinearity of the mode products gives (F.8). Finally, vectorization yields

$$\mathrm{vec}(\mathcal{B}_{\mathrm{dep}}) = (W \otimes W \otimes W \otimes W)\,\mathrm{vec}(\mathcal{E}_\phi).$$

The spectral norm of a Kronecker product is the product of the spectral norms of its factors, which proves (F.9). □

Proposition F.2 (Unrestricted dependent-source non-identifiability)**.** *Suppose* $\mathrm{rank}(W) \geq 2$ *and consider a latent-variable class that permits an arbitrary finite-moment distribution for* ϕ*. Then the observation law of* $h = W\phi + \xi$ *does not identify the individual unit-column directions. More precisely, there is a continuum of pairs* $(\widetilde{W}, \widetilde{\phi})$*, arbitrarily close to* (W, ϕ)*, such that* $\widetilde{W}$ *has unit-norm columns,*

$$\widetilde{W}\widetilde{\phi} = W\phi \qquad \textit{almost surely},$$

and $\widetilde{W}$ *is not related to* W *by signed permutation.*

Proof. Choose two non-collinear columns w_a, w_b, which exist because $\mathrm{rank}(W) \geq 2$. Let $Q(t) \in O(K)$ rotate the (a, b) coordinate plane and fix all other coordinates. For sufficiently small t, every column of $A(t) := WQ(t)$ is nonzero. Set

$$D(t) = \mathrm{diag}(\|A(t)_{:1}\|_2, \ldots, \|A(t)_{:K}\|_2), \tag{F.12}$$

$$\widetilde{W}(t) = A(t)D(t)^{-1}, \tag{F.13}$$

$$\widetilde{\phi}(t) = D(t)Q(t)^\top \phi. \tag{F.14}$$

The columns of $\widetilde{W}(t)$ are unit norm and

$$\widetilde{W}(t)\widetilde{\phi}(t) = WQ(t)D(t)^{-1}D(t)Q(t)^\top \phi = W\phi.$$

Finite moments are preserved by the fixed linear transformation. With a choice of orientation,

$$A(t)_{:a} = \cos(t)w_a - \sin(t)w_b.$$

The derivative at zero of its normalized version is

$$-\big(I - w_a w_a^\top\big)w_b \neq 0,$$

so $t \mapsto \widetilde{W}(t)$ is nonconstant on every sufficiently small interval. Its image is therefore uncountable, whereas the signed-permutation orbit of W is finite. Infinitely many arbitrarily small t give observationally equivalent dictionaries outside that orbit. □

F.2 Proof of Theorem 3.9

Proof. Part (i). Absorb the nonzero weights into the first mode and write

$$\mathcal{T}_h = \sum_{k=1}^{K} (\lambda_k w_k) \otimes w_k \otimes w_k \otimes w_k. \tag{F.15}$$

Thus the four factor matrices are

$$A^{(1)} = W\,\mathrm{diag}(\lambda_1, \ldots, \lambda_K), \tag{F.16}$$

$$A^{(2)} = A^{(3)} = A^{(4)} = W. \tag{F.17}$$

Multiplication of columns by nonzero scalars does not change Kruskal rank, so

$$\mathrm{krank}(A^{(m)}) = \mathrm{krank}(W), \qquad m = 1, 2, 3, 4. \tag{F.18}$$

The N-way Kruskal theorem states that an N-mode, K-term CP decomposition is of rank K and essentially unique whenever

$$\sum_{m=1}^{N} \mathrm{krank}(A^{(m)}) \geq 2K + (N - 1) \tag{F.19}$$

(Kruskal, 1977; Sidiropoulos and Bro, 2000; Bhaskara et al., 2014). With $N = 4$, the hypothesis becomes exactly

$$4\,\mathrm{krank}(W) \geq 2K + 3. \tag{F.20}$$

Hence (F.15) is an essentially unique rank-K CP decomposition.

Now suppose that a second real symmetric decomposition exists:

$$\mathcal{T}_h = \sum_{j=1}^{K} \widetilde{\lambda}_j \widetilde{w}_j^{\otimes 4}, \qquad \|\widetilde{w}_j\|_2 = 1. \tag{F.21}$$

Essential CP uniqueness matches the rank-one summands under one permutation π. In particular, for every k, the corresponding mode vectors are proportional, so $\widetilde{w}_{\pi(k)} = c_k w_k$ for some nonzero real c_k. Unit norms force $c_k \in \{\pm 1\}$. Since the tensor order is even,

$$\widetilde{w}_{\pi(k)}^{\otimes 4} = w_k^{\otimes 4}, \tag{F.22}$$

and equality of the matched rank-one terms then gives $\widetilde{\lambda}_{\pi(k)} = \lambda_k$. This proves the deterministic claim.

Part (ii). Work over $\mathbb{C}$. For each k, choose a fourth root $\alpha_k \in \mathbb{C}$ satisfying $\alpha_k^4 = \lambda_k$ and set $u_k = \alpha_k w_k$. Then

$$\mathcal{T}_h = \sum_{k=1}^{K} u_k^{\otimes 4}, \tag{F.23}$$

which is a Waring decomposition of a quartic in d variables. In the notation of Chiantini et al. (2017), the polynomial degree is 4 and the projective dimension is $d - 1$. Their subgeneric threshold is therefore

$$r_{4,d-1} = \frac{\binom{(d-1)+4}{4}}{(d-1)+1} = \frac{1}{d}\binom{d+3}{4}. \tag{F.24}$$

Their exceptional triples $(\text{degree}, \text{projective dimension}, \text{rank})$ are

$$(6, 2, 9), \qquad (4, 3, 8), \qquad (3, 5, 9). \tag{F.25}$$

For quartics only $(4, 3, 8)$ remains, which translates to $(d, K) = (4, 8)$. Thus, under the stated genericity and rank assumptions, (F.23) is the unique complex Waring decomposition, modulo permutation and multiplication of each u_k by a fourth root of unity.

Any second real weighted symmetric decomposition would, after choosing complex fourth roots of its weights, produce a second complex Waring decomposition and hence contradict complex uniqueness. Restoring the real unit-norm convention reduces the fourth-root ambiguity to the real sign change $w_k \mapsto -w_k$, while the real weights remain unchanged. Finally, the exceptional set is a proper complex algebraic subset of the parameter space. Real points are Zariski dense in that complex space, so its intersection with the real parameter space has Lebesgue measure zero. This proves the almost-everywhere real statement. □

F.3 Proof of Proposition 3.10

Proof. Orthogonal invariance gives

$$Q^\top \phi_G \sim \mathcal{N}(0, cI_r). \tag{F.26}$$

Because every column of W_G has norm one, all columns of $A_I = W_G$ are nonzero. By continuity, the same remains true for every Q in a sufficiently small neighborhood of the identity, so D_Q is invertible. The columns of $\widetilde{W}_G = A_Q D_Q^{-1}$ are unit norm by construction. Moreover,

$$\widetilde{W}_G \widetilde{\phi}_G = A_Q D_Q^{-1} D_Q Q^\top \phi_G \tag{F.27}$$

$$= W_G Q Q^\top \phi_G = W_G \phi_G, \tag{F.28}$$

which proves observational equivalence. Since $\widetilde{\phi}_G$ is a linear transform of a Gaussian vector,

$$\mathrm{Cov}(\widetilde{\phi}_G) = D_Q Q^\top (cI_r) Q D_Q = c D_Q^2. \tag{F.29}$$

This covariance matrix is diagonal; jointly Gaussian coordinates with diagonal covariance are mutually independent.

It remains to prove that the normalized dictionary genuinely changes. By $\operatorname{rank}(W_G) \geq 2$, choose two indices $a \neq b$ such that w_a and w_b are not collinear. Let $Q(t)$ be the one-parameter rotation in the (a, b) coordinate plane, with all other coordinates fixed, and choose the sign convention for which

$$(W_G Q(t))_{:a} = \cos(t) w_a - \sin(t) w_b. \tag{F.30}$$

Define the normalized column curve

$$\widetilde{w}_a(t) = \frac{\cos(t) w_a - \sin(t) w_b}{\|\cos(t) w_a - \sin(t) w_b\|_2}. \tag{F.31}$$

Differentiating the normalization map $x \mapsto x/\|x\|_2$ at the unit vector w_a gives

$$\widetilde{w}_a'(0) = -(I_d - w_a w_a^\top) w_b. \tag{F.32}$$

The right-hand side is nonzero precisely because w_b is not collinear with w_a. Hence $t \mapsto \widetilde{W}_G(t)$ is a nonconstant smooth curve through W_G.

Choose $q \in \mathbb{R}^d$ with $\langle q, \widetilde{w}_a'(0) \rangle \neq 0$. The scalar function $t \mapsto \langle q, \widetilde{w}_a(t) \rangle$ has nonzero derivative at zero and is therefore strictly monotone on a sufficiently small interval. Hence the dictionary curve has an uncountable local image. The signed-permutation orbit of W_G is finite, so uncountably many points of this image lie outside that orbit. Thus a continuum of rotations arbitrarily close to the identity produces unit-column dictionaries not equivalent to W_G by signed permutation, while leaving the observation exactly unchanged. This proves non-identifiability. □

G Finite-Sample Feature Discovery

Population uniqueness identifies the target but does not by itself provide an estimator. This section closes that gap in three stages. First, a robust estimator controls the fourth-cumulant tensor under the eighth-moment assumption. Second, an oracle-aided packing argument records the intrinsic statistical difficulty of estimating the dictionary. Third, global separation and local conditioning turn tensor perturbation into parameter recovery, while a sample-split local refinement shows how the intrinsic parameter dimension reappears after basin entry. The statements are kept separate because a statistical rate for a global minimum-distance estimator is not a computational guarantee for the practical optimizer.

G.1 Robust Cumulant-Based Feature Recovery

The fourth cumulant is translation invariant. Estimating it through an unbiased multi-sample kernel therefore avoids carrying separate errors for the mean, fourth central moment, and covariance products. Let Π_4 denote the set of partitions of $\{1, 2, 3, 4\}$. For a multi-index $\alpha = (i_1, i_2, i_3, i_4)$ and four independent observations $X^{(1)}, \ldots, X^{(4)} \overset{\text{iid}}{\sim} h$, define

$$H_\alpha(X^{(1:4)}) = \sum_{\pi \in \Pi_4} (|\pi| - 1)!(-1)^{|\pi|-1} \prod_{B \in \pi} \prod_{r \in B} X_{i_r}^{(\ell_\pi(B))}, \tag{G.1}$$

where ℓ_π assigns distinct sample labels to the blocks of π. Independence across blocks and the moment–cumulant formula give

$$\mathbb{E} H_\alpha(X^{(1:4)}) = \operatorname{cum}(h_{i_1}, h_{i_2}, h_{i_3}, h_{i_4}). \tag{G.2}$$

Partition the observations into independent quartets, apply scalar median-of-means to each of the D_4 distinct symmetric coordinates, and assemble the resulting tensor $\widehat{\mathcal{T}}$. This coordinatewise construction is deliberately simple and gives a global Frobenius-norm bound whose dimension dependence is visible. The factor $D_4 = \dim \operatorname{Sym}^4(\mathbb{R}^d)$ is therefore a property of this full-tensor estimator and union-bound proof, not an information-theoretic claim that feature recovery intrinsically requires d^4 parameters. After a pilot estimate has entered an identifiable chart, the sample-split tangent-score refinement below estimates only the $p = K + K(d-1) = Kd$ local score coordinates and need not assemble the full fourth-order tensor.

The statistical estimator analyzed below is the global minimum-distance fit on the compact class Θ_B from (7):

$$\widehat{\theta}_{\mathrm{glob}} \in \operatorname*{arg\,min}_{\theta\in\Theta_B} \left\| \widehat{\mathcal{T}} - \Phi_4(\theta) \right\|_F . \tag{G.3}$$

Compactness guarantees existence. Theorem G.6 proves global basin capture and the $n^{-1/2}$ rate for (G.3); Corollary G.7 replaces the continuum optimization by a finite exhaustive search.

The implementation used in the experiments is kept separate from the estimator analyzed above. It uses data-driven initialization followed by projected local refinement and is summarized once, in section N, to avoid duplicating algorithmic claims inside the statistical proof. The finite-sample theorem concerns the global minimum-distance estimator; the practical routine is evaluated empirically, and its restart behavior is not promoted to a global computational guarantee.

G.2 Information Lower Bound

The preceding construction gives an upper route once the cumulant tensor can be estimated and the correct basin can be isolated. The next result asks a different question: even if the latent activations are revealed to the estimator, how many observations are needed to distinguish nearby dictionaries modulo signed permutation? The oracle only makes the problem easier, so a lower bound in this experiment also applies to the original observation model.

Lemma G.1 (Local packing modulo signs and permutations)**.** *Fix $d \geq 2$ and a unit-column dictionary $W^0 = [w_1^0|\cdots|w_K^0]$ with no two columns equal up to sign. There exist $\epsilon_0 = \epsilon_0(W^0) > 0$ and universal constants $c_1, c_2 > 0$ such that, for every $0 < \epsilon \leq \epsilon_0$, there are unit-column dictionaries $W^1, \ldots, W^M$ satisfying*

$$\log M \geq c_1 K(d-1), \tag{G.4}$$

$$d_{\pm,\mathrm{perm}}\big(W^a, W^b\big) \geq 4\epsilon \qquad (a \neq b), \tag{G.5}$$

$$\left\|W^a - W^b\right\|_F^2 \leq c_2 K \epsilon^2 \qquad (a, b \in [M]). \tag{G.6}$$

The entire packing may be chosen inside any prescribed open neighborhood of W^0 in the product manifold $(\mathbb{S}^{d-1})^K$.

Proof. Let $\mathfrak{G}$ be the finite group of signed permutations of the columns. Define the max-column distance

$$d_\infty(U, V) := \max_{1\leq k\leq K} \|u_k - v_k\|_2 . \tag{G.7}$$

By definition, $d_{\pm,\mathrm{perm}}(U, V) = \min_{g\in\mathfrak{G}} d_\infty(U, g\cdot V)$. Because no two columns of W^0 agree up to sign, the action of $\mathfrak{G}$ at W^0 is free. Hence the finite orbit has positive separation

$$\Delta_0 := \min_{g\in\mathfrak{G}\setminus\{\mathrm{id}\}} d_\infty(W^0, g\cdot W^0) > 0. \tag{G.8}$$

For each k, choose an orthonormal matrix $U_k \in \mathbb{R}^{d\times(d-1)}$ spanning the tangent space $T_{w_k^0}\mathbb{S}^{d-1}$ and use the local normalization chart

$$F_k(z) := \frac{w_k^0 + U_k z}{\sqrt{1 + \|z\|_2^2}}, \qquad z \in \mathbb{R}^{d-1}. \tag{G.9}$$

There is a universal radius $r_{\mathrm{ch}} > 0$ such that, whenever $\|z\|_2 , \|z'\|_2 \leq r_{\mathrm{ch}}$,

$$\frac{1}{2}\|z - z'\|_2 \leq \|F_k(z) - F_k(z')\|_2 \leq 2\|z - z'\|_2 . \tag{G.10}$$

To verify this, note first that F_k is smooth and $DF_k(0) = U_k$, an isometry. By continuity of DF_k, after choosing a universal small radius one has $\|DF_k(z)\|_{\mathrm{op}} \leq 2$, which gives the upper bound by integration along the line segment from z to z'. On the image of that ball the inverse chart is

$$F_k^{-1}(u) = \frac{U_k^\top u}{\langle w_k^0,\, u\rangle}. \tag{G.11}$$

Its derivative is uniformly bounded by 2 after shrinking r_{ch} once more, because $\langle w_k^0,\, F_k(z)\rangle = (1+\|z\|_2^2)^{-1/2}$ stays bounded away from zero. Applying the upper Lipschitz bound to the inverse chart gives the lower inequality in (G.10).

Let $m = d-1$. A maximal 8ϵ-separated subset of the Euclidean ball $B_m(0,16\epsilon)$ is an 8ϵ-net of that ball. Comparing volumes of the covering balls shows that its cardinality is at least

$$\frac{\operatorname{vol} B_m(0,16\epsilon)}{\operatorname{vol} B_m(0,8\epsilon)} = 2^m. \tag{G.12}$$

Choose such a packing $\mathcal{P}_k$ independently in every tangent space and map it to the sphere through F_k. If $16\epsilon \le r_{\mathrm{ch}}$, (G.10) shows that distinct points in $F_k(\mathcal{P}_k)$ are separated by at least 4ϵ. Taking the Cartesian product over the K columns gives

$$M \ge 2^{K(d-1)}, \qquad \log M \ge (\log 2)K(d-1), \tag{G.13}$$

so (G.4) holds with $c_1 = \log 2$.

Every packed column lies within 32ϵ of its center by the upper bound in (G.10). Hence, for all a, b,

$$\left\|W^a - W^b\right\|_F^2 \le \sum_{k=1}^{K} (64\epsilon)^2 = 4096K\epsilon^2, \tag{G.14}$$

which proves (G.6) with $c_2 = 4096$.

For identity alignment, two distinct product elements differ in at least one column and therefore have max-column distance at least 4ϵ. For a nonidentity $g \in \mathfrak{G}$, invariance of d_∞ under the group action and the triangle inequality give

$$d_\infty(W^a, g\cdot W^b) \ge d_\infty(W^0, g\cdot W^0) - d_\infty(W^a, W^0) - d_\infty(g\cdot W^b, g\cdot W^0) \ge \Delta_0 - 64\epsilon. \tag{G.15}$$

Choose $\epsilon_0 \le \Delta_0/128$. Then the last display is at least $\Delta_0/2 \ge 4\epsilon$. Thus every signed-permutation alignment has max-column error at least 4ϵ, proving (G.5). Finally, because the construction radius tends to zero with ϵ_0, the packing lies inside any prescribed open neighborhood of W^0 after decreasing ϵ_0 if necessary. □

Theorem G.2 (Oracle-aided minimax lower bound)**.** *Fix integers $d \ge 2$ and $1 \le s \le K/4$, an amplitude $A > 0$, and a noise level $\sigma > 0$. Let W^0 be a unit-column dictionary with no signed duplicate columns and suppose that W^0 belongs to an open set $\mathcal{U}$ on which the fourth-cumulant decomposition is identifiable. Such a center exists, for example, throughout the overcomplete range $d < K \le 2d-2$: a generic real unit-column dictionary has $\operatorname{krank}(W^0) = d$, and the deterministic Kruskal condition remains valid on a sufficiently small neighborhood.*

There exist $\epsilon_0 = \epsilon_0(W^0, \mathcal{U}) > 0$ and universal constants $c, c_0 > 0$ such that, for every $0 < \epsilon \le \epsilon_0$, one can choose a finite family $\mathcal{W}_\epsilon \subset \mathcal{U}$ for which every estimator based on n observations satisfies

$$n \le c\frac{\sigma^2 K(d-1)}{A^2 s\epsilon^2} \quad \Longrightarrow \quad \sup_{W\in\mathcal{W}_\epsilon} \mathbb{P}_W\left\{ d_{\pm,\mathrm{perm}}\left(\widehat{W}, W\right) > \epsilon \right\} \ge c_0. \tag{G.16}$$

The same lower bound holds in the strictly easier oracle experiment in which the estimator is additionally given the latent activation vector for every observation.

Proof. Let

$$\phi_k = AZ_kR_k, \qquad Z_k \sim \mathrm{Bern}(p), \tag{G.17}$$

$$R_k \sim \mathrm{Rad}(\pm 1), \qquad p = \frac{s}{K}. \tag{G.18}$$

with all coordinates independent. Since $p \le 1/4$,

$$\mathrm{cum}_4(\phi_k) = A^4 p(1-3p) > 0, \tag{G.19}$$

so the source belongs to the non-Gaussian identifiable submodel on $\mathcal{U}$. Apply Lemma G.1 inside $\mathcal{U}$ and write $\mathcal{W}_\epsilon = \{W^1, \dots, W^M\}$.

Let Q_a denote the law of one oracle observation (ϕ, h) when the dictionary is W^a. The marginal law of ϕ is independent of a, and, conditionally on ϕ,

$$h \mid \phi \sim \mathcal{N}(W^a\phi, \sigma^2 I_d). \tag{G.20}$$

For compactness, write $\mathcal{N}_a(\phi) = \mathcal{N}(W^a\phi, \sigma^2 I_d)$. The chain rule for relative entropy and the Gaussian KL formula then give, for every a, b,

$$D_{\mathrm{KL}}(Q_a^{\otimes n} \| Q_b^{\otimes n}) = n\mathbb{E}_\phi D_{\mathrm{KL}}(\mathcal{N}_a(\phi) \| \mathcal{N}_b(\phi)) = \frac{n}{2\sigma^2}\mathbb{E}\left\|(W^a - W^b)\phi\right\|_2^2. \tag{G.21}$$

The coordinates are centered and independent, and $\mathbb{E}\phi_k^2 = A^2 p$. Hence

$$\mathbb{E}(\phi\phi^\top) = A^2 p I_K = \frac{A^2 s}{K} I_K. \tag{G.22}$$

Substituting this covariance into (G.21) gives

$$D_{\mathrm{KL}}(Q_a^{\otimes n} \| Q_b^{\otimes n}) = \frac{nA^2 s}{2\sigma^2 K}\left\|W^a - W^b\right\|_F^2 \tag{G.23}$$

$$\leq \frac{c_2 n A^2 s \epsilon^2}{2\sigma^2}. \tag{G.24}$$

Choose an index V uniformly from $[M]$ and draw the oracle sample from $Q_V^{\otimes n}$. For any test $\widehat{V}$, Fano's inequality gives

$$\mathbb{P}\{\widehat{V} \neq V\} \geq 1 - \frac{I(V; \text{data}) + \log 2}{\log M}. \tag{G.25}$$

The mutual information is bounded by the largest pairwise KL divergence, so (G.24) and (G.4) imply

$$\frac{I(V; \text{data})}{\log M} \leq \frac{c_2 n A^2 s \epsilon^2}{2c_1 \sigma^2 K(d-1)}. \tag{G.26}$$

Choose the universal constant c in (G.16) small enough that the right-hand side is at most $1/8$. Since $K \geq 4$ and $d \geq 2$, the construction gives $\log M \geq 4 \log 2$, so $(\log 2)/(\log M) \leq 1/4$. Equation (G.25) then yields

$$\inf_{\widehat{V}} \sup_{a \in [M]} Q_a^{\otimes n}\{\widehat{V} \neq a\} \geq \frac{1}{2}. \tag{G.27}$$

Given any dictionary estimator $\widehat{W}$, define the nearest-packing point test

$$\widehat{V} \in \arg\min_{a \in [M]} d_{\pm,\mathrm{perm}}\left(\widehat{W}, W^a\right). \tag{G.28}$$

Because the packing separation is at least 4ϵ, the event $d_{\pm,\mathrm{perm}}\left(\widehat{W}, W^a\right) \leq \epsilon$ forces $\widehat{V} = a$. Indeed, the triangle inequality gives, for every $b \neq a$,

$$d_{\pm,\mathrm{perm}}\left(\widehat{W}, W^b\right) \geq 3\epsilon. \tag{G.29}$$

Consequently,

$$Q_a^{\otimes n}\{\widehat{V} \neq a\} \leq Q_a^{\otimes n}\{d_{\pm,\mathrm{perm}}\left(\widehat{W}, W^a\right) > \epsilon\}. \tag{G.30}$$

Thus the oracle minimax estimation error is bounded below by a positive universal constant; one may take $c_0 = 1/2$. An estimator that observes only h is a special case of an oracle estimator that ignores ϕ, so the same lower bound holds without the oracle information. □

G.3 Global-to-Local Upper Bounds

The lower bound fixes the scale of the estimation problem; we now return to an achievable route. The proof separates tensor estimation, basin entry, and local inversion. That separation is essential: concentration controls the observed tensor, compactness and population uniqueness create a positive gap to inequivalent basins, and the local Jacobian controls the conversion from tensor error to parameter error.

The robust aggregation steps below follow the median-of-means and robust U-statistic tradition for heavy-tailed estimation (Minsker, 2015; Joly and Lugosi, 2016; Lugosi and Mendelson, 2019).

Lemma G.3 (Scalar median-of-means)**.** *Let $Y_1, \dots, Y_N$ be independent with common mean μ and variance at most v. For every $0 < \delta < 1$, if $N \geq C \log(2/\delta)$, there exists a median-of-means estimator $\widehat{\mu}_{\mathrm{MOM}}$ satisfying*

$$\mathbb{P}\left\{ |\widehat{\mu}_{\mathrm{MOM}} - \mu| > C\sqrt{\frac{v \log(2/\delta)}{N}} \right\} \leq \delta, \tag{G.31}$$

where $C > 0$ is universal.

Proof. Choose an odd number of blocks $b \asymp \log(2/\delta)$ and let each block contain $m = \lfloor N/b \rfloor$ observations; discard at most $b - 1$ observations. For the mean $\overline{Y}_j$ of one block,

$$\mathrm{Var}(\overline{Y}_j) \leq \frac{v}{m}. \tag{G.32}$$

Chebyshev's inequality therefore gives

$$\mathbb{P}\Big\{ |\overline{Y}_j - \mu| > \sqrt{8v/m} \Big\} \leq \frac{1}{8}. \tag{G.33}$$

The block means are independent. If their median lies outside the same interval, at least half of the b blocks must be bad. Hoeffding's inequality for the corresponding Bernoulli indicators gives

$$\mathbb{P}\{ \#\{j : \text{block } j \text{ is bad}\} \geq b/2 \} \leq e^{-cb}. \tag{G.34}$$

Taking $b \geq C_0 \log(2/\delta)$ makes this probability at most δ. Finally, $m \geq N/(2b)$ when $N \geq 2b$, so $\sqrt{8v/m} \leq C\sqrt{vb/N}$, which proves (G.31). □

Lemma G.4 (Robust cumulant concentration)**.** *Under Assumption 3.5, if $n \geq C \log(2D_4/\delta)$, the coordinatewise median-of-means estimator constructed from independent quartets satisfies, with probability at least $1 - \delta$,*

$$\left\| \widehat{\mathcal{T}} - \mathcal{T}_h \right\|_F \leq C M_8^4 \sqrt{\frac{D_4 \log(2D_4/\delta)}{n}}. \tag{G.35}$$

Proof. Let $N = \lfloor n/4 \rfloor$ and form N disjoint quartets. For a fixed symmetric multi-index α, the corresponding quartet kernels are independent and have mean $T_\alpha = (\mathcal{T}_h)_\alpha$ by the moment–cumulant formula.

Fix a partition $\pi \in \Pi_4$ with blocks $B_1, \dots, B_q$. Independence of the observations assigned to distinct blocks gives

$$\mathbb{E}\left| \prod_{B \in \pi} \prod_{r \in B} X_{i_r}^{(\ell_\pi(B))} \right|^2 \tag{G.36}$$

$$= \prod_{j=1}^{q} \mathbb{E}\left[\prod_{r \in B_j} \left| X_{i_r}^{(j)} \right|^2 \right]. \tag{G.37}$$

For a block of size $|B_j|$, the product inside the expectation is at most $\left\| X^{(j)} \right\|_2^{2|B_j|}$. Lyapunov's inequality and the definition of M_8 give

$$\mathbb{E}\, \|X\|_2^{2|B_j|} \leq M_8^{2|B_j|}. \tag{G.38}$$

Since $\sum_j |B_j| = 4$, the second moment of every partition monomial is at most M_8^8. The set Π_4 has fifteen elements and its coefficients $(|\pi| - 1)!$ are bounded by an absolute constant. By Cauchy–Schwarz,

$$\mathrm{Var}(H_\alpha) \leq \mathbb{E} H_\alpha^2 \leq C_0 M_8^8. \tag{G.39}$$

Apply Lemma G.3 to the N quartet kernels with failure probability δ/D_4. For every fixed α,

$$\mathbb{P}\left\{ \left| \widehat{T}_\alpha - T_\alpha \right| > C M_8^4 \sqrt{\frac{\log(2D_4/\delta)}{N}} \right\} \leq \frac{\delta}{D_4}. \tag{G.40}$$

A union bound controls all D_4 distinct symmetric coordinates simultaneously. When these coordinates are expanded into the full ordered tensor, each distinct coordinate appears with multiplicity at most 4!. Hence, on the same event,

$$\left\|\widehat{\mathcal{T}} - \mathcal{T}_h\right\|_F^2 \le 4!\, D_4 \max_\alpha \left|\widehat{T}_\alpha - T_\alpha\right|^2 \tag{G.41}$$

$$\le C M_8^8 \frac{D_4 \log(2D_4/\delta)}{N} \tag{G.42}$$

$$\le C M_8^8 \frac{D_4 \log(2D_4/\delta)}{n}, \tag{G.43}$$

where the last constant absorbs the factor between $N = \lfloor n/4 \rfloor$ and n. Taking square roots proves (G.35). □

Lemma G.5 (Local stability of a competitive fit)**.** *Let θ_* satisfy the regularity assumptions of Lemma D.4. Suppose that $\widehat{\theta}$ belongs, after one signed permutation, to the local chart ball $\{\theta(\vartheta) : \|\vartheta\|_2 \le r_*\}$ and satisfies*

$$\left\|\widehat{\mathcal{T}} - \Phi_4(\widehat{\theta})\right\|_F \le \left\|\widehat{\mathcal{T}} - \Phi_4(\theta_*)\right\|_F. \tag{G.44}$$

If $\mathcal{T}_h = \Phi_4(\theta_*)$, then deterministically

$$d_\Theta(\widehat{\theta}, \theta_*) \le \frac{4}{\gamma_*} \left\|\widehat{\mathcal{T}} - \mathcal{T}_h\right\|_F. \tag{G.45}$$

Proof. Because $\Phi_4(\theta_*) = \mathcal{T}_h$, the objective comparison gives

$$\left\|\widehat{\mathcal{T}} - \Phi_4(\widehat{\theta})\right\|_F \le \left\|\widehat{\mathcal{T}} - \mathcal{T}_h\right\|_F.$$

The triangle inequality therefore yields

$$\left\|\Phi_4(\widehat{\theta}) - \Phi_4(\theta_*)\right\|_F \le \left\|\Phi_4(\widehat{\theta}) - \widehat{\mathcal{T}}\right\|_F + \left\|\widehat{\mathcal{T}} - \mathcal{T}_h\right\|_F \tag{G.46}$$

$$\le 2\left\|\widehat{\mathcal{T}} - \mathcal{T}_h\right\|_F. \tag{G.47}$$

After the stated alignment, Lemma D.4 applies and gives

$$d_\Theta(\widehat{\theta}, \theta_*) \le \frac{2}{\gamma_*} \left\|\Phi_4(\widehat{\theta}) - \Phi_4(\theta_*)\right\|_F \le \frac{4}{\gamma_*} \left\|\widehat{\mathcal{T}} - \mathcal{T}_h\right\|_F.$$

□

Sample-split tangent-score refinement after basin entry. Lemma G.5 is deliberately optimizer agnostic, but its comparison with the unknown truth is not an implementable stopping rule. We now record a constructive post-entry refinement that uses the intrinsic tangent dimension rather than all D_4 tensor coordinates. The statement is left unnumbered so that all theorem, equation, and cross-document indices remain unchanged.

Let a pilot parameter $\widetilde{\theta}$ be computed from one subsample and held fixed while an independent refinement subsample of size n_{ref} is processed. Use an aligned exponential chart $u \mapsto \widetilde{\theta}(u)$ around the pilot, write

$$p = K + K(d-1) = Kd, \qquad \widetilde{\Psi}(u) = \Phi_4(\widetilde{\theta}(u)), \qquad J = D\widetilde{\Psi}(0),$$

and suppose J has full column rank. Set

$$G = J^*J, \qquad Q = JG^{-1/2}, \qquad \widetilde{\gamma} = \sigma_{\min}(J).$$

The columns $Q_1, \dots, Q_p$ of the polar factor Q are Frobenius-orthonormal tensors. For a tensor A, write $Q^*A = (\langle Q_1, A\rangle, \dots, \langle Q_p, A\rangle)^\top$. Writing $H(X^{(1:4)})$ for the symmetric tensor whose coordinates are the kernels in (G.1), estimate each scalar

$$t_q = \langle \mathcal{T}_h, Q_q \rangle$$

by median-of-means applied to $\langle H(X^{(1:4)}), Q_q \rangle$, and collect the estimates in $\widehat{t}$. Define the one-step tangent estimator

$$\widehat{u} = G^{-1/2}\left(\widehat{t} - Q^*\Phi_4(\widetilde{\theta})\right).$$

The contractions need not be implemented by storing the dense tensors Q_q. Since $Q^* = G^{-1/2}J^*$, they can be obtained by first computing the structured tangent contractions J^*H and then applying the $p \times p$ preconditioner $G^{-1/2}$.

Assume that, after one signed permutation, the population parameter has coordinates $\theta_* = \widetilde{\theta}(u_*)$ in this chart and that

$$\sup_{\|u\|_2 \le \widetilde{r}} \left\| D^2\widetilde{\Psi}(u) \right\|_{\mathrm{op}} \le \widetilde{L}, \qquad \|u_*\|_2 \le \widetilde{r}.$$

Then, for a universal constant C, with probability at least $1-\delta$ over the independent refinement sample,

$$\|\widehat{u} - u_*\|_2 \le \frac{\widetilde{L}}{2\widetilde{\gamma}} \|u_*\|_2^2 + \frac{CM_8^4}{\widetilde{\gamma}} \sqrt{\frac{p\log(2p/\delta)}{n_{\mathrm{ref}}}}.$$

If the right-hand side keeps $\widehat{u}$ inside the chosen chart, the same bound, multiplied by the finite Lipschitz constant of the chart, controls $d_\Theta(\widetilde{\theta}(\widehat{u}), \theta_*)$. Thus, conditional on basin entry, the stochastic term has the intrinsic scaling

$$\frac{M_8^4}{\widetilde{\gamma}} \sqrt{\frac{Kd\log(Kd/\delta)}{n_{\mathrm{ref}}}},$$

contains neither D_4 nor the global separation margin Δ_*, and is computable in polynomial time. The quantity $\widetilde{\gamma}$ is obtained directly from the Jacobian at the pilot; M_8 enters the confidence statement but is not an input required to run the refinement.

Proof. For every Frobenius-unit tensor A, the contraction $\langle H, A\rangle$ has variance at most $C_0 M_8^8$. Indeed, each of the fifteen partition terms in (G.1) is a tensor product of independent sample vectors, and Cauchy–Schwarz bounds its contraction by the product of their Euclidean norms. Squaring, factorizing across independent sample labels, and applying Lyapunov's inequality gives total eighth moment at most M_8^8; the finite partition coefficients are absorbed into C_0. Scalar median-of-means followed by a union bound over the p orthonormal contractions therefore yields

$$\left\|\widehat{t} - Q^*\mathcal{T}_h\right\|_2 \le CM_8^4 \sqrt{\frac{p\log(2p/\delta)}{n_{\mathrm{ref}}}}.$$

Taylor expansion in the pilot chart gives

$$\mathcal{T}_h - \Phi_4(\widetilde{\theta}) = Ju_* + R, \qquad \|R\|_F \le \frac{\widetilde{L}}{2}\|u_*\|_2^2.$$

Using $Q^*J = G^{1/2}$ in the definition of $\widehat{u}$ gives

$$\widehat{u} - u_* = G^{-1/2}\left[Q^*R + \widehat{t} - Q^*\mathcal{T}_h\right].$$

The displayed bound follows from $\left\|G^{-1/2}\right\|_{\mathrm{op}} = \widetilde{\gamma}^{-1}$ and the preceding two inequalities. Smoothness of the exponential chart on its compact local ball supplies the final parameter-metric comparison. □

Theorem G.6 (End-to-end minimum-distance recovery)**.** *Fix $0 < \lambda_{\min} < \lambda_{\max} < \infty$ and let Θ_B be the compact parameter class in* (7)*. Suppose that $\theta_* = (\lambda_*, W_*)$ is an interior point of Θ_B, is identifiable in Θ_B modulo signed permutations, and satisfies Lemma D.4 with constants (r_*, γ_*).*

Let $\mathfrak{G}$ denote the finite group of simultaneous permutations of (λ, W) and independent sign changes of the columns of W. Define the aligned local basin

$$\mathcal{U}_* = \Theta_B \cap \bigcup_{g\in\mathfrak{G}} \{g\cdot\theta(\vartheta) : \|\vartheta\|_2 < r_*\} \tag{G.48}$$

and the global separation margin

$$\Delta_* = \inf_{\theta \in \Theta_B \setminus \mathcal{U}_*} \|\Phi_4(\theta) - \Phi_4(\theta_*)\|_F . \tag{G.49}$$

Then $\Delta_* > 0$.

Let $\widehat{\mathcal{T}}$ *be the robust cumulant estimator of Lemma G.4, and let* $\widehat{\theta}_{\mathrm{glob}}$ *be any global minimizer in* (G.3). *Set*

$$\varepsilon_n(\delta) = C M_8^4 \sqrt{\frac{D_4 \log(2D_4/\delta)}{n}}. \tag{G.50}$$

If

$$2\varepsilon_n(\delta) < \Delta_*, \tag{G.51}$$

then, with probability at least $1 - \delta$,

$$d_\Theta(\widehat{\theta}_{\mathrm{glob}}, \theta_*) \le \frac{4\varepsilon_n(\delta)}{\gamma_*}. \tag{G.52}$$

Consequently, $d_\Theta(\widehat{\theta}_{\mathrm{glob}}, \theta_*) \le \epsilon$ *whenever*

$$n \ge C \frac{M_8^8 D_4}{\gamma_*^2 \epsilon^2} \log\left(\frac{2D_4}{\delta}\right) \tag{G.53}$$

and, simultaneously, the right-hand side of (G.50) *is smaller than* $\Delta_*/2$.

Proof. The class Θ_B is a closed subset of $[-\lambda_{\max}, \lambda_{\max}]^K \times (\mathbb{S}^{d-1})^K$ and is therefore compact. The aligned chart union $\mathcal{U}_*$ is relatively open in Θ_B, so its complement is compact. The continuous function

$$\theta \longmapsto \|\Phi_4(\theta) - \Phi_4(\theta_*)\|_F$$

attains its infimum on that complement. If the minimum were zero, there would be $\bar{\theta} \in \Theta_B \setminus \mathcal{U}_*$ satisfying $\Phi_4(\bar{\theta}) = \Phi_4(\theta_*)$. Identifiability in Θ_B would imply $\bar{\theta} = g \cdot \theta_*$ for some $g \in \mathfrak{G}$, but every such orbit point belongs to $\mathcal{U}_*$. This contradiction proves $\Delta_* > 0$.

Work on the event

$$\left\|\widehat{\mathcal{T}} - \mathcal{T}_h\right\|_F \le \varepsilon_n(\delta), \tag{G.54}$$

which has probability at least $1 - \delta$ by Lemma G.4. Since $\mathcal{T}_h = \Phi_4(\theta_*)$ and $\widehat{\theta}_{\mathrm{glob}}$ minimizes the same objective over a class containing θ_*,

$$\left\|\widehat{\mathcal{T}} - \Phi_4(\widehat{\theta}_{\mathrm{glob}})\right\|_F \le \left\|\widehat{\mathcal{T}} - \Phi_4(\theta_*)\right\|_F \le \varepsilon_n(\delta). \tag{G.55}$$

Hence

$$\left\|\Phi_4(\widehat{\theta}_{\mathrm{glob}}) - \Phi_4(\theta_*)\right\|_F \le \left\|\Phi_4(\widehat{\theta}_{\mathrm{glob}}) - \widehat{\mathcal{T}}\right\|_F + \left\|\widehat{\mathcal{T}} - \Phi_4(\theta_*)\right\|_F \le 2\varepsilon_n(\delta). \tag{G.56}$$

Under (G.51), this distance is strictly below Δ_*. By the definition of (G.49), $\widehat{\theta}_{\mathrm{glob}}$ must belong to $\mathcal{U}_*$. Align it by the corresponding element of $\mathfrak{G}$ and apply Lemma G.5; this gives (G.52). Solving $4\varepsilon_n(\delta)/\gamma_* \le \epsilon$ yields (G.53), while (G.51) is the separate sample threshold needed for basin capture. □

Corollary G.7 (Finite-net implementation)**.** *For* $0 < \rho \le 1$, *let* $\mathcal{N}_\rho \subset \Theta_B$ *be a* ρ*-net under* d_Θ, *and define*

$$\widehat{\theta}_\rho \in \operatorname*{arg\,min}_{\theta \in \mathcal{N}_\rho} \left\|\widehat{\mathcal{T}} - \Phi_4(\theta)\right\|_F . \tag{G.57}$$

Set

$$L_B = \sqrt{K}\,(1 + 4\lambda_{\max}). \tag{G.58}$$

If

$$2\varepsilon_n(\delta) + L_B \rho < \Delta_*, \tag{G.59}$$

then, with probability at least $1 - \delta$,

$$d_\Theta(\widehat{\theta}_\rho, \theta_*) \le \frac{2}{\gamma_*}\left(2\varepsilon_n(\delta) + L_B \rho\right). \tag{G.60}$$

Moreover, there is a universal constant $C > 0$ for which one may choose

$$|\mathcal{N}_\rho| \le \left[\frac{C\sqrt{K}(1+\lambda_{\max})}{\rho}\right]^{K(d+1)}. \tag{G.61}$$

Thus (G.57) *is a finite end-to-end algorithm, although its worst-case running time is exponential in the parameter dimension.*

Proof. First we prove the Lipschitz bound. Align two parameters $\theta = (\lambda, W)$ and $\theta' = (\lambda', W')$ by a permutation and signs attaining $d_\Theta(\theta, \theta')$. For unit vectors u, v, telescoping one tensor factor at a time gives

$$u^{\otimes 4} - v^{\otimes 4} = \sum_{r=1}^{4} v^{\otimes(r-1)} \otimes (u - v) \otimes u^{\otimes(4-r)}. \tag{G.62}$$

and therefore

$$\left\|u^{\otimes 4} - v^{\otimes 4}\right\|_F \le 4 \left\|u - v\right\|_2. \tag{G.63}$$

Consequently,

$$\|\Phi_4(\theta) - \Phi_4(\theta')\|_F \le \sum_{k=1}^{K} |\lambda_k - \lambda'_k| + 4\lambda_{\max} \sum_{k=1}^{K} \|w_k - w'_k\|_2 \le \sqrt{K}(1 + 4\lambda_{\max}) d_\Theta(\theta, \theta'). \tag{G.64}$$

which is (G.58).

Choose $\theta_\rho \in \mathcal{N}_\rho$ with $d_\Theta(\theta_\rho, \theta_*) \le \rho$. On the event (G.54),

$$\left\|\widehat{\mathcal{T}} - \Phi_4(\widehat{\theta}_\rho)\right\|_F \le \left\|\widehat{\mathcal{T}} - \Phi_4(\theta_\rho)\right\|_F \tag{G.65}$$

$$\le \varepsilon_n(\delta) + L_B \rho. \tag{G.66}$$

Therefore

$$\left\|\Phi_4(\widehat{\theta}_\rho) - \Phi_4(\theta_*)\right\|_F \le \left\|\Phi_4(\widehat{\theta}_\rho) - \widehat{\mathcal{T}}\right\|_F + \left\|\widehat{\mathcal{T}} - \Phi_4(\theta_*)\right\|_F \le 2\varepsilon_n(\delta) + L_B \rho. \tag{G.67}$$

Condition (G.59) puts the net estimator in $\mathcal{U}_*$. The local inverse bound (D.38) then gives (G.60).

For completeness, cover the weight cube $[-\lambda_{\max}, \lambda_{\max}]^K$ in Euclidean norm at radius $\rho/2$ by at most

$$\left(1 + \frac{4\lambda_{\max}\sqrt{K}}{\rho}\right)^K$$

points. Cover each $\mathbb{S}^{d-1}$ at Euclidean radius $\rho/(2\sqrt{K})$ by at most $(1 + 4\sqrt{K}/\rho)^d$ points and take the K-fold product. The resulting product net has radius at most ρ in the unquotiented product metric, hence also in d_Θ, and its cardinality is bounded by (G.61) after enlarging the universal constant. □

Theorem G.8 (Recovery under approximately diagonal fourth cumulants)**.** *Assume the setting of Proposition F.1, but do not assume coordinate independence. Suppose the diagonal part $\theta_* = (\lambda, W)$ in* (F.7) *belongs to Θ_B, is identifiable and regular as in Theorem G.6, and define*

$$\beta_{\mathrm{dep}} := \|\mathcal{B}_{\mathrm{dep}}\|_F. \tag{G.68}$$

Let $\widehat{\theta}_{\mathrm{glob}}$ minimize (G.3)*. If*

$$2\left(\varepsilon_n(\delta) + \beta_{\mathrm{dep}}\right) < \Delta_*, \tag{G.69}$$

then, with probability at least $1 - \delta$,

$$d_\Theta(\widehat{\theta}_{\mathrm{glob}}, \theta_*) \le \frac{4}{\gamma_*}\left(\varepsilon_n(\delta) + \beta_{\mathrm{dep}}\right). \tag{G.70}$$

Proof. On the event (G.54),

$$\left\|\widehat{\mathcal{T}} - \Phi_4(\theta_*)\right\|_F \le \left\|\widehat{\mathcal{T}} - \mathcal{T}_h\right\|_F + \|\mathcal{B}_{\mathrm{dep}}\|_F \le \varepsilon_n(\delta) + \beta_{\mathrm{dep}}.$$

Global optimality gives the corresponding fitted-residual bound

$$\left\|\widehat{\mathcal{T}} - \Phi_4(\widehat{\theta}_{\mathrm{glob}})\right\|_F \le \varepsilon_n(\delta) + \beta_{\mathrm{dep}}.$$

Hence

$$\left\|\Phi_4(\widehat{\theta}_{\mathrm{glob}}) - \Phi_4(\theta_*)\right\|_F \le \left\|\Phi_4(\widehat{\theta}_{\mathrm{glob}}) - \widehat{\mathcal{T}}\right\|_F + \left\|\widehat{\mathcal{T}} - \mathcal{T}_h\right\|_F + \|\mathcal{B}_{\mathrm{dep}}\|_F \le 2\left(\varepsilon_n(\delta) + \beta_{\mathrm{dep}}\right). \tag{G.71}$$

Condition (G.69) forces the estimator into $\mathcal{U}_*$. After alignment, Lemma D.4 yields (G.70). □

Proposition G.9 (Observable diagonal-model adequacy)**.** *Let*

$$\mathfrak{M}_B = \Phi_4(\Theta_B), \tag{G.72}$$

$$m_* = \mathrm{dist}(\mathcal{T}_h, \mathfrak{M}_B), \qquad \widehat{m} = \mathrm{dist}(\widehat{\mathcal{T}}, \mathfrak{M}_B). \tag{G.73}$$

Then, on the event (G.54)*,*

$$|\widehat{m} - m_*| \le \varepsilon_n(\delta). \tag{G.74}$$

For the finite net $\mathcal{N}_\rho$, define

$$\widehat{m}_\rho = \min_{\theta \in \mathcal{N}_\rho} \left\|\widehat{\mathcal{T}} - \Phi_4(\theta)\right\|_F.$$

Then

$$|\widehat{m}_\rho - m_*| \le \varepsilon_n(\delta) + L_B\rho. \tag{G.75}$$

In particular, $\widehat{m} > \varepsilon_n(\delta)$ certifies $m_* > 0$ on the same event.

Proof. The image $\mathfrak{M}_B$ is compact because Θ_B is compact and Φ_4 is continuous. Distance to any nonempty closed set in a normed space is one-Lipschitz, so

$$\left|\mathrm{dist}(\widehat{\mathcal{T}}, \mathfrak{M}_B) - \mathrm{dist}(\mathcal{T}_h, \mathfrak{M}_B)\right| \le \left\|\widehat{\mathcal{T}} - \mathcal{T}_h\right\|_F.$$

This proves (G.74). The Lipschitz bound (G.58) and the net property imply

$$\widehat{m} \le \widehat{m}_\rho \le \widehat{m} + L_B\rho,$$

which, combined with (G.74), proves (G.75). □

Corollary G.10 (Observed-context extension)**.** *Let $C \in [m]$ be observed. For each context e, center the activations and observations conditionally on $C = e$, and suppose*

$$\mathcal{T}_h^{(e)} = \sum_{k=1}^{K} \lambda_{ek} w_k^{\otimes 4}. \tag{G.76}$$

This condition holds, in particular, when the coordinates are independent conditional on $C = e$ and the residual noise is conditionally Gaussian and independent. Define the stacked map

$$\Phi_{\mathrm{ctx}}(\Lambda, W) = \left(\sum_k \lambda_{1k} w_k^{\otimes 4}, \ldots, \sum_k \lambda_{mk} w_k^{\otimes 4}\right)$$

with the product Frobenius norm. For context e, let

$$M_8^{(e)} = \left(\mathbb{E}\Big[\|h - \mathbb{E}[h \mid C = e]\|_2^8 \,\Big|\, C = e\Big]\right)^{1/8}$$

and, for n_e observations in that context, set

$$\varepsilon_e = C(M_8^{(e)})^4 \sqrt{\frac{D_4 \log(2mD_4/\delta)}{n_e}}.$$

If the stacked parameter is identifiable and regular on a compact parameter class, then Theorem G.6 and Corollary G.7 hold verbatim with Φ_4 replaced by Φ_{ctx} and

$$\varepsilon_{\mathrm{ctx}} = \left(\sum_{e=1}^{m} \varepsilon_e^2\right)^{1/2}. \tag{G.77}$$

Thus the marginal coordinates may be dependent through the observed context.

Proof. Conditional cumulant multilinearity gives (G.76) in each context. Apply the robust cumulant construction separately to each context and take a union bound over the m concentration events, absorbing the allocation of failure probability into the logarithms. The product-norm estimation error is bounded by (G.77). The compactness, separation, objective-comparison, and local-inverse arguments used above depend only on the normed observation space and therefore apply unchanged to the stacked map. □

Remark G.11 (Statistical and computational scope). For fixed sparsity, Theorem G.2 has the information-theoretic $K(d-1)/\epsilon^2$ parameter scaling. There are two deliberately different upper-bound routes. First, Theorem G.6 analyzes a global minimum-distance estimator after coordinatewise estimation of the complete symmetric fourth-order tensor. Its D_4 factor is the ambient cost of that particular full-tensor construction, Δ_* is a population separation threshold used only for global basin capture, and Corollary G.7 makes the estimator finite at the price of exponential worst-case computation. This route is an end-to-end statistical existence benchmark, not the proposed large-scale implementation.

Second, once an independently computed pilot lies in an identifiable chart, the sample-split tangent-score refinement above uses only $p = Kd$ robust scalar scores, is polynomial time, and has an $n^{-1/2}$ stochastic term governed by the Jacobian singular value $\widetilde{\gamma}$ evaluated at the pilot. It does not use D_4 or Δ_*. A sample-split post-entry variant of the practical spectral-initialized CBFR routine is therefore covered, but the theorem does not retroactively certify runs that reuse the pilot sample for refinement, and no polynomial-time global basin-entry guarantee is claimed throughout the full generic Waring range. The local refinement also does not remove the cost of the reported tensor-flattening pilot; it isolates the fact that, after entry, neither the statistical dimension nor the refinement step needs to scale with all D_4 coordinates.

The constants have different epistemic status. The population margin Δ_* is not estimated in the real-model experiments and should not be read as a verified property of those fits. The local singular value $\widetilde{\gamma}$ is directly computable from a returned parameter and can be reported as an a posteriori conditioning diagnostic. The moment scale M_8 is a distributional quantity appearing in confidence radii, not an input needed by either optimizer. Deriving adaptive confidence bounds for M_8 under weaker or self-normalized tail assumptions is separate from the identification result.

H Uniform-Support Geometry Beyond the Endpoint

This subsection analyzes the uniform-support local-coherence index $\mathcal{I}$, not the activation-conditioned recovery rate $\mathcal{R}_\nu$. Its purpose is to quantify how the conservative geometric component defined above degrades after the global endpoint becomes impossible. Exact double counting applies because $\mathcal{I}$ is a graph statistic of threshold-violating pairs; only $\mathcal{I} = 1$ implies recovery for every bounded-sparsity support law through (4).

Theorem H.1 (Uniform-support local-coherence envelope)**.** *Let $a_s = (2s-1)^2$ and $\eta = K/d$. Every unit-column dictionary satisfies*

$$\mathcal{I}(W, s) \le 1 - \frac{(\tau(d, K, s) - 1)_+}{a_s - 1} \tag{H.1}$$

$$= 1 - \frac{\big[(a_s - d)\eta - (a_s - 1)\big]_+}{(a_s - 1)(\eta d - 1)}. \tag{H.2}$$

If $d < a_s$, the continuous Welch threshold is

$$\eta_{\mathrm{W}} = \frac{a_s - 1}{a_s - d}, \tag{H.3}$$

and (H.2) *is locally linear in $\eta - \eta_{\mathrm{W}}$ but saturates for large η.*

Conversely, suppose a base frame $W^* \in \mathbb{R}^{d\times q}$ has $\mu(W^*) \le t_s - \rho$ for some $\rho > 0$ and $q \ge s$. For every sufficiently small cluster radius and $K = mq$, there exists a clustered dictionary satisfying

$$\mathcal{I}(W, s) \ge \frac{\binom{q}{s} m^s}{\binom{mq}{s}}, \tag{H.4}$$

$$\lim_{m\to\infty} \mathcal{I}(W, s) \ge \prod_{i=0}^{s-1} \left(1 - \frac{i}{q}\right) > 0. \tag{H.5}$$

The first bound corrects an apparent linear law: the exact Welch expression contains the denominator $\eta d - 1$, so a global linear decay cannot persist indefinitely. Indeed, when $d < a_s$ and $\eta = \eta_{\mathrm{W}} + \Delta$ with $\Delta \downarrow 0$,

$$\tau - 1 = \frac{(a_s - d)\Delta}{d\eta_{\mathrm{W}} - 1 + d\Delta} \tag{H.6}$$

$$= \frac{(a_s - d)^2}{a_s(d-1)}\Delta + \mathcal{O}(\Delta^2), \tag{H.7}$$

because

$$d\eta_{\mathrm{W}} - 1 = \frac{a_s(d-1)}{a_s - d}. \tag{H.8}$$

Thus the onset is genuinely first order, while

$$\lim_{\eta\to\infty} \frac{\tau - 1}{a_s - 1} = \frac{a_s - d}{d(a_s - 1)} \tag{H.9}$$

is finite. The cluster construction gives the corresponding positive plateau in the local-coherence index.

Why the uniform-support envelope is nonlinear. The numerator in (H.2) measures the excess Welch energy beyond the point at which all pairwise correlations could remain below the threshold. The denominator $\eta d - 1$ converts that excess energy into a fraction of the complete pair set. The expansion in (H.7) gives a first-order dependence on $\eta - \eta_{\mathrm{W}}$ at the threshold. As $\eta \to \infty$, both the number of atom pairs and the forced excess Gram energy scale quadratically in K, so their ratio converges to the constant in (H.2). A globally linear upper bound would eventually become negative and cannot bound a quantity in $[0, 1]$.

The cluster construction explains the same saturation geometrically. Duplicating each base direction into a tight neighborhood creates coherent within-cluster pairs, while supports that select at most one atom from each cluster retain the base frame's inter-cluster separation. The limiting product in (H.5) is the probability that s uniformly selected atoms occupy distinct clusters as the cluster multiplicity diverges. This plateau concerns uniformly selected subsets and does not imply activation-conditioned recovery; external atoms from the same cluster may still violate the full-dictionary ERC.

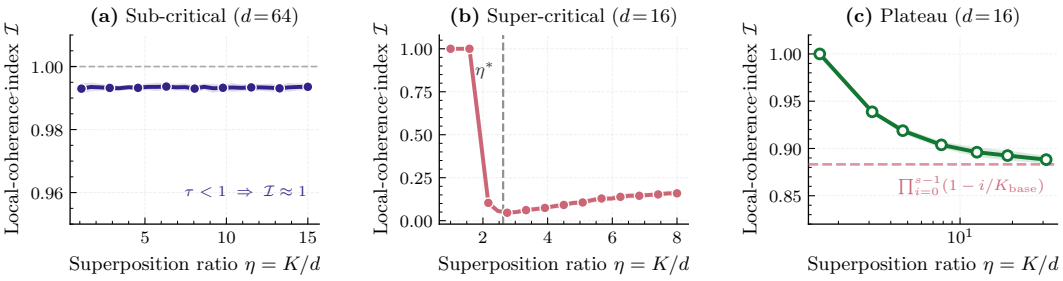


Figure H.1: **Uniform-support coherence geometry.** The figure compares optimized subcritical frames, frames evaluated across the Welch obstruction, and clustered constructions against the positive combinatorial plateau in (H.5).

Remark H.2 (Relation to support-conditioned recovery). Coherence regularization is meaningful as a conservative intervention: when it keeps $\mu_W < t_s$, Proposition 3.1 certifies every support law of sparsity at most s. Beyond that endpoint, model assessment should use $\mathcal{R}_{\widehat{\nu}_n}$ or $\mathcal{C}_{\widehat{\nu}_n}$ on observed activation supports. The uniform statistic $\mathcal{I}$ remains a decoder-geometry diagnostic for the recoverability component above; it is not a substitute for the realized-law quantities $\mathcal{R}_{\widehat{\nu}_n}$ and $\mathcal{C}_{\widehat{\nu}_n}$.

I Proof of the Uniform-Support Coherence Envelope

Proof of Theorem H.1. The upper envelope follows by counting threshold-violating atom pairs in two ways; the construction then shows why a positive plateau can persist beyond the worst-case endpoint. Let

$$\mathcal{E} = \big\{\{j,k\} : 1 \le j < k \le K,\ |\langle w_j,\, w_k\rangle| \ge t_s\big\}, \qquad e = |\mathcal{E}|, \qquad M = \binom{K}{2}. \tag{I.1}$$

Call pairs in $\mathcal{E}$ violating and all other pairs nonviolating. Every violating squared inner product is at most one, whereas every nonviolating squared inner product is strictly below $t_s^2 = 1/a_s$ and hence, in particular, at most $1/a_s$. Combining this with the Welch energy lower bound gives

$$\frac{K(K-d)}{2d} \le \sum_{j<k} |\langle w_j,\, w_k\rangle|^2 \tag{I.2}$$

$$\le e + \frac{M-e}{a_s} \tag{I.3}$$

$$= \frac{M}{a_s} + e\left(1 - \frac{1}{a_s}\right). \tag{I.4}$$

Now

$$\frac{K(K-d)}{2d} - \frac{M}{a_s} = \frac{K(K-1)}{2a_s}\left[\frac{(K-d)a_s}{d(K-1)} - 1\right] \tag{I.5}$$

$$= \frac{M}{a_s}(\tau - 1). \tag{I.6}$$

Subtracting M/a_s from (I.4) and multiplying by $a_s/(a_s-1) > 0$ yields

$$\frac{e}{M} \ge \frac{\tau - 1}{a_s - 1}. \tag{I.7}$$

When $\tau \le 1$, the right-hand side is nonpositive and the nonnegativity bound $e/M \ge 0$ is stronger. Hence, uniformly in all regimes,

$$\frac{e}{M} \ge \frac{(\tau-1)_+}{a_s - 1}. \tag{I.8}$$

Let N_{bad} be the number of size-s supports containing at least one violating edge. Count incidences

$$\mathfrak{I} = \{(E,S) : E \in \mathcal{E},\ |S| = s,\ E \subset S\}. \tag{I.9}$$

Each fixed violating edge is contained in exactly $\binom{K-2}{s-2}$ supports, so

$$|\mathfrak{I}| = e\binom{K-2}{s-2}. \tag{I.10}$$

A bad support contains at most $\binom{s}{2}$ unordered pairs and hence at most $\binom{s}{2}$ violating edges. Therefore

$$|\mathfrak{I}| \le N_{\mathrm{bad}}\binom{s}{2}. \tag{I.11}$$

Combining (I.10) and (I.11), and using $N_{\mathrm{bad}} = \binom{K}{s}(1 - \mathcal{I}(W,s))$, gives

$$1 - \mathcal{I}(W,s) \ge \frac{e\binom{K-2}{s-2}}{\binom{s}{2}\binom{K}{s}} \tag{I.12}$$

$$= \frac{e}{\binom{K}{2}} = \frac{e}{M}. \tag{I.13}$$

The binomial identity in the second line follows from

$$\binom{K}{s}\binom{s}{2} = \binom{K}{2}\binom{K-2}{s-2}, \tag{I.14}$$

which counts a size-s set together with a distinguished pair in two orders. Applying (I.8) to (I.13) proves

$$\mathcal{I}(W, s) \le 1 - \frac{(\tau(d, K, s) - 1)_+}{a_s - 1}. \tag{I.15}$$

To express the same result in terms of $\eta = K/d$, calculate

$$\tau - 1 = \frac{(K-d)a_s - d(K-1)}{d(K-1)} \tag{I.16}$$

$$= \frac{K(a_s - d) - d(a_s - 1)}{d(K-1)} \tag{I.17}$$

$$= \frac{(a_s - d)\eta - (a_s - 1)}{\eta d - 1}. \tag{I.18}$$

This proves (H.2).

For the constructive statement, let $W^* = [w_1^* | \cdots | w_q^*]$ satisfy $\mu(W^*) \le t_s - \rho$. Choose $\varepsilon > 0$ such that

$$\mu(W^*) + 2\varepsilon + \varepsilon^2 < t_s; \tag{I.19}$$

this is possible because the margin ρ is strictly positive. Around each w_a^* choose m distinct unit vectors $z_{a,1}, \ldots, z_{a,m}$ satisfying

$$\|z_{a,r} - w_a^*\|_2 \le \varepsilon. \tag{I.20}$$

Because $s \ge 2$ and $\mu(W^*) < t_s < 1$, the base frame contains at least two noncollinear directions; hence $d \ge 2$. Every spherical neighborhood then contains arbitrarily many distinct points.

If $a \ne b$, then for every r, r', write $e_a = z_{a,r} - w_a^*$ and $e_b = z_{b,r'} - w_b^*$. By Cauchy–Schwarz,

$$|\langle z_{a,r},\, z_{b,r'} \rangle| \le |\langle w_a^*,\, w_b^* \rangle| + \|e_a\|_2 + \|e_b\|_2 + \|e_a\|_2 \, \|e_b\|_2 \le \mu(W^*) + 2\varepsilon + \varepsilon^2 < t_s. \tag{I.21}$$

where the final strict inequality is (I.19). Thus every support selecting at most one member from each cluster passes the local-coherence test. There are $\binom{q}{s}$ ways to choose the occupied clusters and m^s ways to choose one vector from each, so

$$\mathcal{I}(W, s) \ge \frac{\binom{q}{s} m^s}{\binom{mq}{s}}. \tag{I.22}$$

Finally,

$$\frac{\binom{q}{s} m^s}{\binom{mq}{s}} = \frac{q(q-1)\cdots(q-s+1)m^s}{(mq)(mq-1)\cdots(mq-s+1)} \tag{I.23}$$

$$= \prod_{i=0}^{s-1} \frac{m(q-i)}{mq-i} \xrightarrow[m\to\infty]{} \prod_{i=0}^{s-1} \left(1 - \frac{i}{q}\right) > 0. \tag{I.24}$$

This proves the positive asymptotic plateau. □

J Secondary Consequence: Downstream Estimation

The preceding result concerns decoder geometry and support recovery. A separate consequence, useful only after a representation has been fixed, is a standard estimation bound for a norm-controlled probe. Once a representation is fixed, active-subdictionary energy controls the estimation component of a norm-bounded linear probe. Under $\|\phi(x)\|_0 \le s$ and $\|\phi(x)\|_\infty \le B$, the standard Rademacher calculation gives a rate of order $B\sqrt{s\rho_s(W)/n}$, where $\rho_s(W) = \max_{|S|\le s} \lambda_{\max}(W_S^\top W_S)$. This statement concerns estimation for already selected features and supplies no evidence for dictionary identification, coefficient recovery, semantic validity, or causal sufficiency. Raw isotropic residual noise reintroduces a dimension-dependent energy term.

K Experimental Configurations and Data Generation

The empirical study combines controlled SLR experiments, benchmark bottleneck experiments, and public-checkpoint geometry audits. This section specifies the generating distributions, dictionaries, datasets, optimization budgets, independent replicate units, and summary statistics used in each experiment. The representation-level framing is consistent with work on distributed vector representations, linear probes, and the linear representation hypothesis (Mikolov et al., 2013; Alain and Bengio, 2016; Park et al., 2023).

K.1 Sparse Linear Representation (SLR) Data Synthesis

Ground-truth overcomplete dictionaries are initialized with independent columns drawn uniformly from the unit sphere in $\mathbb{R}^d$. For the subcritical capacity experiments, the columns are optimized to reduce the largest absolute off-diagonal Gram entry, and optimization stops when the prescribed coherence threshold is reached or the stated iteration budget is exhausted. For the supercritical plateau construction, a base frame is first optimized and each base direction is then replaced by a cluster of distinct unit vectors lying within spherical radius $\epsilon = 0.02$.

For the cumulant-recovery experiments, sparse activations follow the Bernoulli–Exponential model used in Appendix C:

$$\phi_k(x) = Z_k E_k, \tag{K.1}$$

$$Z_k \sim \text{Bernoulli}(s/K), \qquad E_k \sim \text{Exponential}(\beta_k). \tag{K.2}$$

The support variables are independent, so the expected support size is exactly s. The rates β_k are linearly spaced between 1.0 and 3.0 to avoid nearly coincident component scales in the numerical decomposition. Equation (C.5) verifies that every coordinate has a nonzero fourth cumulant and finite eighth moment. Isotropic Gaussian observation noise $\xi \sim \mathcal{N}(0, \sigma^2 I_d)$ is then added. The Gaussian control replaces the non-Gaussian coordinates by an isotropic Gaussian source subspace in order to realize the rotational ambiguity of Proposition 3.10. The separate probe illustration uses bounded sparse activations and is interpreted only as a downstream scope check.

K.2 Replication and Reporting Conventions

An independent replicate is defined by the source of randomness that changes the fitted object: a stratified split seed for the bottleneck experiments, a dictionary-and-data seed for controlled recovery experiments, and a support-sampling seed for the public-checkpoint support audit. Deterministic quantities, including the all-pairs coherence scan of a fixed released decoder, are computed once per checkpoint and are not assigned artificial replicate counts. Arithmetic means and standard deviations are reported only in tables whose captions specify mean $\pm$ standard deviation; recovery tables based on five trials report the median and interquartile range. Success columns always give the numerator and denominator of the corresponding independent trials.

K.3 Synthetic Operational-Recovery Protocol

The controlled operational-recovery experiment uses $d = 64$, $K = 256$, and $s = 2$, with four independent dictionary seeds per geometry. For each seed and support-law condition, 500 held-out supports evaluate both support-conditioned rates. Separately, 20 independent noiseless basis-pursuit instances estimate coefficient-recovery success. Hazardous supports contain the inserted active–inactive confounder; the safe law samples from the complementary set. The sweep changes only the mixture weight

$$q\nu_{\text{hazard}} + (1-q)\nu_{\text{safe}},$$

leaving the dictionary construction and every other setting fixed.

K.4 Cross-Model Public-SAE Geometry Protocol

The protocol is motivated by the superposition and dictionary-learning view of neural features (Elhage et al., 2022; Bricken et al., 2023). Scaled SAE training and public releases make cross-checkpoint decoder audits possible (Gao et al., 2025; Rajamanoharan et al., 2024; Lieberum et al., 2024; Templeton et al., 2026).

Model	Release	SAE identifier	Hook	d	K	K/d
GPT-2 Small	`gpt2-small-res-jb`	`blocks.2.hook_resid_pre`	L2 resid-pre	768	24,576	32.0
GPT-2 Small	`gpt2-small-res-jb`	`blocks.6.hook_resid_pre`	L6 resid-pre	768	24,576	32.0
Pythia-70M-Deduped	`pythia-70m-deduped-res-sm`	`blocks.3.hook_resid_post`	L3 resid-post	512	32,768	64.0
Gemma-3 270M IT	`gemma-scope-2-270m-it-resid_post`	`layer_12_width_16k_l0_medium`	L12 resid-post	640	16,384	25.6

Table K.1: **Public SAE checkpoints used in the decoder-geometry audit.**

The audit uses the following released checkpoints:

The GPT-2 release provides residual-pre decoder checkpoints across layers, the Pythia release provides a $32{,}768$-feature residual-post SAE, and the Gemma release specifies width $16{,}384$ with target $L_0 = 60$. The two GPT-2 layers constitute a within-family layer comparison; they are not counted as two independent model families.

Let $v_1, \dots, v_K$ be the released decoder vectors and set $w_k = v_k / \|v_k\|_2$. For each checkpoint, the reported analysis computes

$$\mu_W = \max_{j<k} |\langle w_j, w_k \rangle|, \tag{K.3}$$

$$\mathcal{I}(W, 2) = 1 - \frac{\#\{j < k : |\langle w_j, w_k \rangle| \geq 1/3\}}{\binom{K}{2}}, \tag{K.4}$$

$$\rho_2(W) = 1 + \mu_W. \tag{K.5}$$

The pair scan is blockwise, exact up to floating-point arithmetic, and never stores the full $K \times K$ Gram matrix. During the same scan we count pairs above 0.5, 0.8, and 0.95. A separate sample of one million uniformly drawn distinct pairs estimates the median, 90th, 99th, and 99.9th percentiles of $|\langle w_j, w_k \rangle|$. These prevalence statistics distinguish isolated near-duplicate pairs from typical dictionary geometry.

A complementary frame diagnostic uses

$$F = \frac{d}{K} W W^\top. \tag{K.6}$$

For a unit-norm tight frame, $F = I_d$. We report its smallest and largest eigenvalues, condition number, and normalized deviation $d^{-1/2}\|F - I_d\|_F$. This diagnostic complements mutual coherence by summarizing global frame anisotropy.

For every checkpoint and standardized support size $s \in \{8, 16, 32\}$, we draw five independently seeded samples of $5{,}000$ uniformly random supports each. For Gemma we additionally evaluate the released native scale $s = 60$. The reported support statistic is

$$\widehat{\mathcal{I}}(W, s) = \frac{1}{N} \sum_{r=1}^{N} \mathbf{1}\left\{ \mu(W_{S_r}) < (2s-1)^{-1} \right\}, \tag{K.7}$$

with a 95% Wilson interval. For every sampled support we compute $\lambda_{\max}(W_{S_r}^\top W_{S_r})$. Its sample mean and 95th percentile describe typical restricted energy, while the largest observed value obeys

$$\widehat{\rho}_s^{\mathrm{LB}} \leq \rho_s(W) \leq \min\{s, 1 + (s-1)\mu_W\}. \tag{K.8}$$

For $s > 2$, the sampled maximum is a lower bound because the combinatorial support set is not exhaustively searched.

All reported worst-pair quantities use the exact blockwise all-pairs scan. The `--quick` option is used only for implementation checks and labels its sampled maximum as a lower bound on μ_W. Because the releases do not provide matched activation matrices under a common protocol, support sampling uses the stated uniform reference law. Accordingly, this audit does not estimate $\mathcal{R}_\nu$, $\mathcal{C}_\nu$, $\mathcal{T}_h$, or γ_*.

K.5 Benchmark Bottleneck Activations

We use four classification benchmarks spanning image, biomedical, chemical, and biological measurements: the scikit-learn OptDigits subset, Wisconsin Diagnostic Breast Cancer (WDBC), Wine, and Iris. The datasets contain $1{,}797$, 569, 178, and 150 examples, respectively. For each of three fixed split seeds, the data are stratified into 60% training, 15% validation, and 25% test partitions. Input standardization is fitted on the training partition and applied unchanged to validation and test examples.

A supervised rectified MLP produces a frozen bottleneck representation $h \in \mathbb{R}^d$. We use $(d, K) = (16, 32)$ for OptDigits, $(12, 24)$ for WDBC, and $(8, 16)$ for Wine and Iris. A TopK sparse autoencoder with $s = 2$ active coordinates reconstructs the bottleneck activations. The baseline minimizes normalized reconstruction error. The regularized variant adds

$$\lambda_{\mathrm{coh}} \beta^{-1} \log \sum_{i \neq j} \exp(\beta \, |\langle w_i, w_j \rangle|) \,, \qquad \lambda_{\mathrm{coh}} = 0.16, \quad \beta = 30. \tag{K.9}$$

with decoder columns renormalized after every update. Both variants receive 800 Adam updates with the same sparsity and overcomplete ratio $K/d = 2$.

All activation-dependent metrics are evaluated on the held-out test partition. For each test code, let S_i be its TopK activation support and let $\widehat{\nu}_n = n^{-1} \sum_i \delta_{S_i}$. Because $K \leq 32$ and $|S_i| = 2$, both operational quantities are computed exactly:

$$\mathcal{R}_{\widehat{\nu}_n}(W) = \frac{1}{n} \sum_i \mathbf{1}\{\mathrm{ERC}(S_i; W) > 0\}, \tag{K.10}$$

$$\mathcal{C}_{\widehat{\nu}_n}(W) = \frac{1}{n} \sum_i \mathbf{1}\{\Gamma(S_i; W) > 0\}. \tag{K.11}$$

We additionally report the uniform pair statistic

$$\mathcal{I}(W, 2) = \frac{1}{\binom{K}{2}} \sum_{i<j} \mathbf{1}\left\{ |\langle w_i, w_j \rangle| < \frac{1}{3} \right\} \tag{K.12}$$

and count seeds satisfying $\mu_W < 1/3$. These latter quantities are conservative decoder controls and are not substituted for recovery under $\widehat{\nu}_n$.

For the benchmark bottleneck activations, we quantify the higher-order signal with a randomized directional sketch. For $M = 96$ random unit directions a_m, define

$$S_4(H) = \left[\frac{1}{M} \sum_{m=1}^{M} \left(\frac{\widehat{\mathrm{cum}}_4(a_m^\top h)}{\widehat{\mathrm{Var}}(a_m^\top h)^2} \right)^2 \right]^{1/2}. \tag{K.13}$$

For each split, the activation statistic is divided by the median statistic from eight Gaussian samples matched in sample size and empirical covariance. A ratio above one records a fourth-order departure from these finite-sample Gaussian controls; it does not by itself establish the diagonal cumulant model used for identification.

K.6 Neural Architectures and Optimization Baselines

For the controlled representation-learning experiment, we use a tied-weight bottleneck multilayer perceptron. Sparse input coordinates are mapped through a linear encoder followed by a rectified linear unit, and the decoder is the transpose of the encoder. The shared weight matrix therefore defines the learned dictionary W. Training minimizes an importance-weighted mean squared reconstruction loss.

When investigating differentiable coherence regularization, a smooth surrogate penalty is added to the objective. Let $G = W^\top W$ denote the representation Gram matrix. The penalty uses a LogSumExp function over the off-diagonal elements defined as

$$\mathcal{L}_{\mathrm{coh}}(W) = \frac{1}{\beta} \log \sum_{i \neq j} \exp(\beta |G_{ij}|) \,. \tag{K.14}$$

We set $\beta = 20$. This value is fixed across the reported toy-model configurations and controls the smoothness of the LogSumExp approximation to the largest absolute off-diagonal Gram entry.

In the determined calibration $(d, K) = (10, 8)$, CBFR is compared with four SAE variants, K-SVD, Online Dictionary Learning, FastICA, and PCA; every method returns the requested K components without padding. The separate overcomplete benchmark uses $(d, K, s) = (8, 12, 3)$ and compares the spectral-initialized symmetric-CP estimator with the ℓ_1, TopK, Gated, and JumpReLU SAEs. Gated, switch-style, multilevel, and bidirectional-feature variants illustrate the continuing diversification of SAE objectives and parameterizations (Rajamanoharan et al., 2024; Mudide et al., 2025; Bussmann et al., 2025; Zhu et al., 2025). Each SAE uses $1{,}500$ Adam updates, batch size $1{,}024$, learning rate 3×10^{-4}, and unit-normalized decoder columns. These are fixed algorithm-specific budgets, not a claim of equal wall-clock time or equal floating-point cost.

K.7 Evaluation Metrics and Alignment

Dictionary recovery is evaluated modulo the unavoidable column-permutation and sign ambiguities. Given the generating dictionary W and an estimate $\hat{W}$, the Hungarian algorithm selects the minimum-cost bipartite matching under pairwise sign-invariant Euclidean distance. The reported maximum column error is

$$\mathcal{E}(\widehat{W}, W) = \max_{k \in [K]} \min_{z \in \{-1,1\}} \left\| \widehat{w}_{\pi(k)} - z w_k \right\|_2 . \tag{K.15}$$

Here π is the resulting optimal assignment. The metric is therefore invariant to a common column permutation and to independent sign flips.

K.8 Tabulated-Experiment Protocol Summary

Table K.2 consolidates the independent replicate unit, per-replicate budget, and summary statistic for every table reported in Appendix L. Experiment-specific architecture and optimizer settings remain in the preceding subsections.

Experiment	Independent replication	Budget and reporting
Public-SAE support audit	Five support-sampling seeds; all-pairs scan deterministic	$5{,}000$ supports per seed and s; one million pairs per checkpoint for bulk quantiles; exact all-pairs quantities and Wilson intervals for sampled proportions
Benchmark bottleneck activations	Three stratified split seeds	800 Adam updates per SAE; $M = 96$ directions and eight Gaussian controls per split; mean $\pm$ standard deviation and certificate count out of three
Synthetic operational recovery	Four dictionary-and-support seeds	500 held-out supports and 20 noiseless basis-pursuit instances per condition; arithmetic means, with standard deviations in the support-law sweep
Overcomplete identifiability	Five dictionary-and-data trials per (K, n)	Three projected-Adam restarts of 700 steps; median [interquartile range] and success count out of five
Overcomplete method comparison	Five dictionary-and-data trials per sample size	$1{,}500$ Adam updates per SAE; CBFR uses three 700-step restarts; median [interquartile range] and success count out of five

Table K.2: **Replication and reporting protocol for tabulated experiments.** Counts refer to independent experimental replicates, not optimizer restarts within one fitted model.

L Empirical Results

The main paper presents the comparisons needed for the central claims. This section supplies the full protocol-matched evidence: uncertainty summaries, negative controls, additional operating points, and diagnostics that delimit what each experiment can establish. The ordering mirrors the claim hierarchy rather than the chronology of the experiments, so decoder-only geometry appears before activation-conditioned and ground-truth recovery results.

This section evaluates the theory at four distinct levels: decoder geometry, activation-conditioned recovery, controlled support-law effects, and generating-dictionary identification. The ordering is deliberate. Decoder-only measurements test geometric feasibility but cannot establish operational recovery; activation-conditioned experiments add the empirical support law; controlled constructions isolate the active–inactive mechanism; and generating-dictionary experiments assess statistical identification under a known latent model. Section K specifies the data splits, replicate units, support budgets, optimization budgets, and reporting statistics.

L.1 Decoder-Only Geometry Audit of Public SAEs

We evaluate four public checkpoints spanning three language-model families to compare decoder geometry across both models and layers: GPT-2 Small residual-pre SAEs at layers 2 and 6, the Pythia-70M-Deduped layer-3 residual-post SAE, and a Gemma-3 270M IT layer-12 residual-post SAE. The audit is situated relative to scaled SAE studies and released model-wide SAE suites (Gao et al., 2025; Lieberum et al., 2024; Templeton et al., 2026). Their released dimensions range from $(d, K) = (640, 16{,}384)$ to $(512, 32{,}768)$, corresponding to K/d between 25.6 and 64.

For each checkpoint, decoder atoms are unit-normalized and every unordered atom pair is scanned blockwise. This computes μ_W, $\mathcal{I}(W, 2)$, the number of threshold-violating pairs, and $\rho_2(W) = 1 + \mu_W$ exactly up to floating-point arithmetic. A separate sample of one million pairs estimates bulk correlation quantiles, and $(d/K)WW^\top$ measures deviation from a normalized tight frame. For each $s \in \{8, 16, 32\}$, and additionally $s = 60$ for Gemma, five support-sampling seeds generate $5{,}000$ uniformly drawn supports each. These samples estimate $\mathcal{I}(W, s)$ and the mean, 95th percentile, and maximum observed $\lambda_{\max}(W_S^\top W_S)$. The observed maximum is labeled as a lower bound on $\rho_s(W)$ and is reported together with the deterministic upper bound $\min\{s, 1 + (s-1)\mu_W\}$.

The audit separates three quantities that are often conflated: the single worst pair, the prevalence of threshold violations, and the restricted energy of sampled subdictionaries. Their divergence is informative rather than contradictory: a decoder may contain a small number of extreme correlations while most sampled supports remain comparatively well conditioned. Because the checkpoints do not provide native activation supports under a common protocol, however, the audit cannot estimate $\mathcal{R}_\nu$ or $\mathcal{C}_\nu$. Matched activations are also required to evaluate the cumulant tensor $\mathcal{T}_h$ and the inverse-conditioning parameter γ_*.

A separate task experiment uses a public activation matrix from 300 Indirect Object Identification prompts. For a GPT-2 Small layer-8 residual-stream SAE with $24{,}576$ features, 146 features differ between successful and failed prompts after Holm correction, and 105 of those differences have $|d| > 0.8$. A logistic probe on the raw 768-dimensional residual stream obtains ROC AUC 0.929, compared with 0.927 for the top 100 SAE features, while ablating the strongest individual correlate does not restore task performance. The near-equality of probe performance shows that a sparse feature subset preserves substantial predictive information; the failed single-feature intervention shows that this information is not concentrated in one causally sufficient coordinate. This task experiment therefore concerns predictive compression and causal non-equivalence, not dictionary recovery, and is interpreted under the standard cautions on probe controls, task relevance, counterfactual removal, and causal abstraction (Hewitt and Liang, 2019; Ravichander et al., 2021; Elazar et al., 2021; Geiger et al., 2021; Belinkov, 2022).

L.2 Activation-Conditioned Bottleneck Experiment

The main paper reports the fourth-order diagnostic and the geometry–utility intervention, so those numerical summaries are not repeated here. This appendix records the complementary experimental information needed to interpret them. The bottleneck dimensions are $(d, K, s) = (16, 32, 2)$ for OptDigits, $(12, 24, 2)$ for WDBC, and $(8, 16, 2)$ for Wine and Iris. Each result uses three stratified train/test splits; the fourth-order statistic is evaluated only on held-out activations and is normalized by the median of eight covariance-matched Gaussian controls within the same split. The statistic therefore tests for directional fourth-order signal beyond covariance, while leaving the diagonal-core and global-identifiability assumptions as separate questions.

For the coherence intervention, the main text reports changes in mutual coherence, the exact pair statistic, reconstructed-activation task accuracy, and certificate counts. A distinct reconstruction metric

is useful because task accuracy can remain stable even when pointwise activation reconstruction changes. Table L.1 therefore reports only normalized test MSE, a quantity not displayed in the main paper.

Table L.1: **Complementary reconstruction diagnostic for the coherence intervention.** N-MSE is test reconstruction MSE normalized by activation variance, reported as mean ± standard deviation over three stratified split seeds. Geometric quantities and reconstructed-activation task accuracy are reported only in the main paper.

Dataset	TopK N-MSE	Regularized N-MSE	Δ N-MSE
OptDigits	0.101 ± 0.020	0.143 ± 0.012	$+0.042$
WDBC	0.008 ± 0.004	0.014 ± 0.011	$+0.006$
Wine	0.035 ± 0.002	0.034 ± 0.008	-0.001
Iris	0.032 ± 0.011	0.024 ± 0.003	-0.008

The N-MSE diagnostic sharpens the interpretation of the intervention. Coherence regularization increases normalized reconstruction error on OptDigits and WDBC, is essentially neutral on Wine, and improves it on Iris. Thus the geometric gains reported in the main paper do not arise from one common reconstruction tradeoff. The intervention changes the learned decoder geometry consistently, whereas its reconstruction cost depends on the data distribution. This distinction is important because neither low reconstruction error nor high task accuracy is, by itself, evidence for source independence or uniqueness of the latent directions.

L.3 Controlled Operational-Recovery Experiment

The main paper contains the completed support-law sweep, the paired-decoy transition, and the cross-geometry operational comparison. They are not reproduced here. The role of this section is instead to expose the construction that makes the two stress tests logically distinct and to connect the numerical experiments to the exact recovery criterion.

For the localized-defect experiment, the dictionary distribution is held fixed while the support law reallocates probability mass between safe and hazardous supports. Consequently every decoder-only geometric statistic is unchanged across the sweep. Proposition 3.1 implies that the support-conditioned certificate is a lower bound on the exact support-conditioned recovery rate for every empirical support law. Basis pursuit can still recover an individual coefficient vector when that sufficient support-level certificate fails, because the ERC is uniform over all coefficient values on the support. This explains why the empirical basis-pursuit rate need not coincide with either support-level quantity in the main table.

The paired-decoy construction varies geometry while holding the active support fixed. Let $S = \{1, \dots, s\}$, set $w_i = e_i$ for $i \in S$, choose $u \perp \operatorname{span}\{e_i : i \in S\}$, and add

$$w_{\pm} = \frac{\sum_{i=1}^{s} e_i \pm \tau u}{\sqrt{s + \tau^2}}. \tag{L.1}$$

Then

$$\mathrm{ERC}(S; W) = \Gamma(S; W) = 1 - \frac{s}{\sqrt{s + \tau^2}}. \tag{L.2}$$

The active representation of $\sum_{i=1}^{s} e_i$ has ℓ_1 norm s, whereas the two-decoy representation has ℓ_1 norm $\sqrt{s + \tau^2}$. Both the certificate and the basis-pursuit optimum therefore change at the same threshold $\tau = \sqrt{s(s-1)}$. Equality gives two representations with equal ℓ_1 norm and hence a non-unique optimum; strict positive margin selects the active representation. The main table samples this analytic transition on both sides of the threshold.

Together, the two constructions isolate complementary mechanisms. One changes which supports are visited without changing the dictionary; the other changes the active–inactive geometry without changing the active support. Their agreement is therefore structural rather than a consequence of reusing the same experimental degree of freedom.

L.4 Differentiable Coherence Regularization in Network Training

Theorem 3.3 shows that the uniform coherence certificate is impossible when the Welch lower bound reaches the Donoho–Elad threshold. Standard reconstruction training does not directly control this quantity. We therefore add a differentiable coherence penalty to the representation-learning objective. Because maximum coherence is nonsmooth, the implementation uses a LogSumExp surrogate over the off-diagonal entries of the Gram matrix.

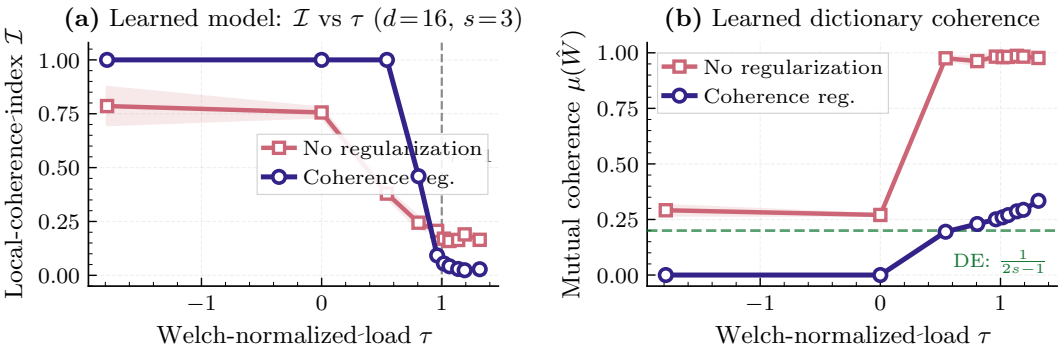


Figure L.1: **Observed certificate transition in the trained toy model.** The regularized and unregularized objectives are compared over the same configuration grid. The figure reports the learned mutual coherence and the resulting certificate status; it is an empirical comparison on this grid, not an estimate of an asymptotic phase-transition location.

Figure L.1 compares the learned dictionaries produced by the two objectives. On the evaluated grid, the unregularized objective develops larger extreme Gram entries and loses the full certificate at lower superposition ratios. The regularized objective delays this loss and keeps the measured local index at one throughout the sampled region below the observed Welch obstruction. This alignment between the analytic obstruction and the learned transition is evidence that coherence is an optimizable bottleneck, not a post hoc diagnostic. The conclusion remains finite-grid: the experiment neither identifies the population optimum nor estimates an asymptotic transition.

The cross-geometry operational comparison is reported only in the main paper. Its role is complementary to this training experiment: the training result shows that the worst-pair bottleneck can be optimized, whereas the operational comparison shows that even a fixed worst-pair geometry can have different consequences under different support laws.

L.5 Statistical Identifiability and Convergence

Theorem 3.9 and Proposition 3.10 distinguish population uniqueness from rotational ambiguity. We evaluate the corresponding recovery behavior in overcomplete regimes with

$$d = 8, \qquad K \in \{10, 12, 14\}, \qquad s = 3, \qquad \sigma = 0.05. \tag{L.3}$$

For each generated dictionary, every d-column submatrix is checked numerically. None is singular at the working precision, which is consistent with $\operatorname{krank}(W) = d$ for these realizations but is not a proof of exact algebraic Kruskal rank. The smallest observed singular values decrease as K increases, documenting weaker numerical conditioning. The implementation estimates the conventional empirical fourth cumulant, uses the spectral initialization in Appendix G, and runs three projected-Adam restarts of 700 steps. At each (K, n), five independent dictionary-and-data trials are evaluated. Recovery is measured by the maximum Hungarian-aligned, sign-invariant column error, and the prespecified reporting threshold is $\mathcal{E}(\widehat{W}, W) < 0.3$.

The main paper reports the $n = 10{,}000$ and $n = 25{,}000$ regimes together with the final success counts. Table L.2 contains only the two lower-sample regimes omitted from the main paper. These rows show where the practical routine has not yet reliably entered the identifiable basin and therefore complement, rather than duplicate, the main finite-sample comparison.

At these lower sample sizes the non-Gaussian settings are not uniformly resolved: increasing K weakens numerical conditioning and delays reliable basin capture, while the Gaussian control remains

Table L.2: **Low-sample overcomplete recovery regimes omitted from the main paper.** Entries are median maximum aligned column error with interquartile range in brackets over five trials.

Source	K	$\mathbf{n = 2.5k}$	$\mathbf{n = 5k}$
Bernoulli–Exp.	10	$1.088\,[1.001, 1.130]$	$0.271\,[0.236, 0.334]$
Bernoulli–Exp.	12	$1.133\,[0.869, 1.156]$	$0.750\,[0.307, 0.810]$
Bernoulli–Exp.	14	$1.001\,[0.611, 1.096]$	$1.087\,[1.052, 1.248]$
Gaussian control	12	$1.001\,[0.982, 1.077]$	$1.096\,[0.999, 1.239]$

rotationally ambiguous regardless of sample size. Read together with the later-sample results in the main paper, this pattern separates two effects that can otherwise look similar in a single run. Finite-sample or optimization failure can disappear as the identifying signal is estimated more accurately; population non-identifiability cannot.

L.6 Fixed-Budget Feature-Recovery Comparison

Sparse autoencoders are widely used in mechanistic interpretability (Bricken et al., 2023; Gao et al., 2025; Rajamanoharan et al., 2024). Their evaluation increasingly emphasizes principled interpretability and control tests, broad benchmark coverage, and sparse-probing utility (Makelov et al., 2025; Karvonen et al., 2025; Kantamneni et al., 2025). The determined-regime calibration and the overcomplete fixed-budget comparison are reported only in the main paper; this appendix records the controls needed to interpret those comparisons.

All overcomplete methods use $(d, K, s) = (8, 12, 3)$ and are evaluated with the same aligned column-error metric. Each SAE variant receives $1{,}500$ Adam updates under the architecture and regularization choices specified in section K.6; CBFR uses three projected local-optimization restarts of 700 steps after its cumulant-based candidate stage. The budgets are intentionally stated in algorithm-native update counts rather than converted into a synthetic common cost. Consequently the comparison supports a claim about recovery under the declared objectives and budgets, not a wall-clock-normalized ranking of algorithms. Hyperparameter-search effort is also not included in the budget, so the result should not be extrapolated to a universal ordering of methods.

The determined regime serves a different purpose from the overcomplete experiment. There, every method can return all requested components, which checks that gross implementation or alignment failures do not explain the comparison. The overcomplete setting then tests the regime central to the paper, where component recovery must remain meaningful when $K > d$. Keeping these roles separate prevents the calibration experiment from being mistaken for evidence about overcomplete identifiability.

L.7 Finite-Grid Sample-Requirement Diagnostic

Theorems G.2 and G.6 give, respectively, an intrinsic information lower bound and a conservative full-tensor upper bound. The sample-split tangent-score refinement gives an intrinsic-dimension local upper bound conditional on basin entry. The experiment below records finite-grid dependence on dimension, feature count, and target error; it is not designed to distinguish these three regimes asymptotically.

Figure L.2 reports the smallest evaluated sample size meeting the fixed recovery criterion for each tested dimension and target accuracy. Across the evaluated grid, larger parameter dimension and smaller target error both increase the observed requirement. The grid is compatible with the intrinsic Kd/ϵ^2 scale appearing in the lower bound and in the conditional local refinement, but it is too small to distinguish that behavior from the more conservative D_4/ϵ^2 full-tensor upper bound, to estimate constants, or to verify the global separation margin. No extrapolation to language-model-scale dictionaries is warranted.

Across the experiments, the evidence supports the paper's separation of geometric feasibility, operational recovery, and statistical identification. Decoder geometry determines conservative uniform certificates; the support law determines how often those certificates matter in use; fourth-order source structure determines whether the generating dictionary is identifiable; and finite-sample optimization

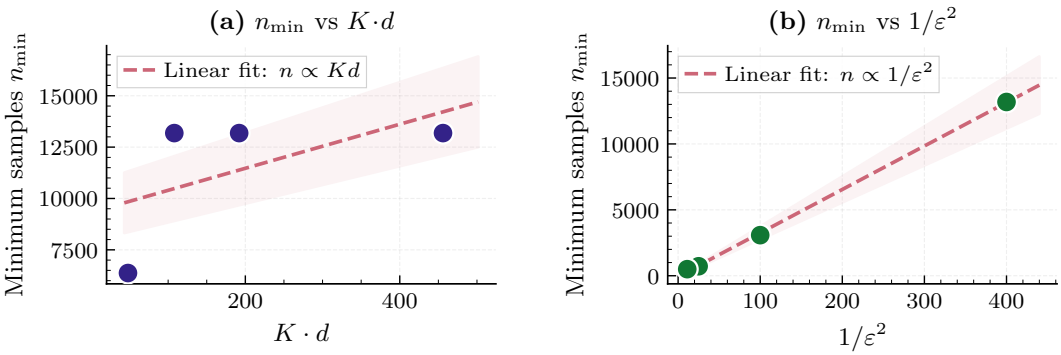


Figure L.2: **Sample-scaling diagnostic.** The figure reports the observed recovery frontier against the intrinsic parameter count Kd and target factor $1/\varepsilon^2$ over the evaluated configuration grid. This finite grid is descriptive and does not identify the ambient full-tensor exponent or certify basin entry.

determines whether the identifiable solution is reached in practice. These layers are empirically related but not interchangeable.

M Reproducibility and Evaluation Protocol

The main experiment suite ran on a cloud instance with a 16-core CPU, one NVIDIA GeForce RTX 4090 GPU with 24 GB of device memory, and 48 GB of host memory. The overcomplete recovery experiments were rerun in a separate CPU environment. Timing measurements, where reported, are compared only within the same hardware setting. All implementations use Python 3.11 and the package versions recorded in the released environment file.

The anonymous repository https://anonymous.4open.science/r/Nips-2-179C contains the experimental code, configuration files, data interfaces, and run configurations used for the reported results. An archival public repository will be released after the double-blind review period.

Claim-aligned evaluation. The empirical analysis follows the paper's decomposition into decoder geometry, activation-conditioned recovery, and generating-dictionary identification. Each claim is paired with the observable that directly tests it: full-dictionary certificates for support recovery, held-out support-law estimates for operational recovery, and fourth-cumulant fit together with aligned parameter error for identification. Population uniqueness, global separation, and local conditioning remain theorem-side conditions rather than being inferred from a successful optimization run. This organization keeps geometric, statistical, and computational conclusions distinct while presenting a single coherent evidence chain.

Controlled comparisons. Method comparisons use matched component count, sparsity, noise level, optimization budget, and model-selection effort. We report update counts and, when measured in a common environment, computational cost. Analyses of feature splitting and absorption, sparsity misspecification, and cross-run feature inconsistency motivate these controls (Chanin et al., 2026; Chanin and Garriga-Alonso, 2025; Song et al., 2026). Complementary work on identifiable SAE objectives, circuit analysis, and cross-model alignment further motivates evaluating decoder geometry, activation recovery, and causal evidence as separate axes (Cui et al., 2025; Nelson et al., 2026; Hindupur et al., 2026; Klindt et al., 2026; Ma et al., 2026a;b; Dunefsky et al., 2024; Marks et al., 2025; Thasarathan et al., 2025). Together, this protocol provides a reproducible basis for testing geometric feasibility, operational recovery, and statistical identification within a common framework.

N Practical CBFR Pseudocode

The routine below is the single algorithmic specification used for the reported cumulant-based recovery experiments. It should be read alongside sections G.1 and G.3. Those sections analyze robust tensor estimation and a global minimum-distance fit; the routine here replaces the global search

by an initialization-and-refinement strategy. This substitution is computationally useful but changes the nature of the guarantee: the theorem controls the target basin once the global fit is selected, whereas the practical optimizer may miss that basin and is therefore assessed through restart-level empirical recovery.

Algorithm N.1 Practical robust cumulant-based feature recovery (CBFR)

Require: observations $h_1, \dots, h_n$; component count K

1: Partition the observations into independent quartets.
2: For each distinct symmetric multi-index α, evaluate the unbiased fourth-cumulant kernel in (G.1).
3: Apply coordinatewise median-of-means aggregation and assemble $\widehat{\mathcal{T}}$.
4: If $K \le d$, initialize with joint diagonalization; if $K > d$, form the $(12) \mid (34)$ flattening and use its leading invariant subspace to construct normalized candidate directions.
5: Refine weights and unit-norm atoms by projected optimization of $\left\| \widehat{\mathcal{T}} - \sum_k \lambda_k w_k^{\otimes 4} \right\|_F$; use independent restarts under the budget reported in sections K.6 and K.8.
6: Return the best fitted restart and report tensor residual and signed-permutation-aligned column error as separate diagnostics.

Relation to the theory. The unbiased kernel and robust aggregation correspond to the tensor-estimation step analyzed in theorem G.4. Once an iterate lies inside the identifiable chart, the local inverse argument in theorem D.4 explains why a small tensor residual can control parameter error, subject to the local conditioning parameter. What the practical routine does not establish is global basin capture: the spectral or joint-diagonalization initialization is a computational heuristic for entering a useful chart, not a substitute for the separation argument in theorem G.6. This is why the empirical tables report both successful and failed restarts rather than conditioning the analysis on optimizer success.

Reporting discipline. The optimization residual, aligned dictionary error, and support-conditioned recovery metrics answer different questions and are reported separately. A low tensor-fit or reconstruction error is therefore never used as evidence for a stronger identification or operational-recovery claim.