IDENTIFIABLE INTERVENTIONS FROM NONIDENTIFIABLE SPARSE CODES

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders may admit multiple nonnegative sparse explanations of one activation, so a feature deletion can depend on the chosen explanation. We study identifying the activation change rather than recovering a unique code. Signed circuits characterize all exactly identifiable diagonal interventions, and a sharp modulus quantifies their noise amplification. An exact quotient theorem shows that certification and minimum-cost selection commute with arbitrary replication of decoder directions. On forest Gram graphs, a balanced-subtree algorithm certifies stability in $O(mk^4)$ arithmetic operations with rational counterexamples; positive paths admit optimal binary deletion over all feature subsets. For single-activation codes on arbitrary dictionaries, optimal attenuation has an explicit shortest-path formula. A separate construction proves an unbounded advantage over every stable binary deletion for any sparsity budget and arbitrarily overcomplete, nonidentifiable codes. Conversely, exact certification of one feature is coNP-complete even for well-conditioned, weakly correlated rational dictionaries. Uniform amplitude bounds transfer the structured certificates to densely perturbed dictionaries and control actual collateral change. Reproducible synthetic validation includes exact algorithm checks, 120 whole-input certification instances, and checkable refusal witnesses.

1 INTRODUCTION

Sparse autoencoders (SAEs) express a neural activation using a sparse combination of decoder directions. These directions are used to modify activations as well as describe them, for example by subtracting selected feature contributions (Cunningham et al., 2023). Yet accurate reconstruction need not determine a unique explanation. Instability across learned dictionaries (Paulo and Belrose, 2025) and ambiguity in fixed-dictionary inference (Carbo and Nalisnick, 2026) motivate a distinction: *must an explanation be unique before an intervention can be well defined?*

Consider identical columns $d_1 = d_2 = u$ and $x = u$. Both $(1, 0)$ and $(0, 1)$ are exact nonnegative codes. Deleting feature 1 produces either 0 or u, whereas deleting both always produces 0. Code recovery asks which coefficient vector is correct; intervention identification asks whether the requested synthesized change is the same for every allowed coefficient vector. The second requirement can hold even when the first is impossible.

We study a fixed decoder and all nonnegative k-sparse explanations within a reconstruction tolerance. This is an encoder-independent robustness standard: a guarantee remains valid whichever admissible explanation an encoder selects. The price is conservatism when some allowed codes never occur on the data distribution. We separate three questions: *which* interventions are determined in the noiseless model, *how much* noise they amplify, and *how efficiently* one can certify and select them.

Signed circuits characterize all exactly identifiable diagonal interventions, and a target-specific modulus gives their sharp worst-case noise amplification. An exact quotient theorem then shows that both certification and minimum-cost selection commute with arbitrary replication of decoder directions. This is the connection to nonidentifiable, highly overcomplete codes: algorithms operate on distinct prototypes, not on a uniquely recovered raw code. Dense deviations from this reference structure are handled by input-specific perturbation certificates with a uniform amplitude bound.

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.

On a forest Gram graph, a balanced-subtree dynamic program certifies stability in $O(mk^4)$ arithmetic operations and returns rational counterexamples. Positive paths admit optimal binary deletion over *all* feature subsets. For $k = 1$ on arbitrary dictionaries, a hyperbolic change of variables gives an explicit shortest-path attenuation envelope. An orthogonal-pair construction further proves an unbounded advantage over every stable binary deletion for *any* sparsity budget and arbitrarily large overcompleteness, while leaving the requested deletion complete. Outside these regimes, exact certification of even one feature is coNP-complete for well-conditioned, weakly correlated rational dictionaries.

Relation to prior work. Paired concept shifts (Joshi et al., 2026), dictionary stabilization (Fel et al., 2025; Nelson et al., 2026), and almost-orthogonal features (Miller et al., 2026) add information or change representations to improve identifiability or isolation. We instead certify agreement among explanations of an unchanged decoder. Our mathematical ingredients include block-sparse subspace geometry (Eldar et al., 2010), conformal circuit decomposition (Müller and Regensburger, 2016), Lipschitz extension (Azagra et al., 2021; McShane, 1934), and forest copositive structure (Shaked-Monderer, 2016). The algorithmic contribution is the preservation of separate positive and negative sparsity budgets, their Schur-complement compression, and exact intervention selection. Certification hardness is already established for RIP, NSP, and spark (Tillmann and Pfetsch, 2014; Bandeira et al., 2013); our reduction isolates a single prescribed deletion under nonnegative two-code constraints and bounded conditioning. The rational-witness argument uses the classical quadratic-programming perspective (Murty and Kabadi, 1987).

2 Intervention identifiability rather than code recovery

Let $D = [d_1, \dots, d_m] \in \mathbb{R}^{n\times m}$ have unit-norm columns. A fixed decoder bias is subtracted before analysis. Define

$$\Sigma_k^+ = \{z \geq 0 : \|z\|_0 \leq k\}, \qquad \mathcal{Z}_\varepsilon(x) = \{z \in \Sigma_k^+ : \|Dz - x\|_2 \leq \varepsilon\}. \tag{1}$$

For $w \in [0,1]^m$ and $W = \operatorname{diag}(w)$, the intervention is $I_w(x; z) = x - DWz$. Binary $w = \mathbf{1}_G$ deletes group G; write $P_G = \operatorname{diag}(\mathbf{1}_G)$. This operation preserves the reconstruction residual; replacing x by $D(I - W)z$ would also remove that residual and is a different operation. When optimizing a deletion target T, we impose $w_T = \mathbf{1}$: the target is never weakened to obtain stability.

The ambiguity and amplification are

$$\Delta_w(x) = \sup_{z,z' \in \mathcal{Z}_\varepsilon(x)} \|DW(z - z')\|_2, \tag{2}$$

$$\kappa_k(w) = \inf\{K \geq 0 : \|DWh\|_2 \leq K\|Dh\|_2 \ \forall h \in \mathcal{H}_k^+\}, \tag{3}$$

where $\mathcal{H}_k^+ = \{h : |\operatorname{supp}(h_+)| \leq k,\ |\operatorname{supp}(h_-)| \leq k\} = \Sigma_k^+ - \Sigma_k^+$ and an infeasible infimum is $+\infty$. All statements involving $\Delta_w(x)$ assume $\mathcal{Z}_\varepsilon(x) \neq \varnothing$.

Theorem 1 (Sharp target-specific modulus)**.** *The following are equivalent: DWz is determined by Dz on Σ_k^+; $\ker D \cap \mathcal{H}_k^+ \subseteq \ker(DW)$; and $\kappa_k(w) < \infty$. For every $\varepsilon > 0$,*

$$\sup_{x:\mathcal{Z}_\varepsilon(x)\neq\varnothing} \Delta_w(x) = 2\varepsilon\kappa_k(w). \tag{4}$$

Without computational restrictions, the minimax error for estimating DWz from $Dz + e$, $\|e\|_2 \leq \varepsilon$, equals $\varepsilon\kappa_k(w)$.

The upper bound follows from $\|D(z - z')\| \leq 2\varepsilon$. The matching lower bound uses a scaled difference and the midpoint of its two reconstructions. Finiteness follows because $\mathcal{H}_k^+$ is a finite union of polyhedral cones. The minimax upper bound is an application of Kirszbraun extension, not a proposed efficient decoder. Complete proofs are in Appendix A. The supremum in (4) is global and unbounded in activation magnitude; $\kappa_k(w) = \infty$ need not imply unbounded ambiguity at each particular x.

2.1 The exact resolution of identifiable interventions

A *circuit* is an inclusion-minimal dependent set of columns, with dependency $Dc = 0$ unique up to scale. It is k-feasible when $|\operatorname{supp}(c_+)|, |\operatorname{supp}(c_-)| \leq k$. Let $H_1, \dots, H_r$ be the connected components of the hypergraph whose edges are these circuit supports, with isolated columns included.

Theorem 2 (Circuit resolution). *For arbitrary real w, DWz is determined by Dz on Σ_k^+ if and only if w is constant on each H_i. In particular, the globally identifiable deletion groups are exactly unions of the H_i.*

For necessity, a cut circuit would leave a nontrivial dependency on a proper subset. For sufficiency, every admissible null vector is a sign-conformal sum of admissible circuit vectors. Appendix B gives a self-contained decomposition proof. Thus the smallest exactly identifiable superset of T is $\bigcup_{H_i \cap T \neq \varnothing} H_i$. This is an exact, noiseless statement; approximate dependencies require quantitative control and need not induce a unique finest stable partition.

2.2 Certification and selection commute with redundancy

Let $G_1, \dots, G_p$ partition the columns, with $d_j = u_{g(j)}$ for $j \in G_{g(j)}$, and let $U = [u_1, \dots, u_p]$. Write $s = \min(k, p)$; a subscript D or U specifies the dictionary used in a modulus or ambiguity.

Theorem 3 (Exact redundancy quotient). *Every finitely stable diagonal intervention is constant within each G_g. For $w_j = v_{g(j)}$,*

$$\kappa_{D,k}(w) = \kappa_{U,s}(v), \qquad \Delta_{D,w}(x) = \Delta_{U,v}(x) \tag{5}$$

at the same reconstruction tolerance. For a raw target T and any finite stability budget, minimum-cost stable selection reduces exactly to the prototype target $\bar{T} = \{g : G_g \cap T \neq \varnothing\}$ with objective $\sum_g c'_g v_g$, where $c'_g = \sum_{j \in G_g \setminus T} c_j$. This holds for binary or continuous weights.

Indeed, $e_i - e_j$ is an admissible null direction for equal columns, forcing $w_i = w_j$. Aggregation $a_g = \sum_{j \in G_g} z_j$ preserves both the reconstruction and the intervention, and every nonnegative s-sparse aggregate is realized by one representative per active group. Thus the entire attainable reconstruction–intervention relation is unchanged (Appendix G.1). The terms c'_g for $g \in \bar{T}$ account for the unavoidable collateral deletion of additional copies of a target direction. After grouping and reading the input, the structured algorithms below run on p prototypes, independently of the number of copies.

3 Exact certification on forest Gram graphs

Write $\Gamma = D^\top D$ and fix $K > 1$. Then

$$\kappa_k(w) \le K \quad \Longleftrightarrow \quad h^\top M_K(w) h \ge 0 \quad \forall h \in \mathcal{H}_k^+, \qquad M_K(w) = K^2\Gamma - W\Gamma W. \tag{6}$$

Requiring $M_K(w) \succeq 0$ is generally too strong: it ignores the separate sparsity budgets of the two explanations. The feasible weights nevertheless form a convex set, since each constraint $\|D\,\mathrm{diag}(h)w\| \le K\|Dh\|$ is convex in w.

Assume the graph with edges ij whenever $\Gamma_{ij} \neq 0$ is a forest. Select signs $\sigma_i \in \{\pm 1\}$ such that $\sigma_i \sigma_j \Gamma_{ij} < 0$ on each edge. A connected induced subtree U is *k-balanced* when it contains at most k vertices of either sign. The choice of the overall sign in each tree is immaterial.

Theorem 4 (Balanced-subtree criterion). *For a forest Gram graph, $w \in [0,1]^m$, and $K > 1$,*

$$\kappa_k(w) \le K \quad \Longleftrightarrow \quad M_K(w)[U] \succeq 0 \quad \textit{for every } k\textit{-balanced subtree } U. \tag{7}$$

Proof. Let $S = \mathrm{diag}(\sigma)$. The off-diagonal entries of $SM_K(w)S$ are nonpositive because $K^2 - w_i w_j > 0$. If a balanced principal submatrix is not PSD, taking coordinatewise absolute values of a negative vector cannot increase its quadratic form. Switching back gives a negative vector with at most k positive and k negative entries, violating (6).

Conversely, take $h \in \mathcal{H}_k^+$. In its nonzero support forest, remove every edge with positive contribution $M_{ij} h_i h_j$. Each remaining connected component has the sign pattern σ or $-\sigma$, hence is k-balanced. A forest has no extra internal edge that was removed: any such edge would create a cycle. Each component's quadratic contribution is nonnegative by hypothesis. Adding back the removed nonnegative terms proves the claim. □

This preserves the two-code sparsity constraint explicitly. It allows singular dictionaries and arbitrarily strong admissible edge correlations, subject only to $\Gamma \succeq 0$. The structural assumption is exact. In fact, a unit-column forest dictionary satisfies $\mathrm{rank}(D) \geq m/2$, and a connected one satisfies $\mathrm{rank}(D) \geq m - 1$ (Appendix C.5). Hence raw dictionaries with $m > 2n$ cannot have forest Gram graphs. Theorem 3 places this assumption on the *quotient* instead; Section 7 transfers its certificate to dense column perturbations rather than thresholding away Gram entries.

3.1 A DYNAMIC PROGRAM WITH EXACT COUNTEREXAMPLES

There can be exponentially many connected subtrees. Root each tree, set $b_v = K^2 - w_v^2 > 0$ and $a_{vu} = |\Gamma_{vu}|(K^2 - w_v w_u)$, and eliminate descendants towards the root. For each budget (p, q), store the smallest root Schur complement $d_v(p, q)$ among rooted connected subtrees with that sign count. If all chosen child states are positive,

$$d_v(p,q) = b_v - \max_{\text{admissible child allocations}} \sum_{u \text{ chosen child}} \frac{a_{vu}^2}{d_u(p_u, q_u)}. \tag{8}$$

The recurrence includes the sign count of v. A smaller positive child denominator is always more dangerous, so one scalar per budget suffices. Child tables are merged by two-dimensional knapsack convolution.

Theorem 5 (Exact forest algorithm)**.** *Given a valid forest Gram matrix, the decision $\kappa_k(w) \leq K$ uses $O(mk^4)$ arithmetic operations. For rational Γ, w, K^2, the algorithm has polynomial bit complexity. Every rejection yields a polynomial-size rational $h \in \mathcal{H}_k^+$ with $h^\top M_K(w) h < 0$.*

A negative Schur complement gives a witness by back-substitution. A zero complement is retained as a singular PSD state, not divided by an artificial regularizer. Connecting such a state to one more node within budget creates a negative witness. The complete invariant, zero-pivot argument, witness construction, and pseudocode are in Appendix C. The displayed operation count excludes Gram construction and validation. Polynomial bit complexity follows because each stored pivot is a ratio of determinants of principal matrices of order at most $2k$; arbitrary-precision operations are not unit-cost (Appendix C).

4 SELECTING THE LEAST COSTLY DELETION

Take a positive path Gram matrix: $\Gamma_{ij} = 0$ for $|i - j| > 1$ and $\Gamma_{i,i+1} > 0$. The switching signs alternate. Connected balanced subtrees are exactly intervals of length at most $2k$. Set $L = \min(2k, m)$.

Theorem 6 (Local windows and optimal path deletion)**.** *A weight vector is globally K-stable if and only if*

$$K^2\Gamma[I] - W_I \Gamma[I] W_I \succeq 0 \quad \textit{for every length-}L\textit{ interval } I. \tag{9}$$

For nonnegative additive costs c_j and target T, the minimum of $\sum_{j \notin T} c_j w_j$ over stable binary w satisfying $w_T = \mathbf{1}$ is computable in $O(m 2^{2k} k^3)$ arithmetic operations.

Every shorter interval is contained in a length-L interval, so PSD of these maximal windows suffices. To optimize, scan left to right, retain the last $L - 1$ bits, and check the new window when appending a bit. Prefixes with the same suffix admit identical future completions, so retaining only the cheapest one is exact. Costs represent declared combinatorial selectivity, not automatically an activation-space norm. No clustering hierarchy is fixed in advance.

For continuous weights, (9) has an exact semidefinite representation. With $\Gamma[I] = R_I^\top R_I$, each window is equivalent to

$$\begin{pmatrix} K^2\Gamma[I] & W_I R_I^\top \\ R_I W_I & I \end{pmatrix} \succeq 0. \tag{10}$$

A linear objective gives an SDP; general convex costs give an exact convex optimization problem. Independently of graph structure, $k = 2$ admits an exact $O(m^4)$-constraint SDP by the classical identity $\mathcal{COP}_r = \mathbb{S}_+^r + \mathbb{N}^r$ for $r \leq 4$; Appendix D states the full construction. The support enumeration, rather than within-support relaxation, is the source of their size.

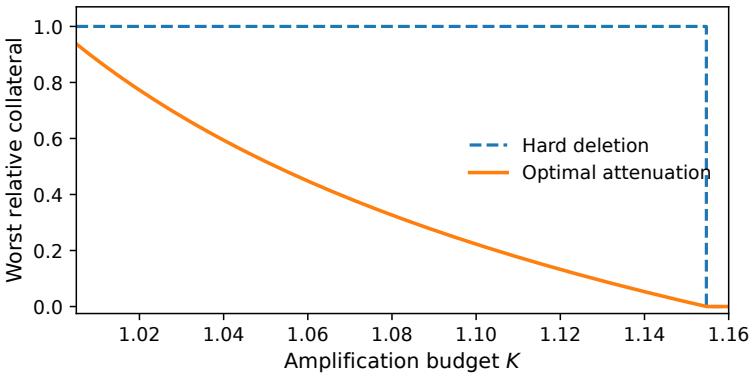


Figure 1: Analytic path family with neighbor correlation $1/2$, $k = 1$, and target $\{1\}$. Binary deletion has worst relative collateral 1 below $K = 2/\sqrt{3}$. The optimal attenuation's collateral approaches zero. Curves evaluate the proved formulas. Replicating every column and targeting all copies of feature 1 leaves both curves unchanged by Theorem 3.

5 Optimal attenuation can outperform every hard group

For $k = 1$ and any dictionary, the full continuous problem has a closed-form solution. Only positive correlations matter because the difference of two nonnegative one-sparse codes has opposite signs on its two coordinates. For each pair with $\rho_{ij} = d_i^\top d_j > 0$, assign the edge length

$$\ell_{ij} = \operatorname{arctanh} \sqrt{1 - \rho_{ij}^2}, \qquad a_K = \operatorname{arctanh}(1/K). \tag{11}$$

Identical directions have zero-length edges; nonpositive correlations impose no edge. Distances are shortest-path distances, with $+\infty$ across disconnected components.

Theorem 7 (Least collateral attenuation)**.** *For $k = 1$, $\kappa_1(w) \le K$ is equivalent to $|s_i - s_j| \le \ell_{ij}$ on every edge, where $s_i = \operatorname{arctanh}(w_i/K)$. For a nonempty target T, the pointwise least feasible attenuation with $w_T = \mathbf{1}$ is*

$$w_j^* = K \tanh\big([a_K - \operatorname{dist}(T, j)]_+\big). \tag{12}$$

It minimizes every coordinatewise nondecreasing cost of extra attenuation.

Indeed, the two-coordinate quadratic condition is $(K^2 - w_i^2)(K^2 - w_j^2) \ge \rho_{ij}^2 (K^2 - w_i w_j)^2$. The hyperbolic tangent subtraction identity transforms it into the edge constraint. Every feasible function must lie above the distance envelope from T; the envelope itself satisfies every edge constraint. The distance envelope is a standard real-valued Lipschitz-extension construction (McShane, 1934); the intervention-specific step is the exact hyperbolic reduction. Appendix E includes antiparallel and duplicate columns.

Binary masks can only take $s_j \in \{0, a_K\}$. Thus edges with $\rho_{ij} > \sqrt{1 - K^{-2}}$ cannot be cut, recovering a hard grouping rule. Continuous attenuation need not be constant on those components.

For a strict separation, take $d_i = (e_i + e_{i+1})/\sqrt{2}$, $i = 1, \dots, m$, and $T = \{1\}$. Neighbor correlation is $1/2$, and all other correlations are zero. At $K = 1.1$, every feasible binary intervention deletes all m features, whereas

$$w^* = (1, 0.2227133298 \dots, 0, \dots, 0). \tag{13}$$

The additive collateral costs are $m - 1$ and $0.2227133298 \dots$. More intrinsically, under $k = 1$, $\sup_{z \in \Sigma_1^+ \setminus \{0\}} \|D(W - P_T)z\|_2 / \|Dz\|_2 = \max_{j \notin T} w_j$. As $K \uparrow 2/\sqrt{3}$ from below, every binary mask still has relative collateral 1, while (12) has collateral tending to zero. The target remains fully deleted throughout. The arbitrary-dictionary formula is specific to $k = 1$: a three-column circuit imposes extra joint constraints at $k = 2$. The next result obtains an all-sparsity separation on a different, explicitly specified family.

Corollary 8 (Separation at every sparsity and redundancy)**.** *For any integer $k \geq 1$, integer expansion factor $r \geq 2$, and $\eta > 0$, there are a rational unit-column dictionary with $m/n = r$, nonunique nonnegative k-sparse codes, a target T, and a rational $K > 1$ such that every K-stable binary deletion with $w_T = \mathbf{1}$ has relative collateral*

$$\mathcal{C}_D(w;T) := \sup_{\substack{z\in\Sigma_k^+ \\ Dz\neq 0}} \frac{\|D(W-P_T)z\|_2}{\|Dz\|_2} = 1,$$

whereas a K-stable attenuation with $w_T = \mathbf{1}$ has $\mathcal{C}_D(w;T) < \eta$.

Take $q \geq k$ orthogonal pairs $u_{2g-1} = e_{2g-1}$ and $u_{2g} = (3e_{2g-1} + 4e_{2g})/5$, and repeat each direction r times. Target every copy of each odd direction. For $1 < K < 5/4$, every stable binary mask deletes all columns, while the optimal extra attenuation in each pair is

$$c_K = \frac{K(1-4K/5)}{K-4/5} \longrightarrow 0 \qquad (K \uparrow 5/4). \tag{14}$$

The pairwise PSD conditions certify this attenuation for *all* difference vectors, not only one-sparse pairs. Nonnegative inter-column correlations give $\mathcal{C}_D = c_K$; one extra feature attains equality, and so do codes on several orthogonal extra directions. At $K = 6/5$, $c_K = 3/25$. Appendix E.2 proves the construction and its exact optimality.

6 A COMPUTATIONAL LIMIT EVEN FOR WELL-CONDITIONED DICTIONARIES

Theorem 9 (Complexity of exact certification)**.** *When k is part of the input, deciding $\kappa_k(\mathbf{1}_{\{0\}}) \leq K$ is coNP-complete, even for rational unit-column dictionaries with $\frac{1}{2}I \preceq D^\top D \preceq \frac{3}{2}I$ and pairwise absolute correlations $O(1/m)$. The budget is represented by rational K^2. The reduction has an inverse-polynomial gap, not a constant-factor approximation gap.*

We reduce from CLIQUE (Karp, 1972). Given an N-vertex graph with adjacency A, let $\ell = \lceil\sqrt{2N}\rceil$, $t = \ell^{-2}$, and

$$\Gamma = \begin{pmatrix} 1 & t\mathbf{1}^\top \\ t\mathbf{1} & I - tA \end{pmatrix}. \tag{15}$$

Diagonal dominance gives the stated eigenvalue bounds. Normalizing a dangerous direction to $h_0 = 1$, all additional positive coordinates can be discarded without increasing the denominator. For a negative support S, the best magnitude is $t(I - tA_S)^{-1}\mathbf{1}$, which is nonnegative. Consequently

$$\kappa_k(\mathbf{1}_{\{0\}})^2 = (1 - \max_{|S|\leq k} r(S))^{-1}, \qquad r(S) = t^2\mathbf{1}^\top(I - tA_S)^{-1}\mathbf{1}. \tag{16}$$

A k-clique attains $r_* = t^2k/[1 - t(k-1)]$. Without one, every candidate satisfies $r(S) \leq r_* - 2t^3$: a missing edge loses $2t^3$ already in the first nonconstant Neumann term, and smaller supports lose at least t^2. Choose $K^2 = [1 - (r_* - t^3)]^{-1}$. A rational dictionary realizing (15) is constructed explicitly in Appendix F. For membership in coNP, minimize a violating signed quadratic form on a simplex. The stationary equations on its positive support have a rational feasible point of polynomial encoding length with the same negative value (Appendix F.4). Thus every unsafe rational instance has a short checkable counterexample, even when finding one is hard. Unlike coefficient nonidentifiability, this obstruction persists for full-column-rank dictionaries.

7 DENSE PERTURBATIONS: UNIFORM INPUT CERTIFICATES AND REFUSAL

The exact quotient extends to approximate groups by controlling *column* errors. This retains the original decoder in the intervention, including every small Gram entry.

Proposition 10 (Perturbation transfer)**.** *Suppose $D = \bar{D} + E$, $\max_j \|e_j\|_2 \leq \delta$, and all $z \in \mathcal{Z}_\varepsilon(x)$ satisfy $\|z\|_1 \leq B$. If $\kappa_{\bar{D},k}(w) \leq K$ and $w \in [0,1]^m$, then*

$$\Delta_w(x) \leq 2K\varepsilon + 2\delta B(K+1). \tag{17}$$

Writing $h = z - z'$ gives $\|h\|_1 \le 2B$; apply the nominal certificate to $\bar{D}h$ and bound both error terms. A near-collinear example in Appendix G shows that the order-δB term is necessary even at zero reconstruction error.

Theorem 11 (Redundancy-independent robust certificate)**.** *Let $\bar{d}_j = u_{g(j)}$, $w_j = v_{g(j)}$, $s = \min(k, p)$, and $\kappa_{U,s}(v) \le K$. Assume $\max_j \|d_j - \bar{d}_j\|_2 \le \delta$ and U has unit columns and entrywise nonnegative Gram matrix. If $\delta < s^{-1/2}$, every feasible code obeys*

$$\|z\|_1 \le B_{\mathrm{cone}}(x) := \frac{\|x\|_2 + \varepsilon}{s^{-1/2} - \delta}. \tag{18}$$

Alternatively, any a satisfying $D^\top a \ge \mathbf{1}$ gives the uniform bound $B_{\mathrm{dual}}(x) = a^\top x + \varepsilon\|a\|_2$, without the restriction on δ. Use either valid bound, or their minimum, as B in (17)*. For a target that is a union of complete groups, $w_T = \mathbf{1}$, and $c = \max_{j \notin T} w_j$ (zero if T is all columns), the actual collateral satisfies*

$$C_{w|T}(x) := \sup_{z \in \mathcal{Z}_\varepsilon(x)} \|D(W - P_T)z\|_2 \le c\big(\|x\|_2 + \varepsilon + 2\delta B\big). \tag{19}$$

Proof. For $\alpha_g = \sum_{j \in G_g} z_j$, nonnegativity of the prototype Gram gives $\|U\alpha\|_2 \ge \|\alpha\|_2 \ge \|\alpha\|_1/\sqrt{s} = \|z\|_1/\sqrt{s}$. Hence $\|Dz\|_2 \ge (s^{-1/2} - \delta)\|z\|_1$, proving (18). A feasible dual vector gives $\|z\|_1 \le a^\top Dz \le a^\top x + \varepsilon\|a\|_2$. For extra weights $0 \le b_g \le c$, the same Gram nonnegativity gives $\|U \operatorname{diag}(b)\alpha\|_2 \le c\|U\alpha\|_2$. Adding the column error and using $\|U\alpha\|_2 \le \|x\|_2 + \varepsilon + \delta B$ proves (19). □

Here B bounds the *entire* feasible set, not just one encoder output or a tested pair. In particular, both guarantees are independent of group sizes. For $A_x = \|x\|_2 + \varepsilon > 0$, using (18) gives the explicit relative ambiguity bound

$$\frac{\Delta_w(x)}{A_x} \le \frac{2K\varepsilon}{A_x} + \frac{2\delta(K+1)}{s^{-1/2} - \delta}.$$

Thus the quotient certificates remain informative in a nonzero neighborhood of the exact reference model. Appendix G gives alternatives for signed prototype correlations, subspace groups with noncollinear redundancy, and separated-prototype grouping.

Refusal requires a lower witness. The triangle inequality gives, for any $w_T = \mathbf{1}$,

$$\Delta_T(x) \le \Delta_w(x) + 2C_{w|T}(x). \tag{20}$$

A feasible pair with target discrepancy greater than $\tau + 2b$ therefore rules out *every* intervention having both ambiguity at most τ and collateral at most b. A certificate upper bound exceeding τ is instead *inconclusive*: it is not a lower witness. Coarsening need not improve stability (Appendix I). For a downstream L_f-Lipschitz readout, a certified activation-change radius r implies an effect radius $L_f r$.

8 Validation, coverage, and limitations

We distinguish implementation verification from the usefulness of a whole-input certificate. All results below were generated by the supplied deterministic scripts and independently checked where indicated; Appendix J specifies the exact and numerical parts.

Exact algorithms and analytic formulas. The algorithm suites compare forest certification with exhaustive balanced subtrees, path selection with all target-containing masks, and the single-activation envelope with an independent LP. The all-sparsity construction is checked in 45 configurations: 276 target-containing binary masks are covered, with 264 explicit obstructions to every mask other than full deletion. The soft candidate passes a full-matrix rational PSD test, hence all sparsity budgets simultaneously.

Check	Instances	Result
Forest DP / exhaustive subtrees	580	0 mismatches
Balanced subtrees / negative witnesses	11,860 / 135	Exact; all witnesses valid
Path optimum / all target-containing masks	72 / 3,024	0 cost mismatches
Distance envelope / independent LP	120	Max error 5.6×10^{-16}
CLIQUE reduction / circuit masks	4,096 / 112	0 mismatches
All-sparsity binary obstructions	264	All exact witnesses valid
Uniform-input certificates / sampled pairs	120 / 1,920	All exact checks valid
Input-specific refusal witnesses	6	All exact witnesses valid

Table 1: Reproducible synthetic verification. Forest, path, all-sparsity, uniform-input, and refusal checks use rational arithmetic. The reduction includes 256 exact-rational cases; the remaining reduction cases and orthant-LP comparisons are numerical.

Whole-input certification under dense perturbations. We use the rational orthogonal-pair prototypes from Corollary 8, $p \in \{4, 8\}$, 3 or 8 copies per prototype, and $k \in \{1, 2, 4\}$. Independent rational unit-sphere perturbations preserve unit columns; every raw Gram graph contains a triangle, including at zero perturbation, so none is a forest. The nominal budget is $K = 6/5$, with extra weight $3/25$. Each instance supplies an exactly feasible dual vector $D^\top a \geq \mathbf{1}$ and outward rational norm bounds. These premises certify the whole feasible code set, not merely the 1,920 sampled feasible pairs.

The decision threshold is $0.3(\|x\|_2 + \varepsilon)$, tested with conservative rounding. Across five perturbation parameters, 77 of 120 instances are certified and 43 are inconclusive. The success counts are $24, 24, 24, 5, 0$ out of 24 at parameters $0, 0.001, 0.005, 0.02, 0.05$, respectively; the generator recomputes the actual maximum column error after perturbation. The loss of coverage makes the robustness–approximation tradeoff visible rather than converting a loose bound into an instability claim. Six separate exact code-pair examples certify the refusal condition (20). A standalone checker verifies all uniform-certificate premises and negative witnesses without importing the certificate generator.

Limitations. The model is global over the declared nonnegative sparse-code family, not distribution-specific. Exact raw forests obey the rank restriction above; large overcompleteness is handled through structured quotients, and dense perturbations use sufficient input certificates. The arbitrary-dictionary distance formula applies only to $k = 1$, while Corollary 8 is an all-sparsity result for orthogonal pairs. Binary path selection minimizes a declared additive cost, not an arbitrary collateral norm. The complexity lower bound has an inverse-polynomial rather than constant gap. The validation establishes implementation consistency and certificate coverage on specified synthetic models, not trained-SAE steering performance. The SDP formulations are exact mathematical descriptions; they are not reported numerical implementations.

Conclusion. Recovering a unique sparse explanation is unnecessary for identifying its intervention. Signed circuits determine the exact intervention units; target-specific moduli quantify their noise sensitivity. Redundancy preserves certification and optimal selection after quotienting, and partial attenuation can be arbitrarily more selective than every stable hard deletion even with many active features and arbitrarily nonunique codes. Forest algorithms and robust input certificates give constructive guarantees, while coNP-completeness delineates the cost of exact generic certification.

AI USE STATEMENT

Generative AI was used extensively in this project, including research exploration, formulation and revision of mathematical claims, proof drafting and checking, literature search, construction of synthetic dictionaries and code pairs, generation of validation code, analysis of computed outputs, figure preparation, and manuscript writing and editing. The computational results reported here were produced by the executable scripts supplied with the paper, rather than supplied as unsupported generated measurements. AI assistance and finite tests do not replace the submitting authors' responsibility for correctness, attribution, and an accurate disclosure in the submission form.

ETHICS STATEMENT

This work studies deterministic ambiguity of representation-relative interventions under an explicitly declared sparse-code model. It does not establish semantic correctness, fairness, safety of deploying a steered model, or the identification of real-world causal variables. Treating a mathematical stability certificate as such an endorsement could be misleading. The validation uses synthetic dictionaries and code pairs only, without human participants, personal data, or model deployment.

REPRODUCIBILITY STATEMENT

The appendix supplies proofs, assumptions, counterexamples, algorithm recurrences, and the synthetic validation protocol. The anonymous supplement contains executable exact-rational certification and path-selection implementations, independent validation routines, deterministic seeds, raw results, negative witnesses, whole-input certificate records, two standalone witness checkers, and figure-generation code. No trained models or external datasets are required. The reported SDP formulations are not presented as implemented numerical experiments. Reproduction commands and the tested dependency versions are included in the supplement.

REFERENCES

Daniel Azagra, Erwan Le Gruyer, and Carlos Mudarra. Kirszbraun's theorem via an explicit formula. *Canadian Mathematical Bulletin*, 64(1):142–153, 2021. doi: 10.4153/S0008439520000314. URL https://doi.org/10.4153/S0008439520000314.

Afonso S. Bandeira, Edgar Dobriban, Dustin G. Mixon, and William F. Sawin. Certifying the restricted isometry property is hard. *IEEE Transactions on Information Theory*, 59(6):3448–3450, 2013. doi: 10.1109/TIT.2013.2248414. URL https://arxiv.org/abs/1204.1580v2.

Alessa Carbo and Eric Nalisnick. Which sparse code? identifiability failures in SAE inference. In *Scientific Methods for Understanding Deep Learning Workshop at ICLR*, 2026. URL https://openreview.net/forum?id=JiyytGKbA9.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. arXiv:2309.08600, 2023. URL https://arxiv.org/abs/2309.08600.

Yonina C. Eldar, Patrick Kuppinger, and Helmut Bölcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Transactions on Signal Processing*, 58(6):3042–3054, 2010. doi: 10.1109/TSP.2010.2044837. URL https://doi.org/10.1109/TSP.2010.2044837.

Thomas Fel, Ekdeep Singh Lubana, Jacob S. Prince, Matthew Kowal, Victor Boutin, Isabel Papadimitriou, Binxu Wang, Martin Wattenberg, Demba Ba, and Talia Konkle. Archetypal SAE: Adaptive and stable dictionary learning for concept extraction in large vision models. arXiv:2502.12892, 2025. URL https://arxiv.org/abs/2502.12892.

Shruti Joshi, Andrea Dittadi, Sébastien Lachapelle, and Dhanya Sridhar. Sparse shift autoencoders for identifying concepts from large language model activations. arXiv:2502.12179v2, 2026. URL https://arxiv.org/abs/2502.12179v2.

Richard M. Karp. Reducibility among combinatorial problems. In Raymond E. Miller, James W. Thatcher, and Jean D. Bohlinger, editors, *Complexity of Computer Computations*, pages 85–103. Plenum Press, 1972. doi: 10.1007/978-1-4684-2001-2_9. URL https://doi.org/10.1007/978-1-4684-2001-2_9.

Edward J. McShane. Extension of range of functions. *Bulletin of the American Mathematical Society*, 40(12):837–842, 1934. doi: 10.1090/S0002-9904-1934-05978-0. URL https://doi.org/10.1090/S0002-9904-1934-05978-0.

Moritz Miller, Florent Draye, and Bernhard Schölkopf. Superposition without interference? towards isolated interventions via almost orthogonal features in language models. arXiv:2602.04718v5, 2026. URL https://arxiv.org/abs/2602.04718v5.

Stefan Müller and Georg Regensburger. Elementary vectors and conformal sums in polyhedral geometry and their relevance for metabolic pathway analysis. arXiv:1512.00267v2, 2016. URL https://arxiv.org/abs/1512.00267v2.

Katta G. Murty and Santosh N. Kabadi. Some NP-complete problems in quadratic and nonlinear programming. *Mathematical Programming*, 39:117–129, 1987. doi: 10.1007/BF02592948. URL https://doi.org/10.1007/BF02592948.

Walter Nelson, Theofanis Karaletsos, and Francesco Locatello. Toward identifiable sparse autoencoders. arXiv:2605.31245, 2026. URL https://arxiv.org/abs/2605.31245.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features. arXiv:2501.16615, 2025. URL https://arxiv.org/abs/2501.16615.

Naomi Shaked-Monderer. SPN graphs: When copositive = SPN. *Linear Algebra and its Applications*, 509:82–113, 2016. doi: 10.1016/j.laa.2016.07.018. URL https://arxiv.org/abs/1604.02172.

Naomi Shaked-Monderer. Corrigendum to "SPN graphs: When copositive = SPN". arXiv:1712.05115, 2017. URL https://arxiv.org/abs/1712.05115.

Andreas M. Tillmann and Marc E. Pfetsch. The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. *IEEE Transactions on Information Theory*, 60(2):1248–1259, 2014. doi: 10.1109/TIT.2013.2290112. URL https://arxiv.org/abs/1205.2081v6.

A PROOFS OF THE SHARP TARGET-SPECIFIC MODULUS

Unless stated otherwise, D is fixed, $1 \le k \le m$, $z \ge 0$, and each code has at most k nonzero coordinates. Column normalization is needed for the explicit Gram formulas, but not for the kernel or circuit statements. A zero column can be removed because its contribution to every diagonal intervention is zero.

A.1 THE DIFFERENCE SET AND FINITE STABILITY

For $h = z - z'$ with $z, z' \in \Sigma_k^+$, every positive coordinate of h is in $\mathrm{supp}(z)$, and every negative coordinate is in $\mathrm{supp}(z')$. Thus $h \in \mathcal{H}_k^+$. Conversely, $h = h_+ - h_-$ realizes any $h \in \mathcal{H}_k^+$ as a difference of two admissible codes. This proves $\mathcal{H}_k^+ = \Sigma_k^+ - \Sigma_k^+$; the separate positive and negative budgets cannot be replaced by an arbitrary $2k$-sparsity budget without changing the model.

The equality $Dz = Dz'$ implies $DWz = DWz'$ for all such pairs if and only if

$$h \in \ker D \cap \mathcal{H}_k^+ \implies DWh = 0. \tag{21}$$

A finite value of $\kappa_k(w)$ plainly implies (21). For the reverse implication, write $\mathcal{H}_k^+$ as a finite union of closed polyhedral cones $C_{P,N}$ specified by disjoint positive and negative index sets with $|P|, |N| \le k$. For each cone, the set

$$L_{P,N} = \{(Dh, DWh) : h \in C_{P,N}\}$$

is a polyhedral cone, and hence closed: a finitely generated representation of a polyhedral cone remains finitely generated under a linear map. If no finite ratio bound holds, some fixed cone contains a sequence with $\|DWh_t\|_2 = 1$ and $Dh_t \to 0$. Passing to a subsequence, $DWh_t \to v$ with $\|v\|_2 = 1$. Closedness gives $(0, v) \in L_{P,N}$, contradicting (21). This proves the equivalences in Theorem 1.

A.2 SHARP WORST-CASE AMBIGUITY

For $z, z' \in \mathcal{Z}_\varepsilon(x)$, the difference $h = z - z'$ satisfies $\|Dh\|_2 \le 2\varepsilon$. Thus $\Delta_w(x) \le 2\varepsilon\kappa_k(w)$. Conversely, take $h \in \mathcal{H}_k^+$ with $Dh \ne 0$, write $h = h_+ - h_-$, and scale it until $\|Dh\|_2 = 2\varepsilon$. At

$$x = \tfrac{1}{2}(Dh_+ + Dh_-)$$

both codes are feasible. Their target difference is $2\varepsilon\|DWh\|_2/\|Dh\|_2$. Taking the supremum proves sharpness when the modulus is finite. If (21) fails, a nonzero target difference at identical reconstructions can be scaled arbitrarily, proving the infinite case. If the target map vanishes identically, both sides are zero. The supremum is over observations x: this is not an assertion that every individual x attains the bound, nor that every input has unbounded ambiguity when the global modulus is infinite.

A.3 THE EXACT MINIMAX ESTIMATION RADIUS

Let an estimator output an arbitrary vector, not necessarily the target of a feasible code, and define

$$R^*(\varepsilon) = \inf_{\widehat{v}} \sup_{z \in \Sigma_k^+,\ \|e\|_2 \le \varepsilon} \|\widehat{v}(Dz + e) - DWz\|_2.$$

Then $R^*(\varepsilon) = \varepsilon\kappa_k(w)$. The midpoint construction above gives the lower bound by the triangle inequality. For the upper bound, assume the modulus κ is finite and define $F(Dz) = DWz$ on $Y = D\Sigma_k^+$. The kernel condition makes F well defined, and it is κ-Lipschitz. A Euclidean Lipschitz extension, a classical consequence of Kirszbraun's theorem (Azagra et al., 2021), gives the desired estimator.

For completeness, only a one-point extension is needed here, and it has the following proof. Fix an observation x and finitely many $y_i \in Y$, with $p_i = F(y_i)$. Minimize the coercive convex function

$$\phi(q) = \max_i \{\|q - p_i\|_2^2 - \kappa^2\|x - y_i\|_2^2\}.$$

At a minimizer q_*, there are nonnegative weights λ_i supported on active indices, summing to one, such that $q_* = \sum_i \lambda_i p_i$. The variance identity gives

$$\begin{aligned}\phi(q_*) &= \tfrac{1}{2}\sum_{i,j}\lambda_i\lambda_j\|p_i - p_j\|_2^2 - \kappa^2\sum_i \lambda_i\|x - y_i\|_2^2 \\ &\le -\kappa^2\Big\|x - \sum_i \lambda_i y_i\Big\|_2^2 \le 0.\end{aligned}$$

Thus the finitely many balls $B(p_i, \kappa\|x - y_i\|_2)$ intersect. Including one fixed ball gives compactness, so the finite-intersection property yields a point common to the balls for every $y \in Y$. Select such a point as $\widehat{v}(x)$. Then $\|\widehat{v}(Dz + e) - DWz\|_2 \le \kappa\|e\|_2$. For $\kappa = 0$, the target is constant (indeed zero), and the conclusion is immediate. The estimator is an arbitrary ambient-space vector, as allowed by the stated minimax problem.

B Exact resolution through signed circuits

A circuit C is a minimally dependent set of nonzero columns. Its nullspace is one dimensional, and its nonzero dependency c has support exactly C. A circuit is k-feasible if $|\operatorname{supp}(c_+)| \le k$ and $|\operatorname{supp}(c_-)| \le k$; this does not depend on multiplying c by a nonzero scalar. View feasible circuits as hyperedges, and include isolated columns as singleton connected components.

Lemma 12 (Conformal circuit decomposition)**.** *Every $h \in \ker D \setminus \{0\}$ is a sum of circuit vectors, each supported in $\operatorname{supp}(h)$ and conformal to h: every nonzero summand coordinate has the sign of the corresponding coordinate of h.*

Proof. Choose a nonzero null vector c conformal to h with minimum support. It must be a circuit. Otherwise its support has a nullspace of dimension at least two, and there is a null vector v on that support not proportional to c. The parameters t for which $c + tv$ has the prescribed signs form an interval containing zero in its relative interior. At least one endpoint is finite. At that endpoint a coordinate vanishes while the vector remains nonzero, contradicting minimality. Thus c is a circuit.

Rescale c positively until $h - c$ remains conformal to h and at least one coordinate vanishes; explicitly use the minimum ratio $|h_i|/|c_i|$ over its support. The residual is in the nullspace and has smaller support. Induction proves the claim. □

This is a standard conformal-decomposition argument; related polyhedral treatments include Müller and Regensburger (2016). It is included to specify exactly which sparsity budgets are preserved.

Proof of Theorem 2. Suppose DWz is determined by Dz. For any feasible circuit vector c, the kernel condition gives $D_C \operatorname{diag}(w_C)c_C = 0$. Since $\ker D_C$ is one dimensional and $c_i \neq 0$ for every $i \in C$, this implies w_i is constant on C. Constancy propagates through hypergraph connectivity.

Conversely, decompose any $h \in \ker D \cap \mathcal{H}_k^+$ by Lemma 12. Every summand is a feasible circuit vector, because its positive and negative supports are contained in those of h. If w is constant on every feasible circuit, each summand c satisfies $DWc = 0$. Summing proves the kernel condition. This argument applies to arbitrary real diagonal weights, not only $[0, 1]$ weights. □

Consequently, exact identifiable binary groups are precisely unions of hypergraph components. If there are r components, the vector space of all exact identifiable diagonal weights has dimension r. For a requested target T, its least exact identifiable superset is the union of the components meeting T. Merging these components further discards independent intervention degrees of freedom; it is not required for exact identifiability.

Low pairwise coherence need not identify the resolution. Let $v_1, \ldots, v_{2k}$ be regular-simplex unit vectors in $\mathbb{R}^{2k-1}$, so $\sum_i v_i = 0$ and $v_i^\top v_j = -1/(2k-1)$ for $i \neq j$. Choose signs s_i with k of each sign and set $d_i = s_i v_i$. Then $\sum_i s_i d_i = 0$ is a feasible circuit joining every column, although every absolute pairwise correlation is only $1/(2k-1)$. Thus a pairwise-threshold heuristic cannot replace feasible circuits for arbitrary k.

C FOREST CERTIFICATION: PROOF, RECURRENCE, AND WITNESSES

C.1 QUADRATIC REDUCTION AND BALANCED SUBTREES

Write $\Lambda = K^2 > 1$. Expanding squared norms gives

$$\|DWh\|_2^2 \le \Lambda \|Dh\|_2^2 \iff h^\top(\Lambda\Gamma - W\Gamma W)h \ge 0.$$

The graph consists of exactly the nonzero off-diagonal entries of Γ. On a forest choose signs recursively so $\sigma_i\sigma_j\Gamma_{ij} < 0$ on every edge. Switching by $S = \mathrm{diag}(\sigma)$ yields a matrix $Z = SM_K(w)S$ with diagonal

$$b_v = \Lambda - w_v^2 > 0$$

and edge entries $-a_{vu}$, where $a_{vu} = |\Gamma_{vu}|(\Lambda - w_v w_u) > 0$.

Here is an expanded justification of Theorem 4. If a balanced connected principal submatrix is not PSD, it has a negative vector. Taking coordinatewise absolute values in the switched matrix does not increase its quadratic form. Switching back gives an admissible negative direction with the required two sign budgets. Conversely, start with any admissible h and cut support edges with a positive quadratic cross term. Each remaining component has signs agreeing with σ or $-\sigma$, so it is balanced. Importantly, it is an induced subtree: an edge cut within the same remaining component would form a cycle. Thus the restriction of the original quadratic form to that component is exactly the relevant principal submatrix. All these restrictions are nonnegative by hypothesis, and all removed edge contributions are nonnegative. Their sum proves the full inequality. This is the point where the forest assumption is essential.

C.2 ROOTED SCHUR-COMPLEMENT DYNAMIC PROGRAMMING

Root each tree. A state at v describes a connected subtree containing v and no ancestors of v. Its budget (p, q) counts the vertices of the two fixed signs σ. A singleton starts with value b_v and its one-vertex budget. Each child is either omitted, or contributes a connected subtree containing that child. Before a negative witness is found, any child subtree that can safely be attached has a positive scalar root Schur complement d_u.

Eliminating all selected child subtrees leaves at the root

$$d_v = b_v - \sum_{u \text{ selected}} \frac{a_{vu}^2}{d_u}. \tag{22}$$

For every budget, store the minimum achievable d_v and backpointers to its choices. This compression is exact: for fixed child budget the term a^2/d decreases in $d > 0$, so the smallest child Schur complement is always the most dangerous choice. A less dangerous state with identical counts can never be needed to detect a failure at an ancestor. Choosing no vertices from a child is always available.

Process children by a two-budget knapsack convolution. If the root or a child intermediate value is negative, stop and recover its witness. A zero child is handled separately below. If a zero root is obtained after some children, it may remain as a singular state when subsequent children are omitted; adding any nonzero child within budget yields a negative state. States with either count exceeding k are discarded. Every connected subtree has a highest vertex, so these tables cover all balanced connected subtrees, including those not containing the root of the entire tree.

C.3 SINGULAR STATES ARE NOT NUMERICAL TOLERANCES

For a represented subtree whose descendant pivots are positive, completing squares expresses its quadratic form as positive terms plus $d_v t_v^2$. When $d_v < 0$, set $t_v = 1$ and set all eliminated coordinates to their minimizing values. Recursively, if the parent value is t_v, the child-root value is $a_{vu}t_v/d_u$. All values are positive and rational for rational inputs, and the quadratic form equals $d_v < 0$.

When $d_v = 0$, the same construction gives a positive rational null vector r with $r_v = 1$. If the budget permits attaching its parent p across an edge of magnitude a, use the vector r on the subtree, the value a/b_p at p, and zero elsewhere. Its quadratic form is

$$b_p(a/b_p)^2 - 2a(a/b_p) = -a^2/b_p < 0.$$

Algorithm 1 Exact forest certification (summary)

Require: Rational normalized PSD Gram matrix Γ with forest graph, $w \in [0,1]^m$, $\Lambda > 1$, $k \geq 1$

Root each tree and find switching signs σ
for v in postorder **do**
  Initialize table at v with the singleton value $\Lambda - w_v^2$
  for each child u **do**
    Keep all states omitting u
    for each pair of states whose combined counts are at most (k, k) **do**
      if the child root value is zero **then**
        return a rational negative witness from the singular child and v
      end if
      Form $d_{\text{new}} = d_{\text{old}} - a_{vu}^2/d_u$
      if $d_{\text{new}} < 0$ **then**
        return its rational negative witness
      end if
      Retain the smallest new value at its count pair, with backpointers
    end for
    Replace the current table by the merged table
  end for
end for
return certified

No other already chosen child is needed for this witness, so its counts remain within the tested budget. This proves that a singular child cannot be divided by, silently discarded, or regularized with a small positive constant. A singular state without an admissible extension is genuinely safe: PSD permits equality.

Switching a negative vector back by σ preserves rationality and gives at most k positive and k negative coordinates. Its sign counts and exact quadratic form can be checked independently without re-running the optimization. Writing $h = h_+ - h_-$ converts it into two admissible explanations; the midpoint construction in Appendix A produces a shared observation at any positive prescribed noise tolerance. This is an all-input counterexample, not necessarily a counterexample at a preselected observed x.

C.4 Operation counts and numerical semantics

There are at most $(k+1)^2$ states per table. One child convolution compares $O(k^4)$ pairs, and a forest has fewer than m child edges. Thus the recurrence takes $O(mk^4)$ arithmetic operations and polynomial state storage, proving Theorem 5. This count starts from a supplied valid Gram matrix and its graph; constructing a dense Gram matrix, checking its PSD property, reading a dense input, and rational bit lengths are separate costs. The reference implementation validates inputs by default and uses arbitrary-precision fractions. A floating implementation that thresholds edges or replaces exact signs by tolerances is not the same certificate.

Polynomial encoding length. The arithmetic bound also yields a polynomial-time bit algorithm for rational input, although it does not give the same numerical bound in bit operations. Let L be the total binary encoding length of Γ, w, Λ and k. For a stored state with selected subtree H and root v, the descendant block is positive definite until the algorithm stops or handles a zero child. Its root Schur complement satisfies

$$d_v = \frac{\det M_K(w)[H]}{\det M_K(w)[H \setminus \{v\}]},$$

with empty determinant equal to one. Sign switching leaves these determinants unchanged. Both orders are at most $2k$, and all entries have bit length polynomial in L. Clearing denominators and applying the determinant expansion or Hadamard's bound shows that each numerator and denominator has polynomial bit length in L and k. The same argument covers a negative candidate state; a zero state has exact encoding zero.

The recurrence stores a realized minimum, not a sum over exponentially many candidate subtrees, so taking minima introduces no new encoding growth. Intermediate child merges represent actual selected subtrees and obey the same determinant bound. Witness back-substitution multiplies and divides polynomial-bit quantities along at most $2k$ vertices. Singular-child witnesses use the same null-vector reconstruction and one additional rational coordinate. Consequently all arithmetic, comparisons, storage, and output have polynomial bit cost. Rational Gram construction and PSD validation are also polynomial-time preprocessing, but are not included in the $O(mk^4)$ recurrence count.

A separating constraint. For completeness, a rejected candidate also separates the convex feasible weight set. Let h be a witness at w_0, set $v_0 = D\,\mathrm{diag}(h)w_0 \neq 0$, and define $g = \mathrm{diag}(h)D^\top v_0/\|v_0\|_2$. Every feasible w obeys

$$g^\top w \leq \|D\,\mathrm{diag}(h)w\|_2 \leq K\|Dh\|_2,$$

whereas $g^\top w_0 = \|v_0\|_2 > K\|Dh\|_2$. This is a mathematical separation oracle under the stated structure.

C.5 Rank and circuit structure of an exact forest

Proposition 13 (Rank restriction of a forest Gram graph)**.** *A unit-column dictionary with a forest Gram graph satisfies* $\mathrm{rank}(D) \geq m/2$*. If its graph is connected,* $\mathrm{rank}(D) \geq m - 1$*. Every singular connected Gram block has a one-dimensional kernel with full support; its only circuit is the entire block.*

Proof. For a connected block H, switch signs so $A = S\Gamma[H]S$ has nonpositive off-diagonal entries. If $|H| \geq 2$, $B = I - A$ is nonnegative and irreducible, and $A \succeq 0$ implies $\rho(B) \leq 1$. By the Perron–Frobenius theorem, if A is singular then $1 = \rho(B)$ is a simple eigenvalue with a strictly positive eigenvector. Equivalently, the smallest eigenvalue of an irreducible symmetric Z-matrix is simple; see Shaked-Monderer (2016, Lemma 2.8). Thus $\mathrm{rank}\,\Gamma[H] \geq |H| - 1$. Its kernel vector, if present, has no zero coordinate, so no proper subset of H is dependent. A singleton has Gram matrix $[1]$ and rank one. Summing the component ranks and using $|H| - 1 \geq |H|/2$ for $|H| \geq 2$ proves the global bound; the connected bound follows directly. □

This also specializes the exact circuit resolution. A singular component is one identifiable unit if the two signs of its null vector each occur at most k times; otherwise that circuit is not k-feasible and each vertex is an isolated exact intervention unit. Different Gram components are orthogonal, so no circuit crosses components. Replicating each prototype after this analysis produces raw Gram cliques and arbitrarily many nonidentifiable code coordinates while preserving the quotient guarantees of Theorem 3.

D Path selection and low-sparsity semidefinite formulations

D.1 Exact path windows and binary optimization

On a positive-correlation path, switching signs alternate. A connected interval has counts differing by at most one and is balanced exactly when its length is at most $2k$. If $L = \min(2k, m)$, every shorter interval is contained in a length-L interval, and PSD passes to principal submatrices. This proves the window characterization in Theorem 6.

For a binary mask scanned left to right, every new length-L window is checked when its final bit is appended. A state records the last $L - 1$ bits and the least prefix cost producing them. Appending a bit prescribed by T is mandatory; other bits allow either choice. Two prefixes with the same suffix have the same possible future completions and future window constraints, so retaining the less costly one is exact. At the end, a minimum-cost surviving state and its backpointers recover the optimum over all masks containing T. For $m < L$ no special case arises because L was capped at m; for $L = m$ the single final window tests the whole mask. The all-one mask is feasible since $M_K(\mathbf{1}) = (K^2 - 1)\Gamma \succeq 0$.

At most 2^{L-1} suffix states and two extensions are examined per position. A dense exact PSD test on each L-window uses $O(L^3)$ arithmetic operations, giving $O(m2^{2k}k^3)$. The reference implementation uses backpointers rather than copying full prefixes. Costs are additive selection costs; a nonlinear norm of collateral activation need not decompose in this way.

D.2 Continuous path formulation

Let $\Gamma[I] = R_I^\top R_I$. The Schur complement of the identity block in

$$\begin{pmatrix} K^2\Gamma[I] & W_I R_I^\top \\ R_I W_I & I \end{pmatrix}$$

is $K^2\Gamma[I] - W_I\Gamma[I]W_I$. Thus the displayed LMI in the main text is equivalent to the exact window condition, including singular Gram matrices. The block entries are affine in w. Any convex objective over these constraints, $0 \le w \le 1$, and $w_T = \mathbf{1}$ is therefore a convex optimization problem, and a linear objective gives an SDP. This is exactness of the formulation; solving irrational optima exactly or establishing validated rounding is a separate numerical task.

D.3 Arbitrary dictionaries at $k = 2$

A symmetric matrix A is copositive when $u^\top Au \ge 0$ for every $u \ge 0$. For each pair of disjoint supports P, N with $|P|, |N| \le 2$, put $J = P \cup N$, let S_J put $+1$ on P and -1 on N, and set

$$B_J = D_J S_J, \qquad Q_J = B_J^\top B_J.$$

All differences in that orthant are $h_J = S_J u$, $u \ge 0$. Consequently exact stability is equivalent to

$$K^2 Q_J - W_J Q_J W_J \in \mathcal{COP}_{|J|}$$

for every such pair. The classical identity $\mathcal{COP}_r = \mathbb{S}^r_+ + \mathbb{N}^r$ for $r \le 4$ states that every such matrix is PSD plus a symmetric entrywise nonnegative matrix; see Theorem 5.1 of Shaked-Monderer (2016). Using this imported identity and a Schur complement, the condition is equivalent to existence of a symmetric entrywise nonnegative N_J with

$$\begin{pmatrix} K^2 Q_J - N_J & W_J B_J^\top \\ B_J W_J & I \end{pmatrix} \succeq 0. \tag{23}$$

There are $O(m^4)$ support–sign pairs, and all entries are affine in the unknown weights and N_J. This proves the exact $k = 2$ SDP formulation without a forest assumption. The identity for $r \le 4$ is a cited classical theorem, not an original result or an assumption about higher-dimensional copositive cones.

Scope of the cited copositive result. Shaked-Monderer (2017) corrects a different statement about graphs formed by multiple triangles sharing an edge. We do not use that statement. The forest proof here is self-contained, and the only cone identity imported above is the order-at-most-four identity. No conclusion about the corrected multi-triangle family is required.

E Optimal single-activation attenuation

Consider $k = 1$, $K > 1$, and $0 \le w \le 1$. Differences on only one coordinate satisfy the desired inequality automatically. Two distinct coordinates have the form $h = \alpha e_i - \beta e_j$ up to overall sign, with $\alpha, \beta \ge 0$. Expanding gives

$$(K^2 - w_i^2)\alpha^2 + (K^2 - w_j^2)\beta^2 - 2\rho_{ij}(K^2 - w_i w_j)\alpha\beta \ge 0. \tag{24}$$

For $\rho_{ij} \le 0$, every term is nonnegative. For $\rho_{ij} > 0$, the diagonal coefficients are strictly positive; minimizing over $\alpha/\beta \ge 0$ shows that (24) is equivalent to

$$(K^2 - w_i^2)(K^2 - w_j^2) \ge \rho_{ij}^2 (K^2 - w_i w_j)^2.$$

Set $u_i = w_i/K \in [0, 1/K]$. The identity $(1 - u_i u_j)^2 - (u_i - u_j)^2 = (1 - u_i^2)(1 - u_j^2)$ transforms the condition to

$$\left| \frac{u_i - u_j}{1 - u_i u_j} \right| \le \sqrt{1 - \rho_{ij}^2}.$$

Since $u_i = \tanh(s_i)$ and $\tanh$ is monotone on nonnegative arguments, this is exactly $|s_i - s_j| \le \ell_{ij}$. Duplicate columns have $\rho = 1$, forcing $s_i = s_j$; antiparallel columns have $\rho = -1$ and impose no constraint. This asymmetry is a consequence of nonnegative codes, not an omitted absolute value.

On the graph with these edge lengths, any feasible s has $s_t = a_K$ for $t \in T$. Summing edge constraints along a path gives $s_j \ge a_K - \mathrm{dist}(T, j)$, and nonnegativity gives the positive-part lower bound. Conversely, the distance-to-set function changes by at most the length of each edge, and $r \mapsto [a_K - r]_+$ is 1-Lipschitz. Thus $s_j^* = [a_K - \mathrm{dist}(T, j)]_+$ is feasible and coordinatewise minimal. Mapping back monotonically proves Theorem 7. For an empty target the minimum is the zero vector. Vertices disconnected from the target also receive zero.

For binary weights, crossing a kept/deleted pair changes s by exactly a_K. Edges of length less than a_K cannot be cut; all other edges may be cut. Since

$$\ell_{ij} < a_K \iff \rho_{ij} > \sqrt{1 - K^{-2}},$$

components of this strict-threshold graph are exactly the minimum hard intervention units. Equality at the threshold permits a cut. More generally, a nonempty binary group has

$$\kappa_1(\mathbf{1}_G) = \left(1 - \rho_G^2\right)^{-1/2}, \qquad \rho_G = \max_{i \in G,\, j \notin G} (d_i^\top d_j)_+,$$

with an empty maximum equal to zero. An empty group has modulus zero. The formula follows by fixing the inside magnitude α and minimizing the denominator over $\beta \ge 0$; the optimum is $\beta = \alpha(\rho_{ij})_+$.

E.1 Hard versus continuous collateral

For the path family $d_i = (e_i + e_{i+1})/\sqrt{2}$, each edge length is $\ell = \operatorname{arctanh}(\sqrt{3}/2)$. With target $\{1\}$ the solution is

$$w_j^* = K \tanh([a_K - (j-1)\ell]_+).$$

At $K = 1.1$, $\ell < a_K < 2\ell$, giving the numerical example in the main text. Its rounded vector $(1, 223/1000, 0, \ldots, 0)$ also passes the independent exact-rational forest certificate at $K^2 = 121/100$; the transcendental evaluation alone is not treated as an exact certificate.

For a one-sparse code $z = \alpha e_j$ with unit decoder columns, the ratio of collateral norm to original reconstruction norm is w_j outside T and zero inside T. Taking the supremum gives $\max_{j \notin T} w_j$. As K increases toward $2/\sqrt{3}$ from below, a_K decreases toward ℓ, so the continuous solution's collateral tends to zero, while the threshold graph remains connected and every binary solution has collateral one. This proves an unbounded ratio separation in actual worst-case relative collateral. It is a $k = 1$ statement; for multiple active columns, cancellations prevent replacing a collateral norm by a coordinatewise monotone cost in general.

E.2 An unbounded separation at every sparsity

Proof of Corollary 8. Fix integers $k \ge 1$, $r \ge 2$, and choose $q \ge k$. Let $n = 2q$ and

$$u_{2g-1} = e_{2g-1}, \qquad u_{2g} = \frac{3}{5} e_{2g-1} + \frac{4}{5} e_{2g}, \qquad 1 \le g \le q.$$

These columns are rational and unit norm. Distinct pairs occupy orthogonal coordinate planes, and each within-pair correlation is $\rho = 3/5$. Form D by repeating every prototype r times, so $m = 2qr$ and $m/n = r$. Let T contain all copies of the odd prototypes. Different representatives of any direction already give distinct one-sparse nonnegative codes for the same activation. An aggregate with k active prototype directions has at least r^k different k-sparse representative codes, so ambiguity is not confined to a single active feature.

For a finite stability budget, Theorem 3 forces equal weights on copies of the same direction. Suppose a binary mask keeps an even direction while deleting its odd partner. The quotient difference $h = e_{2g-1} - (3/5)e_{2g}$ has one positive and one negative coordinate, and

$$\|Uh\|_2^2 = \frac{16}{25}, \qquad \|U \operatorname{diag}(v) h\|_2^2 = 1.$$

It lifts to an admissible raw difference for every $k \geq 1$. Its amplification is $5/4$, so for $1 < K < 5/4$ every stable binary mask must also delete all even directions. Thus its only possible mask is full deletion.

For continuous weights put $v_{2g-1} = 1$ and $v_{2g} = c_K$ from (14). The two-by-two stability matrix for each pair is

$$\begin{pmatrix} K^2 - 1 & (3/5)(K^2 - c_K) \\ (3/5)(K^2 - c_K) & K^2 - c_K^2 \end{pmatrix}.$$

Its diagonal is positive and its determinant is zero, as follows either by substitution or the pairwise hyperbolic identity with $\sqrt{1 - \rho^2} = 4/5$. Therefore the full block-diagonal matrix is PSD. This proves stability for every real quotient difference and hence every raw k-sparse pair. Conversely, each individual pair is admissible for every k, so its one-pair lower bound forces $v_{2g} \geq c_K$ for any stable intervention. This proves coordinatewise minimality, not just feasibility.

For an aggregate $a \geq 0$, the squared reconstruction norm is

$$\|Ua\|_2^2 = \sum_{g=1}^{q} \left(a_{2g-1}^2 + a_{2g}^2 + \frac{6}{5} a_{2g-1} a_{2g} \right) \geq \sum_{g=1}^{q} a_{2g}^2.$$

The squared collateral is $c_K^2 \sum_g a_{2g}^2$, so its relative norm is at most c_K. Equality is attained by a code supported on any even direction, or on any k even directions since $q \geq k$. Full binary deletion has relative collateral one by the same calculation. Finally, $c_K > 0$ for $1 < K < 5/4$ and $c_K \to 0$ as $K \uparrow 5/4$. Choose a rational K sufficiently close to $5/4$ that $c_K < \eta$; then the weights are rational as well. At $K = 6/5$, direct substitution gives $c_K = 3/25$. □

The construction uses isolated interacting pairs, not arbitrary multi-activation dictionaries. It extends the *separation phenomenon* to all k, without extending the arbitrary-dictionary shortest-path formula beyond $k = 1$.

F Proof of the computational classification

We give all steps of Theorem 9, including a rational dictionary rather than merely a PSD Gram matrix. CLIQUE is the standard NP-complete decision problem (Karp, 1972). Restrict to $N \geq 2$ and $2 \leq k \leq N$, which retains the hardness of the unrestricted input family.

Let A be the adjacency matrix, $\ell = \lceil \sqrt{2N} \rceil$, $t = \ell^{-2}$, and $C = I - tA$. Construct Γ as in (15). Every absolute off-diagonal row sum is at most $Nt \leq 1/2$, so Gershgorin's inequality gives $\frac{1}{2} I \preceq \Gamma \preceq \frac{3}{2} I$. In particular all denominators below are strictly positive.

F.1 Evaluating the intervention modulus

Deleting column 0 makes the numerator of the squared ratio h_0^2. Directions with $h_0 = 0$ contribute zero; all others can be rescaled and globally sign-reversed to set $h_0 = 1$. Write the remaining coordinates as $p - q$ with disjoint nonnegative supports. The positive support of p has size at most $k - 1$, and the support of q at most k. The denominator is

$$1 + 2t\mathbf{1}^\top (p - q) + (p - q)^\top C (p - q).$$

Removing p changes it by the nonnegative amount

$$2t\mathbf{1}^\top p + p^\top C p - 2p^\top C q \geq 0.$$

Indeed $C \succ 0$, its off-diagonal entries are nonpositive, and the supports are disjoint. Removing p preserves admissibility, so the optimum has $p = 0$.

For fixed negative support S, minimize $1 - 2t\mathbf{1}^\top q + q^\top C_S q$ over $q \geq 0$. Since $t\|A_S\|_{\mathrm{op}} < 1$, the Neumann series yields $C_S^{-1} \geq 0$ entrywise. The unconstrained minimizer $q = tC_S^{-1}\mathbf{1}$ is thus feasible and gives the minimum $1 - r(S)$, where $r(S) = t^2 \mathbf{1}^\top C_S^{-1} \mathbf{1}$. This proves (16). Empty support gives $r(\varnothing) = 0$.

F.2 GAP FOR A MISSING CLIQUE

For a k-clique, $A_S\mathbf{1} = (k-1)\mathbf{1}$, so

$$r(S) = r_* := \frac{t^2 k}{1 - t(k-1)}.$$

For any k-vertex nonclique, pad or index its adjacency in the same dimension as that of the complete graph. Entrywise domination of nonnegative matrices implies domination of every matrix power. In

$$r(S) = t^2 \sum_{j \geq 0} t^j \mathbf{1}^\top A_S^j \mathbf{1},$$

all terms are at most those of the complete graph, and the $j = 1$ term loses at least $2t^3$ because an undirected edge is missing. For $|S| < k$, padding by isolated zero coordinates for the higher powers gives the same domination, while the $j = 0$ term alone loses at least $t^2 \geq 2t^3$. Hence if the graph has no k-clique, every $|S| \leq k$ satisfies $r(S) \leq r_* - 2t^3$.

The rational budget

$$\Lambda = K^2 = \frac{1}{1 - (r_* - t^3)}$$

has polynomial encoding length and exceeds one. In the clique case the intervention is unsafe; in the no-clique case it is strictly safe. Thus the complement of CLIQUE reduces to safe certification, proving coNP-hardness. Membership in coNP is proved in Appendix F.4. The separating gap is inverse polynomial, of order t^3, and does not establish constant-factor inapproximability.

F.3 RATIONAL UNIT-COLUMN REALIZATION

For every nonzero off-diagonal entry $\Gamma_{ij} = \pm t$, allocate a fresh row with $1/\ell$ in column i, $\operatorname{sign}(\Gamma_{ij})/\ell$ in column j, and zeros elsewhere. If a column participates in r_j such rows, add $\ell^2 - r_j$ private rows, each with $1/\ell$ in that column only. Since $r_j \leq N \leq \ell^2$, this is possible. Each column has squared norm $\ell^2/\ell^2 = 1$, each desired pairwise inner product is exactly $\pm t$, and all other inner products are zero. The number of rows is $O(N^2)$, and every nonzero entry is rational with polynomial encoding length. This realizes (15) by an explicit dictionary and completes the hardness reduction.

F.4 POLYNOMIAL-SIZE RATIONAL NEGATIVE WITNESSES

The membership direction follows from the same rational quadratic-programming structure underlying classical copositivity complexity (Murty and Kabadi, 1987). We give the argument to retain the two separate sparsity budgets.

Lemma 14 (Short witnesses on a nonnegative orthant)**.** *If a rational symmetric matrix A has $u^\top A u < 0$ for some $u \geq 0$, it has such a nonnegative rational vector whose binary encoding length is polynomial in the encoding length and order of A.*

Proof. By homogeneity, minimize $u^\top A u$ on the simplex $u \geq 0$, $\mathbf{1}^\top u = 1$. A minimizer u^* exists and has value $\lambda^* < 0$. Let $J = \operatorname{supp}(u^*)$. The minimizer lies in the relative interior of the simplex on J, so its equality-constrained stationary equations are

$$A[J]u_J^* = \lambda^* \mathbf{1}, \qquad \mathbf{1}^\top u_J^* = 1.$$

Consider the nonempty rational polyhedron

$$\mathcal{P} = \{(v, \lambda) : v \geq 0,\ \mathbf{1}^\top v = 1,\ A[J]v = \lambda \mathbf{1}\}.$$

For any two feasible points (v, λ) and (y, μ), symmetry gives $\mu = v^\top A[J] y = y^\top A[J] v = \lambda$. Thus all feasible points have the same scalar $\lambda = \lambda^*$. The polyhedron is bounded: v is in a simplex and λ is fixed. It has a rational vertex obtained by solving a full-rank subsystem of its rational equalities and active nonnegativity constraints. Clearing denominators and applying Cramer's rule and determinant bounds yields polynomial encoding length for that vertex $(\widehat{v}, \lambda^*)$. Moreover, $\widehat{v}^\top A[J]\widehat{v} = \lambda^* < 0$. Extend $\widehat{v}$ by zero outside J to obtain the required witness. □

Now suppose a rational instance D, w, K^2 violates stability. Select the disjoint positive and negative supports P, N of a violating $h \in \mathcal{H}_k^+$, and put $J = P \cup N$. Let S_J have sign $+1$ on P and -1 on N. The rational signed matrix

$$A = S_J\big(K^2 D_J^\top D_J - W_J D_J^\top D_J W_J\big)S_J$$

is not copositive. Lemma 14 supplies a polynomial-size nonnegative rational u with negative quadratic form. Taking $h_J = S_J u$ and zero elsewhere preserves at most k positive and k negative entries. The matrix A has polynomial encoding length in the original rational data, so the same is true of h.

A verifier checks the two support counts and the strict rational inequality $K^2\|Dh\|_2^2 - \|DWh\|_2^2 < 0$ in polynomial time. Hence unsafe certification belongs to NP, and safe certification belongs to coNP, including the single-deletion and well-conditioned restriction. Combined with the reduction above this proves Theorem 9. The argument asserts existence of short negative witnesses, not an efficient procedure for finding them on arbitrary dictionaries.

G Redundant groups, geometric certificates, and approximate structure

G.1 Exact duplicate quotient and optimal selection

Proof of Theorem 3. For i, j in the same group, $D(e_i - e_j) = 0$ and $e_i - e_j \in \mathcal{H}_1^+ \subseteq \mathcal{H}_k^+$. A finite modulus forces $DW(e_i - e_j) = (w_i - w_j)u_{g(i)} = 0$. Since the prototype is nonzero, $w_i = w_j$. Thus group-constancy is necessary even when it was not imposed as a design restriction.

For group-constant weights, define $a_g = \sum_{j \in G_g} z_j$. Then $a \geq 0$, $\|a\|_0 \leq s = \min(k, p)$, $Dz = Ua$, and $DWz = U\,\mathrm{diag}(v)a$. Conversely every such aggregate a is realized by assigning its mass to one representative in each active group. In particular,

$$\{(Dz, DWz) : z \in \Sigma_k^+(\mathbb{R}^m)\} = \{(Ua, U\,\mathrm{diag}(v)a) : a \in \Sigma_s^+(\mathbb{R}^p)\}.$$

This equality of attainable pairs gives equality of their differences, hence the moduli. Restricting to reconstructions within ε of any fixed x gives identical feasible target sets, hence identical ambiguity and information-theoretic estimation problems.

A finite-budget intervention with $w_T = \mathbf{1}$ must have $v_g = 1$ for every $g \in \bar{T}$. Its additive collateral objective transforms exactly as

$$\sum_{j \notin T} c_j w_j = \sum_{g=1}^{p} \left(\sum_{j \in G_g \setminus T} c_j \right) v_g = \sum_{g=1}^{p} c'_g v_g.$$

Conversely every feasible prototype weight vector lifts to a feasible raw intervention with the same objective. The feasible optimization problems are therefore identical, including binary constraints. If a target meets but does not contain a whole group, the cost c'_g on that group is an unavoidable fixed contribution, not a term that may be silently omitted. If prototypes themselves share further admissible circuits, those constraints remain present in the quotient modulus. □

For a target made of complete groups, actual collateral target sets also commute with aggregation. If each prototype has at least two copies, no individual active-copy index is uniquely recoverable, even when the prototype matrix has full column rank. Reading and grouping raw columns still costs time; only the subsequent structured certification and optimization depend on the quotient size rather than multiplicity.

G.2 Subspace separation without within-group coefficient recovery

For a partition $G_1, \ldots, G_p$, let U_g be an orthonormal basis of $\mathrm{span}(D_{G_g})$, and let U_J concatenate these bases. Put $q = \min(2k, p)$ and

$$\alpha_q^2 = \min_{1 \leq |J| \leq q} \lambda_{\min}(U_J^\top U_J).$$

Assume $\alpha_q > 0$. For $h = z - z' \in \mathcal{H}_k^+$, at most q groups are active. Set $v_g = D_{G_g} h_{G_g} = U_g a_g$ on those groups. Then

$$\|Dh\|_2^2 = \Big\| \sum_g U_g a_g \Big\|_2^2 \geq \alpha_q^2 \sum_g \|a_g\|_2^2 = \alpha_q^2 \sum_g \|v_g\|_2^2.$$

Since $\|Dh\|_2 \leq 2\varepsilon$ for a feasible pair, this proves the simultaneous bound

$$\Big(\sum_g \|D_{G_g}(z - z')_{G_g}\|_2^2 \Big)^{1/2} \leq \frac{2\varepsilon}{\alpha_q}. \tag{25}$$

No D_{G_g} is required to have independent columns. The bases represent synthesized contributions, not within-group coefficients. These are target-specific uses of familiar block-sparse subspace geometry (Eldar et al., 2010).

A polynomially checkable sufficient bound uses $C_{gh} = \|U_g^\top U_h\|_{\mathrm{op}}$ for $g \neq h$. Let β_q be the largest, over rows, sum of their largest $q-1$ off-diagonal entries. Expanding a norm and applying $2ab \leq a^2 + b^2$ gives

$$\Big\| \sum_{g \in J} U_g a_g \Big\|_2^2 \geq \sum_{g \in J} \Big(1 - \sum_{h \in J \setminus g} C_{gh} \Big) \|a_g\|_2^2 \geq (1 - \beta_q) \sum_{g \in J} \|a_g\|_2^2.$$

Thus $\alpha_q^2 \geq 1 - \beta_q$. A nonpositive lower bound is inconclusive, not a counterexample. For two unit directions separated by an acute angle θ, deleting the first in the $k = 1$ model has exactly $\kappa = 1/\sin\theta$: the competing codes have magnitudes a and $a\cos\theta$. Consequently geometric noise amplification cannot in general be eliminated by a different proof.

G.3 Approximate groups and a uniform amplitude certificate

Suppose $\widetilde{d}_j \in \mathrm{span}(U_{g(j)})$ and $\|d_j - \widetilde{d}_j\|_2 \leq \delta$. If all feasible codes obey $\|z\|_1 \leq B$, then for $h = z - z'$,

$$\|\widetilde{D}h\|_2 \leq 2\varepsilon + 2\delta B, \qquad \|(D - \widetilde{D})_{G_g} h_{G_g}\|_2 \leq 2\delta B.$$

Applying the same subspace argument to $\widetilde{D}h$ yields

$$\Delta_{G_g}(x) \leq \frac{2(\varepsilon + \delta B)}{\alpha_q} + 2\delta B. \tag{26}$$

The more general transfer in Proposition 10 follows from

$$\|DWh\|_2 \leq \|\bar{D}Wh\|_2 + \|EWh\|_2 \leq K(2\varepsilon + 2\delta B) + 2\delta B,$$

using $\|Wh\|_1 \leq \|h\|_1 \leq 2B$. Neither assertion follows by merely thresholding entries of a Gram matrix.

In a near-collinear nonnegative model, B can be derived rather than imposed. Suppose unit prototypes obey $\|d_j - u_{g(j)}\|_2 \leq \delta$. Let $s = \min(k, p)$ and let α_s be the restricted smallest singular value for these prototype columns. For $a_g = \sum_{j \in G_g} z_j$, we have $\|a\|_0 \leq s$ and $\|a\|_1 = \|z\|_1$. Therefore

$$\|Dz\|_2 \geq \alpha_s \|a\|_2 - \delta \|z\|_1 \geq \big(\alpha_s / \sqrt{s} - \delta \big) \|z\|_1.$$

If $\gamma = \alpha_s/\sqrt{s} - \delta > 0$, then the entire feasible set satisfies

$$\|z\|_1 \leq B(x) := \frac{\|x\|_2 + \varepsilon}{\gamma}.$$

This uses nonnegativity essentially; group coefficients with opposing signs could cancel before aggregation.

Nonnegative prototype Gram and collateral. Theorem 11 avoids restricted singular values when $U^\top U$ is entrywise nonnegative. For an aggregate $\alpha \geq 0$ on at most s groups, expanding its squared norm yields $\|U\alpha\|_2^2 \geq \|\alpha\|_2^2 \geq \|\alpha\|_1^2/s$. This argument remains valid if U is singular: it only concerns nonnegative combinations, not arbitrary signed differences. If a valid uniform amplitude bound is available from either the cone argument or a dual vector, then for $0 \leq b_g \leq c$,

$$\|U \operatorname{diag}(b)\alpha\|_2^2 = \sum_{g,h} b_g b_h \alpha_g \alpha_h (u_g^\top u_h) \leq c^2 \|U\alpha\|_2^2.$$

For a target made of complete groups, $b_g = v_g - \mathbf{1}_{\{G_g \subseteq T\}}$ are precisely its extra weights. The error contribution is at most $\delta c\|z\|_1$, giving (19). This inequality controls the actual activation-space norm rather than an additive surrogate.

For signed prototype correlations with $(u_g^\top u_h) \geq -\nu$, $g \neq h$, the mass bound has the extension

$$\|U\alpha\|_2^2 \geq (1+\nu)\|\alpha\|_2^2 - \nu\|\alpha\|_1^2 \geq \frac{1-(s-1)\nu}{s}\|\alpha\|_1^2.$$

When $1-(s-1)\nu > 0$, its square root replaces $s^{-1/2}$ in the amplitude argument. The collateral comparison above, however, specifically uses entrywise nonnegative Gram entries and is not asserted under this signed extension.

G.4 A DATA-DRIVEN GROUPING GUARANTEE

Assume unknown unit prototypes have pairwise distances at least $a > 4\delta_0$ and every dictionary column is within δ_0 of its prototype. In a distance graph with threshold $t \in [2\delta_0, a - 2\delta_0)$, within-group pairs are adjacent by the triangle inequality, while between-group pairs are not. The graph components recover exactly the prototype groups. This is a sufficient separated-prototype model, not a claim of recovery on arbitrary correlated dictionaries.

Choose an observed column from each recovered group as its estimated prototype $\widehat{u}_g$. Every column is then within $2\delta_0$ of its group's estimate. For any subset J of at most q prototypes,

$$\|\widehat{U}_J - U_J\|_{\mathrm{op}} \leq \|\widehat{U}_J - U_J\|_F \leq \sqrt{q}\,\delta_0,$$

so $\widehat{\alpha}_q \geq \alpha_q - \sqrt{q}\,\delta_0$. The observed prototypes and observed approximation errors can consequently be inserted into the preceding certificates. The signs and positivity of their computed denominators must still be checked. If the correct threshold is unknown, a full single-linkage hierarchy contains the separating cut when these assumptions hold; certification chooses among its cuts rather than declaring similarity itself sufficient.

G.5 THE AMPLITUDE-DEPENDENT ERROR IS NECESSARY

Take orthonormal u, v and

$$d_1 = \frac{u+tv}{\sqrt{1+t^2}}, \quad d_2 = \frac{u-tv}{\sqrt{1+t^2}}, \quad d_3 = v, \qquad t > 0.$$

The first two columns are near the direction u, and the third equals v. Nevertheless at $x = Ad_1$, both

$$z = (A, 0, 0), \qquad z' = (0, A, 2At/\sqrt{1+t^2})$$

are nonnegative, at most two-sparse, exact explanations. Their contributions to group $\{1, 2\}$ differ by $2At/\sqrt{1+t^2}$. The column approximation error is of order t and the magnitudes of these codes are of order A, so a contribution of order δB persists even at $\varepsilon = 0$. A pure reconstruction-error guarantee would be false.

H INPUT-SPECIFIC CERTIFICATES WITHOUT CHOOSING A UNIQUE CODE

The exact global algorithms above do not cover every decoder. The following convex certificates can be useful on a fixed input without assuming forest structure. They are sufficient, not complete.

H.1 NONNEGATIVE DUAL BOUNDS

For any $c \in \mathbb{R}^m$ and any $a \in \mathbb{R}^n$ satisfying $D^\top a \geq c$ entrywise, nonnegativity implies

$$c^\top z \leq a^\top D z \leq a^\top x + \varepsilon \|a\|_2 \qquad (z \in \mathcal{Z}_\varepsilon(x)).$$

Thus the SOCP value

$$U_x(c) = \inf_{a : D^\top a \geq c} \{a^\top x + \varepsilon \|a\|_2\}$$

gives $-U_x(-c) \leq c^\top z \leq U_x(c)$. Any feasible dual vectors already give valid bounds, even before optimization has converged. These constraints ignore sparsity and may be infinite; no dual strong-duality assertion is needed for their validity.

Let $q_1, \ldots, q_r$ form an orthonormal basis of $\mathrm{range}(DW)$, and put $c_l = WD^\top q_l$. Obtain a finite certified interval $[L_l, U_l]$ for each $c_l^\top z$. With $b_l = U_l - L_l$, every pair has target distance at most $(\sum_l b_l^2)^{1/2}$. The vector

$$\widehat{v} = \sum_l q_l (L_l + U_l)/2$$

has error at most half this quantity against every feasible target. It need not be the target of any one feasible code. To exploit sparsity completely one may enumerate all supports of size at most k and optimize the linear objective over each corresponding nonnegative residual ball, which is an SOCP; this is a combinatorial alternative, not a general polynomial-time algorithm.

Setting $c = \mathbf{1}$ gives a uniform amplitude certificate. A feasible a implies

$$\|z\|_1 \leq a^\top x + \varepsilon \|a\|_2 \quad \text{for every } z \in \mathcal{Z}_\varepsilon(x).$$

Absence of such a simple dual certificate does not prove the sparse feasible set unbounded.

H.2 A LINEAR ESTIMATOR WITH A CERTIFIED RADIUS

Suppose a valid B is available. For any $R \in \mathbb{R}^{n \times n}$, let

$$\eta(R) = \max_j \|(w_j I - R) d_j\|_2.$$

Then

$$\begin{aligned} \|DWz - Rx\|_2 &\leq \|R(Dz - x)\|_2 + \|(DW - RD)z\|_2 \\ &\leq \varepsilon \|R\|_{\mathrm{op}} + B\eta(R) =: r(R). \end{aligned}$$

Thus $\Delta_w(x) \leq 2r(R)$. Minimizing $\varepsilon t + B\eta$ over R, t, η subject to $\|R\|_{\mathrm{op}} \leq t$ and the column norm constraints is convex, with standard semidefinite and second-order cone representations. If $RD = DW$ exactly, this reduces to the noise amplification of a linear projection. Approximate fitting trades residual noise amplification against unresolved feature directions. The estimate $x - Rx$ is a common approximation of intervention outcomes, not necessarily an outcome generated by a particular feasible code.

I COLLATERAL COST, REFUSAL, AND COUNTEREXAMPLES

I.1 A NECESSARY COST OF STABILIZATION

For $w_T = \mathbf{1}$, write $H = D(W - P_T)$. Since $DP_T = DW - H$, any two feasible codes satisfy

$$\|DP_T(z - z')\|_2 \leq \|DW(z - z')\|_2 + \|Hz\|_2 + \|Hz'\|_2.$$

Taking suprema proves (20). Hence an actual feasible pair with target discrepancy greater than $\tau + 2b$ rules out every such intervention with ambiguity at most τ and collateral at most b. Unlike the global modulus, this refusal is specific to the observed x. An upper certificate for Hz can be constructed by Appendix H; if $\|Hz - \widehat{v}_H\|_2 \leq r_H$ uniformly, then $C_{w|T}(x) \leq \|\widehat{v}_H\|_2 + r_H$.

An additive weight objective in the path dynamic program is a different declared cost. Its exact optimality does not by itself give the minimum of $C_{w|T}(x)$. For $k = 1$ the relative norm cost reduces to the maximum extra weight, as proved above; the distinction matters for larger k.

I.2 Coarsening is not monotonically stabilizing

Let $D = [(1,1,1)^\top/\sqrt{3}, e_1, e_2]$ and $k = 3$. Since D is invertible and $\mathcal{H}_k^+ = \mathbb{R}^3$, the modulus of a group is the operator norm of DP_GD^{-1}. For $x = (x_1, x_2, x_3)$, the code is

$$z_1 = \sqrt{3}x_3, \qquad z_2 = x_1 - x_3, \qquad z_3 = x_2 - x_3.$$

These formulas evaluate the linear operators on all differences; individual codes in a midpoint witness are still nonnegative. Direct calculation gives

$$\kappa(\{1\}) = \sqrt{3}, \quad \kappa(\{2\}) = \kappa(\{3\}) = \sqrt{2},$$
$$\kappa(\{1,2\}) = \kappa(\{1,3\}) = \sqrt{2}, \qquad \kappa(\{2,3\}) = \sqrt{3}.$$

For example, group $\{1,2\}$ maps x to (x_1, x_3, x_3), while $\{2,3\}$ maps it to $(x_1 - x_3, x_2 - x_3, 0)$. Their squared norms have maximum eigenvalues 2 and 3. With $K = 1.5$, the partitions $\{\{1,2\},\{3\}\}$ and $\{\{1,3\},\{2\}\}$ are feasible but no common refinement is feasible, because it would isolate feature 1. Also, merging the individually feasible groups $\{2\}$ and $\{3\}$ becomes infeasible. Thus no general noisy "unique finest partition" or monotone merge rule is valid.

I.3 Single-activation constraints are insufficient for multiple activations

For $d_1 = e_1$, $d_2 = (e_1 + e_2)/\sqrt{2}$, and $d_3 = e_2$, the vector $(1, -\sqrt{2}, 1)$ is a feasible null direction when $k = 2$. The circuit theorem forces every finitely stable diagonal weight to be constant on all three coordinates. Pairwise constraints at a finite K alone can permit nonconstant weights and therefore cannot certify this setting. The forest theorem captures the balanced three-vertex obstruction; its $k = 1$ specialization does not and is not intended to.

I.4 A conservative hierarchy when exact selection is unavailable

For an arbitrary candidate grouping tree, suppose each node G has a valid bound $U_G(x) \geq \Delta_G(x)$ and an additive partition cost $c(G)$. Among tree prunings whose selected nodes satisfy $U_G(x) \leq \tau$, the recurrence

$$F(G) = \min \left\{ c(G) \text{ if } U_G(x) \leq \tau, \sum_{H \in \text{children}(G)} F(H) \right\}$$

finds the least-cost certified pruning. A leaf without an available feasible choice has value $+\infty$. A pruning either chooses G itself or partitions through its children, so induction proves optimality within that tree. This neither proves arbitrary-partition optimality nor permits a union of accepted nodes to inherit the same bound without re-certification. The path theorem is stronger precisely because it optimizes over all subsets rather than over a prescribed hierarchy.

I.5 Downstream statements and their limits

Suppose $\|DWz - \widehat{v}\|_2 \leq r$ uniformly on $\mathcal{Z}_\varepsilon(x)$, and a scalar readout f is L_f-Lipschitz on all the relevant intervention and estimated states. Then

$$|f(x - DWz) - f(x - \widehat{v})| \leq L_f r.$$

Consequently the estimated effect $f(x) - f(x - \widehat{v})$ has the same sign for every feasible code if its magnitude exceeds $L_f r$. For a linear readout, one can instead certify the scalar target $a^\top DWz$ directly using the dual bounds, potentially obtaining a tighter guarantee than certifying the full vector. Stability of one readout does not establish stability of every readout. These are representation-relative intervention results; they identify neither human concepts nor real-world causal variables and do not control unspecified off-manifold behavior of a downstream network.

J Validation protocol, artifacts, and interpretation

All results in Table 1 were generated by the supplied scripts in a fresh deterministic run. No external dataset, trained language model, GPU, or human annotation is used. The files `results/summary.json` and the per-suite CSV files contain the recorded counts; `forest_witnesses.json` contains every rational rejection witness. The theoretical proofs do not depend on these finite tests.

Exact boundary tests. Thirteen checks cover PSD equality at a rational two-column boundary, attaching a zero-Schur-complement child, duplicate directions with equal or unequal intervention weights, and simple accepted cases. Two additional invalid inputs (a cyclic nonzero graph and an invalid Gram input) must raise errors. Equality is not replaced by a numerical epsilon.

Forest certification. Seed 7019 generates rational diagonally dominant forest Gram matrices with both edge signs, $m = 2, \ldots, 9$, several sparsity budgets and binary or quarter-grid weights, and rational squared budgets including $101/100$, $121/100$, $9/4$, and 4. All 580 configurations are compared to independently enumerated balanced connected subtrees with exact rational PSD tests. The enumerator examines 11,860 subtrees. The 135 rejected configurations each have an exact rational witness whose support counts and negative quadratic form are independently re-evaluated. There are 445 accepted configurations and zero decision mismatches. Input construction ensures PSD.

Optimal path masks. Seed 9141 generates positive rational path Gram matrices with $m = 3, \ldots, 8$, $k \in \{1, 2, 3\}$, singleton targets, positive integer costs, and multiple squared stability budgets. For each of 72 problems, every target-containing binary mask is checked using the independent exhaustive forest criterion, not the path-window implementation being tested. The cumulative number of candidates is 3,024. The dynamic-programming optimum agrees with the exhaustive optimum in every case. Ties need not return the same mask.

Single-activation attenuation. Seed 4912 generates 120 dense normalized random dictionaries, periodically with exact duplicate columns, targets, and budgets. In the transformed variables s, an independent linear program minimizes a strictly positive linear objective subject to all edge constraints and target equalities. Its solution agrees with the shortest-path envelope to maximum coordinate discrepancy 5.6×10^{-16}; the largest observed pair-constraint residual is below 2.7×10^{-15}. These are floating-point cross-checks, not interval-certified evaluations of transcendental functions. The separately rounded rational path example is exactly certified. Figure 1 evaluates the proved path formulas at 180 budgets and is regenerated from `attenuation_frontier.csv`.

Hardness reduction. All 1,024 labeled graphs on five vertices are tested at four clique sizes, giving 4,096 reduction decisions. Clique existence is enumerated, and all candidate supports in the reduction value are examined. There are no mismatches. A subset of 256 decisions also uses exact rational inverses; the largest discrepancy from their floating evaluations is below 3.5×10^{-18}. These tests check the finite reduction calculations; only the stated subset uses exact inverses.

Circuit characterization. Four small rational dictionaries and sparsity budgets 1 and 2 give 112 mask decisions. Minimal dependent supports and signed circuits are computed with exact symbolic arithmetic; predictions are compared with linear optimization over each admissible sign orthant, normalized by $\|h\|_1 \leq 1$, imposing $Dh = 0$. No mismatches occur. These dictionaries need not have unit columns because the kernel/circuit statement is invariant under positive column rescaling. The LP comparisons are numerical, whereas circuit enumeration is exact.

Perturbation transfer. Seed 324 constructs orthonormal prototypes, repeats each three times, perturbs and renormalizes each column, and measures the resulting maximum column perturbation. For 240 generated pairs of nonnegative sparse codes, choose their reconstruction midpoint, set ε to half their reconstruction discrepancy, and take B as the larger of the two code ℓ_1 norms. All tested pairs satisfy the transfer inequality; the largest observed left-to-right ratio is approximately 0.827. This original suite checks the pairwise algebra only. The whole-input suite below replaces its pair-derived amplitude by a uniformly valid bound.

All-sparsity separation and exact binary obstructions. The suite uses $q \in \{1, 2\}$ orthogonal pairs, replication $r \in \{2, 3\}$, and $K \in \{6/5, 31/25, 1249/1000\}$, for 12 dictionary–budget combinations. For each, the soft candidate (14) passes an independent full-matrix exact PSD test. Recording the distinct budgets $k \in \{1, \min(2, m), \min(4, m), m\}$ yields 45 configurations. All 276 target-containing binary masks are considered. Apart from the 12 full-deletion masks, every mask is rejected by an explicit rational vector with one positive and one negative coordinate. Cutting copies gives a null direction with nonzero target; otherwise splitting an interacting pair gives the direction

from Appendix E.2. All 264 witnesses are checked by directly evaluating $K^2\|Dh\|_2^2 - \|DWh\|_2^2$. The largest displayed binary-to-soft collateral ratio is approximately 449.36; the unbounded-ratio statement follows from the limit in the proof, not this finite value.

Rational dense perturbation generator. The whole-input suite uses $p \in \{4, 8\}$ pair prototypes, $r \in \{3, 8\}$ copies per prototype, $k \in \{1, 2, 4\}$, five parameters $t \in \{0, 1/1000, 1/200, 1/50, 1/20\}$, and two replicates. The seed is

$$90217 + 100000p + 1000r + 100k + 2(10000t) + \text{rep}, \qquad \text{rep} \in \{0, 1\}.$$

For each copy of a rational unit prototype u, draw a vector with coordinates in $\{-3, -2, \ldots, 3\}/p$ and project it orthogonally to u, obtaining a rational $b \perp u$. Define

$$d(t) = \frac{(1 - t^2\|b\|_2^2)u + 2tb}{1 + t^2\|b\|_2^2}.$$

All columns have exactly unit norm. The reference matrix repeats the unperturbed prototypes. The generator computes the maximum squared column error exactly and rounds its square root upward to a rational δ; t itself is not reported as the error bound. For every instance the first three copies have nonzero pairwise inner products, providing a triangle that rules out an exact raw forest.

Uniform certificates, rather than a pair-derived mass bound. Let $\mu = \min_j \mathbf{1}^\top d_j$, which is positive for every generated case. The rational vector $a = \mathbf{1}/\mu$ satisfies $D^\top a \geq \mathbf{1}$ exactly. Draw k active prototype groups and positive integer masses in $\{1, \ldots, 5\}$, then normalize the masses to sum one. Their nominal reconstruction is $x = U\alpha$, and $\varepsilon = \delta + 1/1000$. Every code obtained by choosing one raw representative per active group is feasible, since its reconstruction differs from x by at most $\delta\|\alpha\|_1 = \delta$. The uniform amplitude bound is the minimum of the valid dual bound and the cone bound, using outward rational norm enclosures; it is computed before selecting any tested pairs.

The nominal intervention has $v_g = 1$ on odd prototypes and $v_g = 3/25$ on even prototypes, with $K = 6/5$. Exact nominal certification and Theorem 11 give the ambiguity upper bound R and actual collateral upper bound. We compare R to $0.3(\|x\|_2 + \varepsilon)$ using a rational *lower* enclosure for $\|x\|_2$ on the threshold side. This yields 77 certified and 43 inconclusive instances, with success counts $24, 24, 24, 5, 0$ in order of increasing t. A separate sample of 16 feasible pairs per instance gives 1,920 exact checks of the ambiguity and collateral inequalities. These samples are consistency checks; uniform validity comes from the exact dual or cone premises.

All norm enclosures use integer square roots on a denominator grid 10^{10} and are rechecked by squaring. The file `uniform_input_certificates.json` records the actual rational dictionaries, dual vectors, error bounds, observations, noise tolerances, weights, radii, and sampled codes. The independent `verify_robust.py` imports only the Python standard library; it checks the premises and conclusions without calling the generator or forest DP.

Exact input-specific refusal. For $t \in \{1/100, 1/20, 1/10\}$ and $A \in \{1, 10\}$, use the rational unit columns

$$d_1 = (p, q)^\top, \quad d_2 = (p, -q)^\top, \quad d_3 = (0, 1)^\top, \qquad p = \frac{1 - t^2}{1 + t^2}, \quad q = \frac{2t}{1 + t^2}.$$

At $x = Ad_1$ and $\varepsilon = 0$, the codes $(A, 0, 0)$ and $(0, A, 2Aq)$ are exact and nonnegative with support size at most two. For target $\{1, 2\}$ their discrepancy is $g = 2Aq$. Setting $\tau = b = g/4$ gives $g > \tau + 2b$, which is a genuine refusal witness for the observed input. All six witnesses are verified exactly. In contrast, the 43 inconclusive robust instances carry no such impossibility claim.

Reproducibility and implementation coverage. The supplement contains deterministic generators, dependency versions, exact witness and uniform-certificate records, two standalone checkers, raw outputs, and the analytic figure script. Reproduction commands are in the package README. Exact certification and witness verification require no conic optimizer. SciPy's LP solver is used only for the independent numerical comparisons in the original suites. The results files separate those comparisons from exact rational decisions. The main-text limitations specify the unimplemented SDP formulations and the scope of synthetic coverage.