# KOLO: EXECUTABLE CHALLENGE BASES FOR NEURAL TRAINING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Standard diagnostics for monitoring neural-network training, such as loss curves, validation accuracy, and gradient norms, are descriptive rather than constructive. They fail to answer whether a superior model remains accessible within the same architecture and objective. We address this by studying *executable challenges*, which preserve selected intermediate representations of a trained network and reconstruct or re-optimize the remaining computation to evaluate the resulting complete model under the original objective. While monitoring $P$ representations induces an exponential lattice of $2^P$ possible retention patterns, we introduce *KOLO* (Keep One, Leave One): a linear-size family of just $2P + 1$ challenges consisting of route zero, every single retained representation, and every single released representation. We show that, without structural assumptions, superior models can remain hidden in untested routes, but KOLO becomes a valid measurement basis when governed by a coherent complete-model compiler and bounded higher-order interactions. Theoretically, we characterize when pairwise interactions are exactly identified by KOLO—proving exact identification on trees and odd-unicyclic graphs—and geometrically distinguish *route reconstruction* (requiring linear-span coverage) from *direct threshold certification* (requiring convex-hull coverage). Operationally, Keep-One separates representation deficit from conditional optimization slack, while Leave-One can reveal more accessible finite-budget optimization paths. Finally, reconstructed route evidence transfers to a global suboptimality bound through explicit solver, comparison-class, and coverage defects. Empirically, on a trained CIFAR-10 network with $P = 8$, KOLO evaluates only 17 of 256 routes yet captures an average of $95.6\%$ of the exhaustive optimization headroom in the task-oriented challenge regime, retrieving the exact best route in 7 of 9 checkpoint–seed cases. These results establish executable challenge bases as a principled way to monitor not only what training achieved, but what better solutions remain constructively accessible.

## 1 INTRODUCTION

Training is commonly monitored through numbers produced by the optimizer: training and validation loss, accuracy, gradient norms, or parameter updates. These diagnostics are useful, but they are descriptive rather than constructive. When progress stalls, they cannot tell a genuinely strong checkpoint apart from one that has merely exhausted the specific training path used to reach it.

We study a complementary question: *can a declared procedure construct a better complete model while keeping the architecture, comparison class, data, and objective fixed?* We call such a procedure an *executable challenge*. It starts from a saved checkpoint, modifies or re-optimizes part of the computation, and counts as evidence only after it materializes a complete model and reevaluates the original objective. A lower-valued endpoint is therefore an explicit counterexample to the claim that the current training run has exhausted the declared alternatives. This complete-model convention appears in the broader Training Under Challenge framework (Yeganegi et al., 2026); the present paper is self-contained and develops a separate theory for choosing a small, informative family of challenges.

Consider a network divided at $P$ monitored internal boundaries. Let $Z_i^t$ denote the checkpoint representation at boundary $i$ on the certification data. A *route* selects a subset $S \subseteq [P]$: the



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.

representations indexed by $S$ must be reproduced, while the intervening parts of the network may be rebuilt or re-optimized. There are $2^P$ such retention patterns. Exhaustively testing them is therefore impossible once the number of monitored boundaries is even moderately large.

KOLO selects the two extreme layers of this route lattice, together with a baseline:

$$\mathcal{A}_{\mathrm{KOLO}} = \{\varnothing\} \cup \{\{i\} : i \in [P]\} \cup \{[P] \setminus \{i\} : i \in [P]\}. \tag{1}$$

Route zero, $\varnothing$, retains no intermediate representation. A *Keep-One* (KO) route preserves exactly one checkpoint representation and asks whether the rest of the network can be rebuilt around it. A *Leave-One* (LO) route preserves every monitored representation except one and asks whether releasing that interface creates an easier route to a better complete model. KOLO therefore uses only $2P + 1$ executions.

The central problem is not the route count itself. It is whether these observations say anything rigorous about the unexecuted middle routes. Our answer has two parts. First, there is no assumption-free guarantee: all KOLO routes may look harmless while an unexecuted route is substantially better. Second, KOLO becomes a genuine *measurement basis* once three conditions hold: its routes are generated by one coherent complete-model compiler, higher-order interactions between boundaries are bounded, and the declared low-dimensional route model is identifiable from the shell alone. Even then, reconstructing a hidden route's value and certifying that it passes a threshold remain different tasks: recovering a value only needs a signed linear combination of shell observations, while certifying passage needs a nonnegative (convex) combination, or some other one-sided structural argument.

The rest of the paper develops both the basis theorem and what the two shells reveal about training. KO has exact, representation-preserving semantics: it rules out every local escape that leaves even one audited representation unchanged. LO adds no new exact guarantee on top of that, but it can be far easier to optimize under a fixed budget, and it supplies exactly the co-singleton equations needed to identify pairwise route interactions. A final transfer theorem then states precisely which additional solver, comparison-class, and coverage guarantees are needed before route-level evidence becomes a global statement about the checkpoint. Deep-linear prediction images give a nontrivial setting in which this upgrade is exact.

**Contributions.**

- We prove the negative-positive basis theorem: KOLO cannot control the hidden middle without structure, but a coherent compiler, certified interaction tail, and identifiable route model give valid reconstruction intervals and minimal adaptive repairs.
- We distinguish identification from passage and formalize KO/LO semantics through exact dominance, finite-budget complementarity, representation accounting, and collective local escape.
- We give a directed route-to-global transfer theorem, an exact deep-linear island, and an exhaustive 256-route neural census measuring challenge strength and reconstruction accuracy per execution.

**Note:** To facilitate navigation through the manuscript, a table of terms is provided in Section A.

## 2 From a checkpoint to route values

The same route can take several different values, and conflating them leads to wrong conclusions. The *ideal constrained value* is the best objective achievable if the retained representations were enforced and the rest of the network optimized perfectly. The *executed value* is what a specific finite algorithm actually finds. The *coupled route cube* is a single coherent family of complete models used to infer values at routes that were never executed.

Formally, let $[P] = \{1, \dots, P\}$ index the monitored boundaries, $\Theta_{\mathrm{cert}} \subseteq \mathbb{R}^d$ the fixed comparison class, and $J : \Theta_{\mathrm{cert}} \to \mathbb{R}$ the unchanged objective (smaller is better). Let $\theta_t$ be a checkpoint with value $J_t = J(\theta_t)$, and let $X_{\mathrm{cert}}$ denote the certification sample (population versions later use sans-serif notation). For $\theta \in \Theta_{\mathrm{cert}}$, $Z_i(\theta; X_{\mathrm{cert}})$ is the representation that $\theta$ induces at boundary $i$ on $X_{\mathrm{cert}}$, and $Z_i^t := Z_i(\theta_t; X_{\mathrm{cert}})$ is the checkpoint's representation there. A retained representation is called *complete* when it includes every shared variable, side path, stateful buffer, source of randomness, and objective term through which the released part of the network could still affect the retained suffix. A formal factorization is given in Section B.

**Definition 1** (Ideal constrained value and finite executed value)**.** For $S \subseteq [P]$, define

$$\mathcal{C}_t(S) = \{\theta \in \Theta_{\mathrm{cert}} : Z_i(\theta; X_{\mathrm{cert}}) = Z_i^t \ \forall i \in S\}, \qquad V_t(S) = \inf_{\theta \in \mathcal{C}_t(S)} J(\theta). \tag{2}$$

The checkpoint itself belongs to $\mathcal{C}_t(S)$, so the class is nonempty. The number $V_t(S)$ is the best objective allowed by the retained representations, whether or not it can be computed. A budget-$B$ route algorithm instead returns a finite archive $\mathcal{E}_{t,B}^{\mathrm{nat}}(S) \subseteq \mathcal{C}_t(S)$ that contains the checkpoint as a fallback, giving the *executed value*

$$\widehat{V}_{t,B}^{\mathrm{nat}}(S) = \min_{\theta \in \mathcal{E}_{t,B}^{\mathrm{nat}}(S)} J(\theta). \tag{3}$$

Thus $V_t(S) \le \widehat{V}_{t,B}^{\mathrm{nat}}(S) \le J_t$.

If the constraints in $\mathcal{C}_t(S)$ are not satisfied exactly, the error-bound bridge of Section C applies.

**Definition 2** (Coupled route cube and cross-crediting)**.** Inferring unexecuted routes needs one shared construction, not a separate optimizer per subset. A single $C^3$ complete-model compiler $\Theta_t : U \to \Theta_{\mathrm{cert}}$ on an open $U \supset [0,1]^P$ defines

$$f_t(x) = J(\Theta_t(x)), \qquad F_t(S) = f_t(\mathbf{1}_S), \tag{4}$$

where $x_i = 1$ activates route coordinate $i$, so the Boolean vertices $x = \mathbf{1}_S$ correspond to routes. We record these values before any checkpoint fallback or cross-branch minimum, since otherwise they need not lie on one smooth route function. Separately, every discovered complete endpoint is *cross-credited*: it counts as evidence for every constraint set it satisfies, not only the one it was built for.

A route *passes* when it fails to improve the checkpoint by more than a declared tolerance. More precisely, a family $\mathcal{R}$ passes exactly at tolerance $\tau$ when $V_t(S) \ge J_t - \tau$ for every $S \in \mathcal{R}$. Finite or *native* passage replaces $V_t$ by the values actually found in Equation (3). *Raw shell* passage tests only the coupled values $F_t(A)$ at the observed KOLO routes $A \in \mathcal{A}_{\mathrm{KOLO}}$ (Equation (1)). Once a route model $\widehat{F}$ has been fit to these values (Section 3), *reconstructed all-route passage* is summarized by the tolerance

$$\tau_{\mathrm{rec}} = \left[ J_t - \min_{S \subseteq [P]} \widehat{F}(S) \right]_+ . \tag{5}$$

Only the last notion can enter the general route-to-global theorem without an additional pass-preserving property.

## 3 When a few routes determine the rest

### 3.1 Passing the observed routes need not certify the hidden routes

Suppose every route we execute has objective at least $J_t$, so none improves the checkpoint. Does the same conclusion follow for the routes we did not execute? Not without more structure, and the counterexample is simple. Take four boundaries and the pairwise interaction function

$$r(x) = \sum_{i<j} b_{ij} x_i x_j = -x_1x_2 + x_2x_3 - x_3x_4 + x_1x_4 = (x_3 - x_1)(x_2 - x_4), \tag{6}$$

where $b_{12} = b_{34} = -1$, $b_{23} = b_{14} = +1$, $b_{13} = b_{24} = 0$, and the last equality is a direct factorization. Because $r$ contains only products of two boundaries, the term $b_{ij}x_ix_j$ is active on a route $S$ only when both $i$ and $j$ belong to $S$. Hence $r(\mathbf{1}_S) = \sum_{\{i,j\} \subseteq S} b_{ij}$, an empty sum when $|S| \le 1$. This function vanishes at route zero and at every singleton (no pair is present), and at every co-singleton (each contains two edges of the 4-cycle with opposite signs, plus a diagonal pair with zero coefficient; e.g. $r(\{1,2,3\}) = b_{12} + b_{23} + b_{13} = -1 + 1 + 0 = 0$), and at the full route. Yet $r(\{1,2\}) = b_{12} = -1$, a strictly better hidden route for a lower-is-better objective. The symmetric pair matrix of $r$ is the matrix $B$ with $B_{ij} = B_{ji} = b_{ij}$ and $B_{ii} = 0$, so that $r(x) = \frac{1}{2} x^\top B x$:

$$B = \begin{pmatrix} 0 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 \\ 1 & 0 & -1 & 0 \end{pmatrix}.$$

Rows 3 and 4 are the negatives of rows 1 and 2, so $B$ has rank two. Scaling $r$ makes the hidden deviation arbitrary while leaving the shell unchanged: for the route function $F_c(S) = J_t + c\, r(\mathbf{1}_S)$ with any $c > 0$, every shell route has value $J_t$ regardless of $c$, but $F_c(\{1,2\}) = J_t - c$. Any lower bound $L$ on the unexecuted routes computed from shell values is therefore violated once $c > J_t - L$, so the shell places no limit on the hidden routes. A multilinear polynomial realizes every Boolean table smoothly, and even strongly convex quadratic subset elimination can generate nonzero interactions of every order (Section D.1). Therefore low degree, low rank, smoothness, or convex inner optimization alone does not turn shell passage into a certificate for the unobserved routes.

### 3.2 A COHERENT COMPILER MAKES INTERACTION CONTROL POSSIBLE

To say anything about unobserved routes, the route values must belong to a single coherent family. This is precisely the role of the compiler $\Theta_t$ introduced in Definition 2. Let $f$ be the objective value produced by a shared complete-model compiler as its continuous selector $x \in [0,1]^P$ varies, and let $F(S) = f(\mathbf{1}_S)$ denote the resulting Boolean function. Every such $F$ has a *multilinear extension* $\widetilde{F}$, defined as the unique polynomial of degree at most $P$ in $x$ that agrees with $F$ at all $2^P$ vertices. This polynomial decomposes $F$ into a constant term, one term for each boundary, pairwise interaction terms, and higher order interaction terms. Let $Q_2$ denote the sum of the first three components, namely the constant, linear, and pairwise terms.

**Theorem 3** (Boolean mixed-derivative bound)**.** *For distinct $i, j, k \in [P]$, let $M_{ijk} = \sup_{x\in[0,1]^P} |\partial_{ijk} f(x)|$ be the largest magnitude of the third mixed partial derivative of $f$ over the whole cube. Then every route $S$ satisfies*

$$|F(S) - Q_2(S)| \le \eta_S^{(3)} := \sum_{\{i,j,k\}\subseteq S} M_{ijk}. \tag{7}$$

*In particular, the largest such gap over all $2^P$ routes is at most $\sum_{i<j<k} M_{ijk}$.*

This is a global finite-difference statement holding at every Boolean vertex simultaneously, not a Taylor approximation at the checkpoint. The same argument holds at every order: the degree-$d$ Boolean Möbius truncation is controlled by $(d+1)$-fold mixed derivatives (Theorem 13).

For a neural compiler $f = J \circ \Theta_t$, the third-order chain rule gives an explicit bound,

$$M_{ijk} \le L_3 a_i a_j a_k + L_2(b_{ij}a_k + b_{ik}a_j + b_{jk}a_i) + L_1 c_{ijk}, \tag{8}$$

where $a_i, b_{ij}, c_{ijk}$ bound the first-, second-, and third-order route derivatives of the compiled model, and $L_1, L_2, L_3$ bound the corresponding derivatives of the objective on a certified tube around the compiled route. For *additive contractive* compilers (Appendix E.1 gives the explicit recurrence), these bounds stay uniform in depth and scale cubically in the route radius. In the degenerate case where the whole selector cube stays inside one affine activation cell and the terminal objective is quadratic, the route function is *exactly* pairwise, and consequently all third order derivatives $M_{ijk}$ vanish.

### 3.3 RECONSTRUCTING ROUTES AND PRESERVING THRESHOLDS

The interaction bound above is only useful if the observed shell actually pins down the low-order model $Q_2$. Suppose that, before seeing any route values, we declare a feature vector $\phi(S) \in \mathbb{R}^q$ capturing the effects we intend to model, for instance a constant, one coefficient per boundary, and a chosen set of pairwise terms. Let $M_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|\times q}$ stack the feature vectors of the observed shell $\mathcal{A}$ as rows, and write

$$y_{\mathcal{A}} = M_{\mathcal{A}}\beta + r_{\mathcal{A}} + e_{\mathcal{A}}, \qquad F(S) = \phi(S)^\top \beta + r(S), \tag{9}$$

where $\beta$ is the true coefficient vector of the declared feature model, $r$ is the (certified, bounded) model-misspecification remainder, and $e_{\mathcal{A}}$ is execution or measurement noise. Given positive weights $\Pi$ (a diagonal matrix with one entry $\pi_A > 0$ per observed route $A \in \mathcal{A}$), set $Q = M_{\mathcal{A}}^\top \Pi M_{\mathcal{A}}$ and estimate $\beta$ by the minimum-norm weighted least-squares solution $\widehat{\beta} = Q^\dagger M_{\mathcal{A}}^\top \Pi y_{\mathcal{A}}$, where $Q^\dagger$ is the Moore–Penrose pseudoinverse of $Q$.

**Theorem 4** (Querywise stable shell reconstruction)**.** *Fix a route $S$ whose feature is identifiable from the shell,*

$$\phi(S) \in \mathrm{range}(M_{\mathcal{A}}^\top) = \mathrm{range}(Q). \tag{10}$$

*If* $\|e_{\mathcal{A}}\|_{\Pi} \le \varepsilon_{\text{exec}}$ *and* $|r(T)| \le \eta_T$*, then*

$$|\phi(S)^\top \widehat{\beta} - F(S)| \le \ell^\dagger_{\mathcal{A}}(S)(\varepsilon_{\text{exec}} + \eta_{\mathcal{A}}) + \eta_S, \tag{11}$$

*where* $\ell^\dagger_{\mathcal{A}}(S) = \sqrt{\phi(S)^\top Q^\dagger \phi(S)}$, $\eta_{\mathcal{A}} = \left(\sum_{A\in\mathcal{A}} \pi_A \eta_A^2\right)^{1/2}$*. The leverage* $\ell^\dagger_{\mathcal{A}}(S)$ *is the exact linear amplification of weighted shell error at query* $S$*. If Equation* (10) *fails, a coefficient direction invisible on the shell can change* $F(S)$ *arbitrarily, so no finite uniform reconstruction bound is possible without additional assumptions.*

The proof of Theorem 4 can be found in Appendix F. Full column rank of $M_{\mathcal{A}}$ is sufficient to reconstruct *every* route in the declared feature class, but it is not necessary for any *one particular* query. More importantly, reconstruction and passage remain different problems: a route can be reconstructed exactly by a signed linear combination of shell values even when the shell values alone do not preserve a common lower threshold.

**Theorem 5** (Identification versus pass preservation)**.** *Assume the first coordinate of every* $\phi(S)$ *is one, and consider the exact linear class* $F_\beta(S) = \phi(S)^\top \beta$*. For a query* $S$*:*

*(i)* $F_\beta(S)$ *is determined by the shell for every* $\beta$ *if and only if* $\phi(S) \in \operatorname{span}\{\phi(A) : A \in \mathcal{A}_{\text{KOLO}}\}$*.*

*(ii) For every* $\beta$ *and threshold* $c$*, shell passage* $F_\beta(A) \ge c$ *for all* $A \in \mathcal{A}_{\text{KOLO}}$ *implies* $F_\beta(S) \ge c$ *if and only if*

$$\phi(S) \in \operatorname{conv}\{\phi(A) : A \in \mathcal{A}_{\text{KOLO}}\}. \tag{12}$$

*Equivalently, there are weights* $w_{S,A} \ge 0$ *summing to one with* $\phi(S) = \sum_A w_{S,A}\phi(A)$*.*

*With* $|r(T)| \le \eta_T$ *and* $|e_A| \le \varepsilon_A$*, such weights give the robust one-sided bound*

$$F(S) \ge \sum_A w_{S,A} y_A - \sum_A w_{S,A}(\varepsilon_A + \eta_A) - \eta_S. \tag{13}$$

The distinction is strict even for the simplest possible model, namely a constant plus one coefficient per boundary: for $P \ge 4$, no genuine middle route lies in the convex hull of route zero, the singletons, and the co-singletons (Corollary 14). KOLO is therefore naturally a *measurement basis*, good for reconstruction; certifying passage directly requires additional sign, ordering, or convexity structure on top of it.

The theorem yields a complete geometric distinction: signed span coefficients reconstruct, while convex coefficients preserve lower thresholds. The distinction is strict even under exact identification: on the path 1–2–3–4, $p(S) = 1 + x_4 - x_1x_2 + x_2x_3 - x_3x_4$ is exactly identified by route zero plus KOLO, every measured route has value at least one, yet $p(\{1,2\}) = 0$. A simple positive island is also immediate: if $F(\varnothing) = J_t$, the exact pair coefficients are nonnegative, and every singleton has value at least $J_t$, then every route has value at least $J_t$; an interaction-tail bound subtracts only the certified remainder.

### 3.4 When pairwise interactions follow a known graph

A common low-order model gives each boundary a main effect and allows a pairwise term only between boundaries that can plausibly interact. Represent the allowed pairs by a known simple graph $G = (V, E)$ with $P = |V| \ge 3$ vertices, and consider

$$F(S) = c + \sum_{i\in V} a_i x_i + \sum_{\{u,v\}\in E} b_{uv} x_u x_v. \tag{14}$$

Route zero determines the baseline $c$; the KO routes determine each main effect $a_i$; the LO routes determine, for each vertex, the *sum* of the pair coefficients on its incident edges. Recovering the individual edge coefficients $b_{uv}$ is therefore a graph incidence problem.

**Theorem 6** (Known-graph identifiability and minimum repair)**.** *Route zero plus KOLO identifies every coefficient of Eq.* (14) *if and only if every connected component of* $G$ *is a tree or is* odd-unicyclic *(has exactly one cycle, of odd length). In general, the number of coefficients left undetermined by the shell, called the* shell nullity*, is*

$$\nu(G) = |E| - |V| + b(G), \tag{15}$$

*where $b(G)$ is the number of bipartite connected components (isolated vertices count as bipartite). Exactly $\nu(G)$ additional scalar route measurements, chosen to reveal rank, are necessary and sufficient to complete the recovery. A component containing exactly one even cycle has one undetermined "alternating" direction, repaired by a single well-chosen pair-route measurement.*

The graph $G$ itself must be predeclared, fit on separate development data, or reported with explicit model-selection uncertainty: a perfect fit to the shell cannot, by construction, reveal an interaction direction that the shell cannot see at all. Exact identifiability also does not imply numerical stability. For a path or an odd cycle on $n$ vertices, the smallest relevant singular value of the incidence matrix is $2\sin(\pi/(2n))$, so worst-case noise amplification grows like $n/\pi$ (Corollary 16). The underlying rank fact about unsigned incidence matrices is classical (Godsil and Royle, 2001); what is new here is its interpretation as a route-shell recovery problem, together with the repair count and stability analysis. Figure 2 illustrates these points: the hidden-middle failure (a), exact tree recovery (b), the compiler's cubic error scaling (c), and the one-query even-cycle repair (d).

## 4 What Keep One and Leave One reveal about training

The preceding section treated KOLO as a measurement design over route values. We now ask a different question: what does success or failure of an individual KO or LO route say about the checkpoint and the training process? These claims use the hard constraints and the finite route algorithms directly; they do not rely on the route-reconstruction model.

### 4.1 Keep-One separates representation loss from optimization loss

Let $J^\star = V_t(\varnothing)$ be the best objective in the full comparison class. Let $J_i^\star = V_t(\{i\})$ be the best objective achievable while the checkpoint representation at cut $i$ is held fixed. Finally, let $K_i$ be the value of the complete model that the actual KO algorithm found after checkpoint fallback, so $J_i^\star \le K_i \le J_t$. The three gaps below answer three different questions: how much improvement was demonstrated, how much conditional optimization may remain, and how much the retained representation itself may exclude.

**Theorem 7** (Executable KO ledger)**.** *For every cut $i$,*

$$J_t - J^\star = \underbrace{J_t - K_i}_{\text{demonstrated gain}} + \underbrace{K_i - J_i^\star}_{\text{conditional solve slack}} + \underbrace{J_i^\star - J^\star}_{D_i^{\mathrm{rep}}\text{: representation deficit}}. \tag{16}$$

*If $L_0 \le J^\star$, then*

$$(K_i - J_i^\star) + (J_i^\star - J^\star) \le K_i - L_0. \tag{17}$$

The identity is algebraically simple but operationally useful. If KO finds an improvement while $Z_i^t$ is unchanged, that improvement cannot require replacing the retained representation. Conversely, if a valid global lower bound $L_0$ lies close to $K_i$, then the total unresolved combination of representation loss and conditional solve error is small.

Under population evaluation, this representation term has familiar statistical meanings. If $\mathsf{Z}_i$ is a function of $\mathsf{X}$ and both comparison classes realize the relevant Bayes rules, then log loss gives

$$D_i^{\mathrm{rep}} = I(\mathsf{Y}; \mathsf{X} \mid \mathsf{Z}_i),$$

while unhalved squared loss gives the conditional-mean projection error $\mathbb{E}\,\|\mathbb{E}[\mathsf{Y} \mid \mathsf{X}] - \mathbb{E}[\mathsf{Y} \mid \mathsf{Z}_i]\|^2$. These identities require population risk, complete cuts, integrability, and Bayes-complete prediction classes; restricted decoders add explicit approximation terms (Section H).

### 4.2 What exact optimization says, and why Leave-One can still help

If the route-constrained problems could be solved exactly, keeping fewer representations can only enlarge the feasible set. This elementary order has an important consequence.

**Proposition 8** (Exact anchored dominance)**.** *If $i \in S$, then $\mathcal{C}_t(S) \subseteq \mathcal{C}_t(\{i\})$ and $V_t(\{i\}) \le V_t(S)$. Consequently, singleton KO passage at tolerance $\tau$ implies passage of every nonempty exact hard-anchor route at the same tolerance. For $P \ge 2$, LO adds no new exact hard-anchor feasible-set coverage.*

In a deterministic serial model with a complete retained downstream cut, terminal-only objective, no cross-cut sharing, and no side path, releasing an earlier equality leaves the exact constrained value unchanged. Thus the *marginal improvement* of that interior LO release is zero. Nonzero LO evidence instead reveals finite search accessibility or a violated shielding assumption, not automatically a causally defective layer.

Finite route algorithms need not preserve the exact order. Let

$$\begin{aligned} V^{\mathrm{KO}} &= \min_i V_t(\{i\}), & V^{\mathrm{LO}} &= \min_i V_t([P] \setminus \{i\}), \\ \widehat{V}_B^{\mathrm{KO}} &= V^{\mathrm{KO}} + \varepsilon_B^{\mathrm{KO}}, & \widehat{V}_B^{\mathrm{LO}} &= V^{\mathrm{LO}} + \varepsilon_B^{\mathrm{LO}}, \end{aligned} \tag{18}$$

where each aggregate uses the best route in its shell and fallback makes the solve gaps nonnegative.

**Theorem 9** (Finite reversal and matched-budget separation)**.** *Although* $V^{\mathrm{KO}} \le V^{\mathrm{LO}}$*,*

$$\widehat{V}_B^{\mathrm{LO}} < \widehat{V}_B^{\mathrm{KO}} \quad \Longleftrightarrow \quad \varepsilon_B^{\mathrm{KO}} - \varepsilon_B^{\mathrm{LO}} > V^{\mathrm{LO}} - V^{\mathrm{KO}}. \tag{19}$$

*For* $P = 3$*, checkpoint* $(0, 0, 0)$*,* $M = n^2$*, and*

$$J_M(x, y, z) = \tfrac{1}{2}(x-1)^2 + \tfrac{M}{2}(y - M^{-1})^2 + \tfrac{M}{2}(z - M^{-1})^2, \tag{20}$$

*run the same projected steepest-descent rule, restarted from the checkpoint on each affine route, with exact line search and budget* $B = n$*. Then*

$$V^{\mathrm{KO}} = \frac{1}{2n^2} < \frac{1}{n^2} = V^{\mathrm{LO}}, \qquad J_t - \widehat{V}_B^{\mathrm{LO}} = \frac{1}{2}, \qquad J_t - \widehat{V}_B^{\mathrm{KO}} \le \frac{2}{n}. \tag{21}$$

The winning LO restriction has condition number one, while the best KO restriction retains eigenvalues $1$ and $M$. No stronger optimizer or unequal oracle budget is assigned to LO. The exact crossover budget, derived in Corollary 18, grows as $M \log M/4$. Conversely, if a declared simulation maps every budget-$B$ LO transcript to a KO transcript within budget $cB$ and objective slack $\delta$, then $\widehat{V}_{cB}^{\mathrm{KO}} \le \widehat{V}_B^{\mathrm{LO}} + \delta$. A finite LO advantage therefore requires the absence of such an accurate comparable-budget simulation under the declared model. Cross-crediting then assigns every LO-discovered endpoint to each weaker KO class it actually satisfies while preserving its LO provenance.

### 4.3 After Keep-One passage, hidden local escape is collective

Let $A_i$ be the Jacobian of the $i$th waypoint equality and $T_i = \ker A_i$. At a stationary checkpoint with Hessian $H$:

**Theorem 10** (Collective escape)**.** *If* $v^\top H v \ge 0$ *on every* $T_i$*, then*

$$v^\top H v < 0 \implies A_i v \neq 0 \; \forall i, \qquad n_-(H) \le \min_i \operatorname{rank}(A_i).$$

*The bound is sharp. For* $P \ge 2$*, every LO tangent space is contained in retained KO tangent spaces, so LO adds no infinitesimal coverage after complete KO passage.*

Proofs, the nonlinear Lagrangian form, and quantitative visibility bounds are in Section J.

## 5 From route evidence to a global conclusion

Reconstructing every route is still not the same as proving global optimality. Three additional gaps may remain. First, the compiled route family may not solve the ideal route problems exactly. Second, the route family may cover only a narrower target comparison class. Third, that target class may still be smaller than the full certified model class. We keep these gaps separate and give each one a direction.

Let $F_{\mathrm{comp}}(S)$ be the value of the complete model produced by the coupled compiler at route $S$, and let $V_{\mathrm{ideal}}(S)$ be the ideal value of the corresponding declared route problem. Define

$$F_{\mathrm{comp}}^\star = \min_{S \subseteq [P]} F_{\mathrm{comp}}(S), \qquad V_{\mathrm{route}}^\star = \inf_{S \subseteq [P]} V_{\mathrm{ideal}}(S), \tag{22}$$

let $V^\star_{\mathrm{target}}$ be the optimum of a declared target comparison class, and recall that $J^\star = \inf_{\theta\in\Theta_{\mathrm{cert}}} J(\theta)$ is the optimum over the full certified class. Assume

$$0 \le F^\star_{\mathrm{comp}} - V^\star_{\mathrm{route}} \le \varepsilon_{\mathrm{sol}}, \tag{23}$$

$$0 \le V^\star_{\mathrm{route}} - V^\star_{\mathrm{target}} \le \varepsilon_{\mathrm{class}}, \tag{24}$$

$$0 \le V^\star_{\mathrm{target}} - J^\star \le \delta_{\mathrm{cov}}. \tag{25}$$

The first term includes any certified correction from approximate anchor residuals. The second is a route-class transfer, and the third is the independent coverage gap from the target class to $\Theta_{\mathrm{cert}}$.

**Theorem 11** (Routewise lower-confidence transfer)**.** *Suppose every route has a certified reconstruction interval*

$$|F_{\mathrm{comp}}(S) - \widehat{F}(S)| \le E_{\mathrm{rec}}(S).$$

*Define the routewise lower confidence value and its checkpoint gap by*

$$\underline{F}(S) = \widehat{F}(S) - E_{\mathrm{rec}}(S), \qquad \tau_{\mathrm{LCB}} = \left[ J_t - \min_{S\subseteq[P]} \underline{F}(S) \right]_+ . \tag{26}$$

*Then*

$$J_t - J^\star \le \tau_{\mathrm{LCB}} + \varepsilon_{\mathrm{sol}} + \varepsilon_{\mathrm{class}} + \delta_{\mathrm{cov}}. \tag{27}$$

*For the shell reconstruction in Theorem 4, one may take*

$$E_{\mathrm{rec}}(S) = \ell^\dagger_{\mathcal{A}}(S)(\varepsilon_{\mathrm{exec}} + \eta_{\mathcal{A}}) + \eta_S \tag{28}$$

*for every identifiable route. The lower-confidence form is never weaker than the simpler bound $\tau_{\mathrm{rec}} + \max_S E_{\mathrm{rec}}(S)$. Raw shell passage can replace reconstructed all-route passage only when Theorem 5 or another proved one-sided law applies.*

The theorem gives the monitor's final, mechanically interpretable output. The quantity $\tau_{\mathrm{LCB}}$ searches the entire reconstructed route landscape while respecting each route's own uncertainty. The remaining terms say how far the compiled routes may be from ideal route optimization, how far the route comparison class may be from the intended target class, and how far that target class may be from the full certified model class. If any bridge is unavailable, the conclusion stops at the preceding level rather than being silently promoted to a global claim.

### 5.1 An exact deep-linear island

Let $\mathcal{P}^\rho_X = \{Q : \mathrm{row}(Q) \subseteq \mathrm{row}(X), \mathrm{rank}(Q) \le \rho\}$ be the full prediction image and $\mathcal{P}^s_H = \{Q : \mathrm{row}(Q) \subseteq \mathrm{row}(H), \mathrm{rank}(Q) \le s\}$ the image after retaining representation $H$.

**Theorem 12** (Prediction-image completeness)**.** *If $\mathcal{P}^s_H = \mathcal{P}^\rho_X$, exact passage of the retained KO route is globally optimal for every output-only objective. Under squared loss, an exact KO minimizer has gap*

$$J_t - J^\star = \tfrac{1}{2}\left[ \sum_{j=1}^{\rho} \sigma_j^2(YP_X) - \sum_{j=1}^{s} \sigma_j^2(YP_H) \right].$$

The image condition is necessary for target-uniform completeness; details are in Section L.

## 6 Exhaustive neural route census

We trained a smooth residual classifier on CIFAR–10 and audited $P = 8$ boundaries at epochs $20, 48, 80$ for three seeds. Every one of the 256 routes was materialized and evaluated on the same 8,000-example audit split. We compare KO, LO, KOLO, and all $|S| \le 2$ routes under a task-shaped compiler and a state-faithful compiler; capture is the fraction of exhaustive headroom recovered by the suite.

With only 17 routes, KOLO captures 95.6% of task-shaped headroom on average (s.d. 11.9%; median 100.0%), contains an exhaustive minimizer in 7/9 cases, and reaches at least $90\%$ in 8/9. The 37-route $|S| \le 2$ family reaches $100\%$, but KOLO yields $2.08\times$ more normalized headroom per route. KO

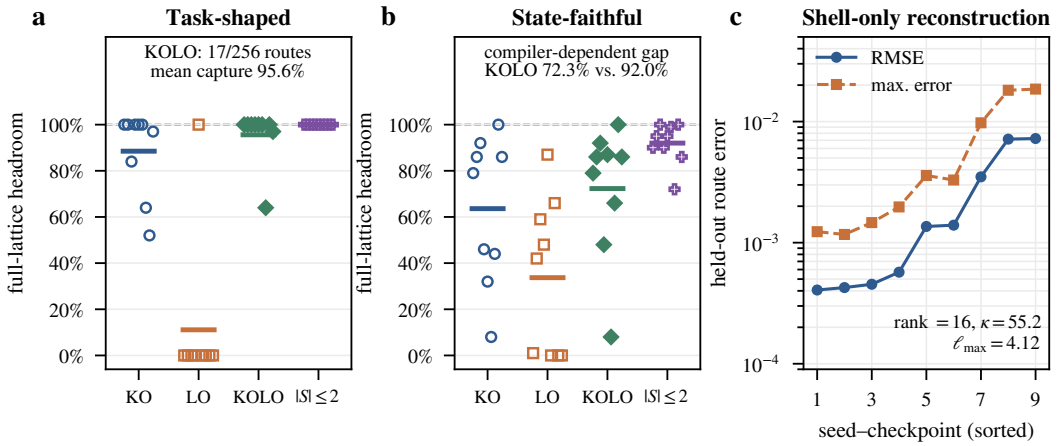



**Figure 1: Sparse monitoring against exhaustive ground truth.** (a) KOLO nearly saturates task-shaped headroom with 17 of 256 routes. (b) State-faithful routes expose genuine middle-route failures, showing that efficiency is compiler-dependent. (c) A coupled model fitted from the shell alone predicts all 256 values with small held-out error; rank, conditioning, and leverage quantify stability.

alone captures 88.5%; the paired increase shows that LO can add finite-search evidence. Under state-faithful reconstruction, KOLO captures 72.3% versus 92.0% for $|S| \leq 2$, confirming that strength depends on the compiler rather than route count alone. Fig. 1 summarizes this comparison.

A separate coupled model was fit using route zero plus KOLO only; all middle routes were held out. Its feature system has rank 16, condition number 55.2, and maximum leverage 4.12. Across nine cases, full-lattice RMSE is 4.1e-04–7.2e-03and maximum route error is 1.2e-03–1.9e-02. Only 6.6% of route features lie in the shell convex hull, empirically separating accurate signed reconstruction from raw pass preservation. Full per-run results and heatmaps appear in Section O, Table 4, and Figure 3.

## 7 Related Work

Local training, probes, stitching, causal interventions, and model editing use or alter internal states (Lee et al., 2015; Jaderberg et al., 2017; Hewitt and Liang, 2019; Bansal et al., 2021; Geiger et al., 2021; Meng et al., 2022). KOLO instead requires complete same-class, same-objective endpoints and treats routes as a measurement design. Boolean analysis, sparse set-function learning, and factorial design provide the closest reconstruction viewpoint (Rota, 1964; O'Donnell, 2014; Stobbe and Krause, 2012; Box et al., 2005); the graph proof uses a classical incidence-rank fact (Godsil and Royle, 2001). Training verification and deep-linear theory address complementary questions (Jia et al., 2021; Srivastava et al., 2024; Mao et al., 2025; Kawaguchi, 2016; Laurent and von Brecht, 2018; Shamir, 2019).

**Conclusion.** KOLO becomes a measurement basis only under explicit interaction and identifiability control; direct passage requires a one-sided law. The exhaustive census shows that this calibrated distinction is useful in practice: a linear shell can recover nearly all task-shaped challenge headroom while exposing when its compiler or assumptions are insufficient.

## AI Use Disclosure

Generative AI systems assisted with literature search, proof exploration, counterexample search, code drafting, experimental design, and copyediting. The authors selected the claims, reconstructed and checked the proofs, and independently reran the deterministic tests. All authors agreed on the final manuscript.

## References

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations. In *Advances in Neural Information Processing Systems*, vol-

ume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/01ded4259d101feb739b06c399e9cd9c-Abstract.html.

George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley-Interscience, 2 edition, 2005.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Chris Godsil and Gordon Royle. *Algebraic Graph Theory*. Springer, 2001.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

Max Jaderberg, Wojciech Marian Czarnecki, Simon Osindero, Oriol Vinyals, Alex Graves, David Silver, and Koray Kavukcuoglu. Decoupled neural interfaces using synthetic gradients. *Proceedings of the 34th International Conference on Machine Learning*, 70:1627–1635, 2017. URL https://proceedings.mlr.press/v70/jaderberg17a.html.

Hengrui Jia, Mohammad Yaghini, Christopher A. Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot. Proof-of-learning: Definitions and practice. In *2021 IEEE Symposium on Security and Privacy*, pages 1039–1056. IEEE, 2021. doi: 10.1109/SP40001.2021.00106.

Kenji Kawaguchi. Deep learning without poor local minima. In *Advances in Neural Information Processing Systems*, volume 29, 2016. URL https://proceedings.neurips.cc/paper/2016/hash/f2fc990265c712c49d51a18a32b39f0c-Abstract.html.

Thomas Laurent and James von Brecht. Deep linear networks with arbitrary loss: All local minima are global. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2902–2907. PMLR, 2018. URL https://proceedings.mlr.press/v80/laurent18a.html.

Chen-Yu Lee, Saining Xie, Patrick Gallagher, Zhengyou Zhang, and Zhuowen Tu. Deeply-supervised nets. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, volume 38 of *Proceedings of Machine Learning Research*, pages 562–570. PMLR, 2015. URL https://proceedings.mlr.press/v38/lee15a.html.

Yuhao Mao, Stefan Balauca, and Martin Vechev. CTBench: A library and benchmark for certified training. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 42984–43010. PMLR, 2025. URL https://proceedings.mlr.press/v267/mao25a.html.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

Ryan O'Donnell. *Analysis of Boolean Functions*. Cambridge University Press, 2014. doi: 10.1017/CBO9781139814782.

Gian-Carlo Rota. On the foundations of combinatorial theory i: Theory of möbius functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 2:340–368, 1964. doi: 10.1007/BF00531932.

Ohad Shamir. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. In *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2691–2713. PMLR, 2019. URL https://proceedings.mlr.press/v99/shamir19a.html.

Megha Srivastava, Simran Arora, and Dan Boneh. Optimistic verifiable training by controlling hardware nondeterminism. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3030. URL https://papers.nips.cc/paper_files/paper/2024/hash/ad885a9caafff30ee9cafdf0ee42fda2-Abstract-Conference.html.

Peter Stobbe and Andreas Krause. Learning fourier sparse set functions. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*. PMLR, 2012.

Farhang Yeganegi, Arian Eamaz, and Mojtaba Soltanalian. Training under challenge: Executable certificates and challenge-closed optimality for neural networks. *arXiv preprint arXiv:2608.12655*, 2026. doi: 10.48550/arXiv.2608.12655. URL https://arxiv.org/abs/2608.12655.

## A A TABLE OF NOTATIONS AND DEFINITIONS

**Terminology, in the order it is needed.** Readers unfamiliar with this framework may want to skim this table before Section 1; each term is also defined inline the first time it appears in the running text.

| Term | Definition |
|---|---|
| Checkpoint $\theta_t$ | A saved snapshot of a network's parameters at some point during training, with objective value $J_t = J(\theta_t)$. |
| Boundary (or *cut*) $i \in [P] := \{1, \dots, P\}$ | A chosen point inside the network's computation graph, typically the output of a layer, where we record an intermediate value for monitoring. |
| Representation $Z_i(\theta)$ | The value produced at boundary $i$ by parameters $\theta$, evaluated on the fixed certification data; $Z_i^t := Z_i(\theta_t)$ is the checkpoint's own value there. |
| Route $S \subseteq [P]$ | A declared retention pattern: a subset of boundaries whose representations must be reproduced exactly, while every other part of the network is free to be rebuilt or re-optimized. |
| Complete model | Any full, end-to-end network that can be run and scored under the original objective $J$, as opposed to a partial computation or an internal representation alone. |
| Route lattice | The set of all $2^P$ possible routes, ordered by set inclusion. |
| KOLO shell $\mathcal{A}_{\mathrm{KOLO}}$ | The linear-size family of $2P+1$ routes consisting of route zero (nothing retained), every *Keep-One* (KO) route (exactly one boundary retained), and every *Leave-One* (LO) route (every boundary retained except one). |
| Compiler $\Theta_t$ | A single smooth function that turns a continuous "activation" vector $x \in [0,1]^P$ into a complete model, in a way that agrees with the route-constrained optimum whenever $x$ is a $0/1$ vertex. It is what lets us treat routes as points sampled from one coherent surface rather than $2^P$ unrelated optimization problems. |
| Route cube | The domain $[0,1]^P$ of the compiler; its $2^P$ Boolean vertices $x = \mathbf{1}_S$ correspond exactly to the $2^P$ routes. |
| Cross-crediting | Reusing a single discovered complete model as valid evidence for *every* route whose constraints it happens to satisfy, not only the route it was optimized for. |
| Passage | A route *passes* at tolerance $\tau$ if its value does not beat the checkpoint's objective by more than $\tau$; a route that passes has not demonstrated a better model. |
| Headroom | The gap between the checkpoint's current objective value and the best value that some declared procedure can find or certify. |
| Shielding | The assumption that, once a downstream representation is held fixed, released upstream parameters have no further way to affect the objective. |
| Leverage $\ell^{\dagger}_{\mathcal{A}}(S)$ | In the reconstruction theorem, the exact factor by which measurement error on the observed shell is amplified when predicting an unmeasured route $S$; the same notion as leverage in weighted least squares. |

## B STANDING ASSUMPTIONS AND IMPLEMENTATION SEMANTICS

**Spaces and finiteness.** For the finite-dimensional optimization statements, $\Theta_{\mathrm{cert}} \subseteq \mathbb{R}^d$ and each vectorized waypoint map is $h_i : \mathbb{R}^d \to \mathbb{R}^{m_i}$. Euclidean inner products are used on parameter space and Frobenius inner products on batched states. The objective is bounded below on every class appearing in a difference, so

$$-\infty < J^\star \le V_t(S) \le J_t < \infty.$$

We use infima unless a finite archive or an explicit attainment result justifies a minimum. Every hard-anchor class is nonempty because it contains the checkpoint.

**Complete cuts.** At a complete cut $i$, all downstream dependence on the released prefix factors through a retained state $R_i = (Z_i, U_i)$:

$$f_\theta(X_{\mathrm{cert}}) = g_{\theta_{>i}}(R_i(\theta; X_{\mathrm{cert}})), \tag{29}$$

where $U_i$ includes shared parameters exposed to both sides, side-path values, stateful buffers, random seeds or distributions, and every prefix-dependent quantity in the objective. The shielding proposition additionally assumes that, conditional on $R_i$, the objective is independent of released prefix parameters. If this factorization fails, the cut is incomplete and shielding or information language must be amended.

**Population notation.** Population identities are stated on a probability space with random variables $(\mathsf{X}, \mathsf{Y})$ and $\mathsf{Z}_i = Z_i(\theta_t; \mathsf{X})$. Conditional entropies are assumed finite for log loss. For squared loss, $\mathsf{Y}$ is square integrable in a separable Hilbert space. The empirical sample $X_{\mathrm{cert}}$ is never substituted for the population random variable.

**Passage.** For lower-is-better objectives, a route value $W(S)$ passes at tolerance $\tau$ when $W(S) \geq J_t - \tau$. Exact, native, raw-cube, and reconstructed passage use different $W$ and are never interchanged. For $P = 1$ there is no distinct LO route. For $P = 2$, singleton and co-singleton routes coincide and do not identify an unrestricted pair coefficient. Graph results assume $P \geq 3$. The product lower bound assumes $K > 2$.

## C Approximate anchors and cross-crediting

Practical compilers may satisfy waypoint equalities only approximately. Let $h_S(\theta) = (h_i(\theta))_{i \in S}$ and suppose, on a certified tube, an error bound and an objective Lipschitz condition hold:

$$\operatorname{dist}(\theta, \mathcal{C}_t(S)) \leq \kappa_S \left\| h_S(\theta) \right\|, \qquad |J(\theta) - J(\vartheta)| \leq L_J \left\| \theta - \vartheta \right\|. \tag{30}$$

Then for every compiled endpoint $\theta_S$ there exists $\bar{\theta}_S \in \mathcal{C}_t(S)$ with

$$J(\bar{\theta}_S) \leq J(\theta_S) + L_J \kappa_S \left\| h_S(\theta_S) \right\|. \tag{31}$$

The right-hand correction is charged to $\varepsilon_{\mathrm{sol}}$ in Theorem 11; it is not silently treated as exact feasibility.

The cross-credited archive always contains the checkpoint. For a finite discovered set $\mathcal{D}_B$, define

$$\widehat{V}^{\mathrm{cc}}_{t,B}(S) = \min\{J(\theta) : \theta \in \mathcal{D}_B \cup \{\theta_t\},\ \theta \in \mathcal{C}_t(S)\}. \tag{32}$$

If $S \subseteq T$, every endpoint feasible for $T$ is feasible for $S$, hence $\widehat{V}^{\mathrm{cc}}_{t,B}(S) \leq \widehat{V}^{\mathrm{cc}}_{t,B}(T)$.

## D No-free results

*Verification of Equation* (6). Direct substitution gives zero at $\varnothing$, at all four singletons, at each co-singleton, and at $[4]$. At $S = \{1, 2\}$, only the first pair term is active, giving minus one. The symmetric pair matrix is the signed adjacency of an alternating four-cycle and has rank two. Multiplication by any scalar preserves all shell values and scales the hidden deviation. □

Every Boolean table $G : 2^{[P]} \to \mathbb{R}$ has the multilinear polynomial extension

$$\widetilde{G}(x) = \sum_{A \subseteq [P]} \mu_A \prod_{i \in A} x_i,$$

so smoothness or analyticity alone cannot couple unobserved vertices to the shell.

### D.1 STRONGLY CONVEX ELIMINATION CREATES ALL INTERACTION ORDERS

For $S \subseteq [P]$ and $\lambda > 0$, consider

$$G_\lambda(S) = \min_{u \in \mathbb{R}^S} \left\{ \frac{1}{2} \left( 1 - \sum_{i \in S} u_i \right)^2 + \frac{\lambda}{2} \sum_{i \in S} u_i^2 \right\}. \tag{33}$$

The Hessian is $\mathbf{1}\mathbf{1}^\top + \lambda I \succ 0$. If $m = |S| > 0$, symmetry and the first-order condition give $u_i = (\lambda + m)^{-1}$ and

$$G_\lambda(S) = \frac{\lambda}{2(\lambda + m)}, \qquad G_\lambda(\varnothing) = \frac{1}{2}. \tag{34}$$

After subtracting the empty-set value, the cardinality function is $-m/[2(\lambda + m)]$. Its $k$th forward difference is nonzero for every $k \le P$ because

$$\Delta^k \frac{1}{\lambda + m} = \frac{(-1)^k k!}{\prod_{j=0}^{k} (\lambda + m + j)}. \tag{35}$$

Thus route-dependent elimination of a strongly convex quadratic can produce Boolean Möbius coefficients of every order. The regression code uses Equation (34), including the correct empty-set normalization.

## E BOOLEAN INTERACTION AND CONTRACTIVE NEURAL BOUNDS

*Proof of Theorem 3.* Let $\widetilde{F}(x) = \sum_A \mu_A \prod_{i \in A} x_i$ be the multilinear extension of the vertex table. For distinct $i, j, k$,

$$\partial_{ijk} \widetilde{F}(x) = \sum_{T \subseteq [P] \setminus \{i,j,k\}} \Delta_i \Delta_j \Delta_k F(T) \prod_{r \in T} x_r \prod_{r \notin T \cup \{i,j,k\}} (1 - x_r). \tag{36}$$

The weights are nonnegative and sum to one. Three applications of the fundamental theorem of calculus give

$$\Delta_i \Delta_j \Delta_k F(T) = \int_{[0,1]^3} \partial_{ijk} f(\mathbf{1}_T + s_i e_i + s_j e_j + s_k e_k)\, ds, \tag{37}$$

so $|\partial_{ijk} \widetilde{F}(x)| \le M_{ijk}$. For fixed $S$, let $g(t) = \widetilde{F}(t \mathbf{1}_S)$. The degree-two Taylor polynomial of $g$ at zero is exactly $Q_2(S)$. Repeated-coordinate derivatives of a multilinear function vanish, and

$$g'''(t) = 6 \sum_{\{i,j,k\} \subseteq S} \partial_{ijk} \widetilde{F}(t \mathbf{1}_S).$$

Taylor's integral remainder and $\frac{1}{2} \int_0^1 (1 - t)^2 dt = 1/6$ therefore yield Equation (7). □

**Theorem 13** (Arbitrary-order Boolean mixed-derivative bound)**.** *Let $d \in \{0, \dots, P - 1\}$, and let $Q_d(S)$ be the degree-$d$ Boolean Möbius truncation of $F(S) = f(\mathbf{1}_S)$. For a set $A$ of distinct coordinates, write $\partial_A = \prod_{i \in A} \partial_i$. Then*

$$|F(S) - Q_d(S)| \le \sum_{\substack{A \subseteq S \\ |A| = d+1}} \sup_{x \in [0,1]^P} |\partial_A f(x)|. \tag{38}$$

*The constant one is sharp, as shown by a single $(d + 1)$-way monomial.*

*Proof.* Let $\widetilde{F}$ be the multilinear extension of the Boolean table and set $g(t) = \widetilde{F}(t \mathbf{1}_S)$. The degree-$d$ Taylor polynomial of $g$ at zero is exactly $Q_d(S)$. Taylor's integral remainder gives

$$F(S) - Q_d(S) = \frac{1}{d!} \int_0^1 (1 - t)^d g^{(d+1)}(t)\, dt.$$

Multilinearity eliminates repeated-coordinate derivatives, so

$$g^{(d+1)}(t) = (d+1)! \sum_{\substack{A \subseteq S \\ |A|=d+1}} \partial_A \widetilde{F}(t\mathbf{1}_S).$$

For each $A$, the derivative $\partial_A \widetilde{F}$ is a Bernstein convex combination of the finite differences $\Delta_A F(T)$. Repeated application of the fundamental theorem of calculus writes each finite difference as an integral of $\partial_A f$ over the corresponding face of the cube. Hence

$$|\partial_A \widetilde{F}(x)| \le \sup_{y \in [0,1]^P} |\partial_A f(y)|.$$

Finally,

$$\frac{(d+1)!}{d!} \int_0^1 (1-t)^d \, dt = 1,$$

which proves Equation (38). Equality holds for $f(x) = \prod_{i \in A} x_i$ with $|A| = d+1$ and $S = A$. □

*Proof of Equation* (8). For distinct $i, j, k$,

$$\begin{aligned}\partial_{ijk}(J \circ \Theta) = {} & D^3 J[\Theta_i, \Theta_j, \Theta_k] + D^2 J[\Theta_{ij}, \Theta_k] + D^2 J[\Theta_{ik}, \Theta_j] \\ & + D^2 J[\Theta_{jk}, \Theta_i] + DJ[\Theta_{ijk}].\end{aligned}$$

Apply the declared induced multilinear operator norms and the triangle inequality. □

### E.1 Aggregate contractive recurrences

Consider the additive selector compiler

$$z_r = G_r(z_{r-1}) + x_r u_r, \qquad \|DG_r\| \le q < 1, \quad \left\|D^2 G_r\right\| \le h, \quad \left\|D^3 G_r\right\| \le t, \quad \|u_r\| \le \rho. \tag{39}$$

Define aggregate derivative masses

$$A_r = \sum_{i \le r} \sup_x \|\partial_i z_r(x)\|, \tag{40}$$

$$B_r = \sum_{i,j \le r} \sup_x \|\partial_{ij} z_r(x)\|, \tag{41}$$

$$C_r = \sum_{i,j,k \le r} \sup_x \|\partial_{ijk} z_r(x)\|. \tag{42}$$

The sums are over ordered indices; this makes product-rule aggregation exact up to the displayed inequalities. Additive injection gives

$$A_r \le q A_{r-1} + \rho, \tag{43}$$

$$B_r \le q B_{r-1} + h A_{r-1}^2, \tag{44}$$

$$C_r \le q C_{r-1} + 3h A_{r-1} B_{r-1} + t A_{r-1}^3. \tag{45}$$

Starting from zero,

$$A_r \le a := \frac{\rho}{1-q}, \tag{46}$$

$$B_r \le b := \frac{h\rho^2}{(1-q)^3}, \tag{47}$$

$$C_r \le c := \frac{3h^2\rho^3}{(1-q)^5} + \frac{t\rho^3}{(1-q)^4}. \tag{48}$$

For scalar $J \circ z_P$ or vector states under induced norms, the sum of unordered distinct third-route derivatives is at most one sixth of the ordered aggregate. Combining this fact with the third-order chain rule yields

$$\sum_{i<j<k} M_{ijk} \le \frac{1}{6}\left(L_3 a^3 + 3L_2 ab + L_1 c\right). \tag{49}$$

Heterogeneous $(q_r, h_r, t_r, \rho_r)$ use the same forward recurrences without replacing them by global maxima. The bound remains depth-uniform because old route derivatives are geometrically attenuated before each new injection.

**Fixed-gate exact pairwise island.** If the entire selector cube stays in one certified activation cell, the compiled parameter or state is affine in $x$, the repairs enter additively, the head is affine, and the terminal objective is quadratic, then $f(x)$ is a polynomial of degree at most two. Stable gates alone are insufficient for a nonquadratic terminal objective.

## F STABLE SHELL INVERSION AND POSITIVE INTERPOLATION

*Proof of Theorem 4.* Write $M = M_{\mathcal{A}}$, $h = r_{\mathcal{A}} + e_{\mathcal{A}}$, and $Q = M^\top \Pi M$. Let $P_Q = Q^\dagger Q$ be the orthogonal projector onto $\mathrm{range}(Q) = \mathrm{range}(M^\top)$. Since

$$\widehat{\beta} = Q^\dagger M^\top \Pi (M\beta + h) = P_Q \beta + Q^\dagger M^\top \Pi h,$$

an identifiable query $\phi = \phi(S) \in \mathrm{range}(Q)$ satisfies $\phi^\top (P_Q \beta - \beta) = 0$. We then have

$$\phi^\top (\widehat{\beta} - \beta) = \phi^\top Q^\dagger M^\top \Pi h.$$

We can write

$$\phi^\top Q^\dagger M^\top \Pi h = (Q^{\dagger/2}\phi)^\top (Q^{\dagger/2} M^\top \Pi h).$$

Thus, by Cauchy–Schwarz,

$$\begin{aligned} \left|\phi^\top (\widehat{\beta} - \beta)\right| &\le \|Q^{\dagger/2}\phi\|_2 \|Q^{\dagger/2} M^\top \Pi h\|_2 \\ &= \sqrt{\phi^\top Q^\dagger \phi}\sqrt{h^\top \Pi M Q^\dagger M^\top \Pi h}. \end{aligned}$$

Now,

$$h^\top \Pi M Q^\dagger M^\top \Pi h = (\Pi^{1/2} h)^\top \left(\Pi^{1/2} M Q^\dagger M^\top \Pi^{1/2}\right) (\Pi^{1/2} h).$$

The matrix

$$P := \Pi^{1/2} M Q^\dagger M^\top \Pi^{1/2}$$

is the orthogonal projector onto $\mathrm{range}(\Pi^{1/2} M)$. Therefore,

$$(\Pi^{1/2} h)^\top P (\Pi^{1/2} h) \le \|\Pi^{1/2} h\|_2^2 = h^\top \Pi h = \|h\|_\Pi^2.$$

Consequently,

$$\left|\phi^\top (\widehat{\beta} - \beta)\right| \le \sqrt{\phi^\top Q^\dagger \phi}\, \|h\|_\Pi.$$

Using

$$\ell_{\mathcal{A}}^\dagger(S) = \sqrt{\phi^\top Q^\dagger \phi} \qquad \text{and} \qquad \|h\|_\Pi \le \varepsilon_{\mathrm{exec}} + \eta_{\mathcal{A}},$$

we obtain

$$\left|\phi^\top (\widehat{\beta} - \beta)\right| \le \ell_{\mathcal{A}}^\dagger(S)(\varepsilon_{\mathrm{exec}} + \eta_{\mathcal{A}}).$$

Finally, since $|r(S)| \le \eta_S$, adding the residual term gives

$$|\phi(S)^\top \widehat{\beta} - F(S)| \le \ell_{\mathcal{A}}^\dagger(S)(\varepsilon_{\mathrm{exec}} + \eta_{\mathcal{A}}) + \eta_S,$$

which proves Equation (11).

**Note:** If $\phi(S) \notin \mathrm{range}(M^\top)$, then there exists $v \in \ker M$ with $\phi(S)^\top v \neq 0$. Replacing $\beta$ by $\beta + cv$ leaves every shell observation unchanged while changing the query by $c\,\phi(S)^\top v$. Letting $|c| \to \infty$ proves that no finite uniform query bound is possible without further restrictions. □

*Proof of Theorem 5.* Let

$$\mathcal{V}_{\mathcal{A}} = \mathrm{span}\{\phi(A) : A \in \mathcal{A}_{\mathrm{KOLO}}\}.$$

**(i) Identification.** Suppose first that

$$\phi(S) \in \mathcal{V}_{\mathcal{A}}.$$

Then there exist coefficients $\{\gamma_A\}$ such that

$$\phi(S) = \sum_{A \in \mathcal{A}_{\mathrm{KOLO}}} \gamma_A \phi(A).$$

Therefore,

$$\begin{aligned} F_\beta(S) &= \phi(S)^\top \beta \\ &= \sum_A \gamma_A \phi(A)^\top \beta \\ &= \sum_A \gamma_A F_\beta(A). \end{aligned}$$

Hence, $F_\beta(S)$ is completely determined by the shell values.

Conversely, suppose that

$$\phi(S) \notin \mathcal{V}_{\mathcal{A}}.$$

Then there exists a vector $v \in \mathcal{V}_{\mathcal{A}}^{\perp}$ such that

$$\phi(A)^\top v = 0 \quad \text{for all } A \in \mathcal{A}_{\mathrm{KOLO}}, \qquad \phi(S)^\top v \neq 0.$$

For any $\beta$ and $t \in \mathbb{R}$, define

$$\widetilde{\beta} = \beta + tv.$$

Then, for every $A \in \mathcal{A}_{\mathrm{KOLO}}$,

$$F_{\widetilde{\beta}}(A) = \phi(A)^\top(\beta + tv) = F_\beta(A),$$

whereas

$$F_{\widetilde{\beta}}(S) = F_\beta(S) + t\phi(S)^\top v.$$

Thus, the shell values remain unchanged while the value at $S$ varies with $t$. Hence, $F_\beta(S)$ cannot be determined from the shell for every $\beta$.

**(ii) Pass preservation.** Suppose first that

$$\phi(S) \in \mathrm{conv}\{\phi(A) : A \in \mathcal{A}_{\mathrm{KOLO}}\}.$$

Then there exist weights $w_{S,A} \geq 0$ satisfying

$$\sum_A w_{S,A} = 1, \qquad \phi(S) = \sum_A w_{S,A}\phi(A).$$

Consequently,

$$F_\beta(S) = \sum_A w_{S,A} F_\beta(A).$$

Therefore, if

$$F_\beta(A) \geq c \qquad \text{for all } A \in \mathcal{A}_{\mathrm{KOLO}},$$

then

$$F_\beta(S) \geq \sum_A w_{S,A} c = c.$$

Thus, the convex-hull condition is sufficient for pass preservation.

Conversely, suppose that

$$\phi(S) \notin \mathrm{conv}\{\phi(A) : A \in \mathcal{A}_{\mathrm{KOLO}}\}.$$

Since the convex hull of the finite set of shell feature vectors is closed and convex, the separating hyperplane theorem implies that there exist $v$ and $c$ such that

$$v^\top \phi(A) \geq c \qquad \text{for all } A \in \mathcal{A}_{\mathrm{KOLO}},$$

while

$$v^\top \phi(S) < c.$$

Choosing $\beta = v$, we obtain

$$F_\beta(A) \geq c \qquad \text{for all } A \in \mathcal{A}_{\mathrm{KOLO}},$$

but

$$F_\beta(S) < c.$$

Thus, pass preservation for every $\beta$ and $c$ requires

$$\phi(S) \in \operatorname{conv}\{\phi(A) : A \in \mathcal{A}_{\mathrm{KOLO}}\}.$$

Finally, suppose

$$F(T) = F_\beta(T) + r(T), \qquad |r(T)| \le \eta_T,$$

and

$$y_A = F(A) + e_A, \qquad |e_A| \le \varepsilon_A.$$

Using

$$F_\beta(S) = \sum_A w_{S,A} F_\beta(A),$$

we obtain

$$\begin{aligned} F(S) &= F_\beta(S) + r(S) \\ &= \sum_A w_{S,A}\big(F(A) - r(A)\big) + r(S) \\ &= \sum_A w_{S,A}\big(y_A - e_A - r(A)\big) + r(S). \end{aligned}$$

Since

$$e_A \le \varepsilon_A, \qquad r(A) \le \eta_A, \qquad r(S) \ge -\eta_S,$$

it follows that

$$F(S) \ge \sum_A w_{S,A} y_A - \sum_A w_{S,A}(\varepsilon_A + \eta_A) - \eta_S,$$

which proves (13). □

**Corollary 14** (KOLO is not universally pass preserving even for modular routes)**.** *Let $P \ge 4$ and let*

$$\phi_1(S) = \big(1, x_1(S), \dots, x_P(S)\big)$$

*be the constant-plus-unary feature map. Consider the KOLO shell consisting of the zero route, all singletons, and all co-singletons. Then every genuine middle route $S$ with*

$$2 \le |S| \le P - 2$$

*satisfies*

$$\phi_1(S) \notin \operatorname{conv}\{\phi_1(A) : A \in \mathcal{A}_{\mathrm{KOLO}}\}.$$

*The same conclusion holds if the full route is also observed. Consequently, raw KOLO-shell passage cannot universally imply middle-route passage even for modular route functions.*

*Proof of Corollary 14.* Fix a middle route $S$ with

$$2 \le |S| \le P - 2.$$

Since $P - |S| \ge 2$, there exist distinct $a, b \notin S$, and hence

$$x_a(S) = x_b(S) = 0.$$

Suppose, for contradiction, that

$$\phi_1(S) = \sum_{A \in \mathcal{A}_{\mathrm{KOLO}}} w_A \phi_1(A), \qquad w_A \ge 0, \qquad \sum_A w_A = 1.$$

Looking at the $a$-th unary coordinate gives

$$0 = x_a(S) = \sum_A w_A x_a(A).$$

Since $w_A \ge 0$ and $x_a(A) \in \{0, 1\}$, every route with $w_A > 0$ must satisfy

$$x_a(A) = 0.$$

Similarly, from $x_b(S) = 0$,

$$w_A > 0 \quad \Longrightarrow \quad x_b(A) = 0.$$

Thus, every route with positive weight must avoid both $a$ and $b$. Among the zero route, singletons, and co-singletons, only the zero route and singletons that avoid $a$ and $b$ have this property. In particular, every route with positive weight has cardinality at most one. Therefore,

$$\begin{aligned} |S| &= \sum_{i=1}^{P} x_i(S) \\ &= \sum_{A} w_A \sum_{i=1}^{P} x_i(A) \\ &= \sum_{A} w_A |A| \le \sum_{A} w_A = 1. \end{aligned}$$

This contradicts $|S| \ge 2$. Hence

$$\phi_1(S) \notin \operatorname{conv}\{\phi_1(A) : A \in \mathcal{A}_{\mathrm{KOLO}}\}.$$

If the full route is also observed, the argument is unchanged: the full route contains both $a$ and $b$, so it cannot receive positive weight. □

**Corollary 15** (Nonnegative pair interactions)**.** *Suppose* $F(S) = c + \sum_i a_i x_i + \sum_{i<j} b_{ij} x_i x_j + r(S)$, $F(\varnothing) = J_t$, $F(\{i\}) \ge J_t - \tau_i$, $b_{ij} \ge 0$, *and* $|r(S)| \le \eta_S$ *with* $r(\varnothing) = r(\{i\}) = 0$. *Then*

$$F(S) \ge J_t - \sum_{i \in S} \tau_i - \eta_S.$$

*In particular, exact singleton passage with zero tail is pass-preserving. The complement statement applies to nonpositive interactions when the full route and all co-singletons are measured relative to the same threshold.*

## G KNOWN-GRAPH PAIRWISE RECOVERY

*Proof of Theorem 6.* Route zero gives

$$c = F(\varnothing),$$

and the singleton routes give

$$a_i = F(\{i\}) - c.$$

Let

$$B_1 = \sum_{e \in E} b_e, \qquad d_i = \sum_{e \ni i} b_e.$$

For the co-singleton route $V \setminus \{i\}$,

$$F(V \setminus \{i\}) = c + \sum_{j \ne i} a_j + B_1 - d_i.$$

Hence the shell determines

$$q_i := B_1 - d_i.$$

Since every edge is incident to exactly two vertices,

$$\sum_i d_i = 2B_1,$$

and therefore

$$\sum_i q_i = (P-2)B_1.$$

Because $P \ge 3$,

$$B_1 = \frac{1}{P-2} \sum_i q_i, \qquad d_i = B_1 - q_i.$$

Thus the remaining pair coefficients satisfy the unsigned incidence system

$$Nb = d,$$

where $N$ is the vertex–edge incidence matrix.

For a connected component $C$, the real rank of its unsigned incidence matrix is

$$\operatorname{rank}(N_C) = \begin{cases} |V_C| - 1, & C \text{ bipartite}, \\ |V_C|, & C \text{ nonbipartite}. \end{cases}$$

Hence, if $b(G)$ denotes the number of bipartite connected components,

$$\operatorname{rank}(N) = |V| - b(G), \qquad \nu(G) := \dim \ker N = |E| - |V| + b(G).$$

Therefore $\nu(G) = 0$ precisely when every bipartite component has $|E_C| = |V_C| - 1$ and every nonbipartite component has $|E_C| = |V_C|$. Since the components are connected, the former are trees, while the latter are unicyclic; being nonbipartite, their unique cycles are odd. Thus the shell identifies the pair coefficients exactly iff every component is a tree or odd-unicyclic.

It remains to characterize the minimum number of additional pair measurements. After subtracting the already known constant and unary terms, querying the pair route $\{u, v\}$ measures the coordinate functional

$$\ell_e(b) = b_e, \qquad e = \{u, v\}.$$

The shell determines $b$ only up to the null space $\mathcal{N} = \ker N$. Hence a collection $R \subseteq E$ of additional pair routes identifies $b$ iff

$$z \in \mathcal{N}, \quad z_e = 0 \quad \forall e \in R \quad \Longrightarrow \quad z = 0.$$

Equivalently, the restrictions $\{\ell_e|_{\mathcal{N}} : e \in R\}$ must span $\mathcal{N}^*$. Since

$$\dim \mathcal{N}^* = \dim \mathcal{N} = \nu(G),$$

at least $\nu(G)$ additional scalar measurements are necessary, and a rank-revealing choice of $\nu(G)$ pair routes is sufficient.

In particular, an even-unicyclic component has a one-dimensional null space generated by the alternating edge vector

$$(1, -1, 1, -1, \ldots, 1, -1).$$

Any pair query on an edge with nonzero coefficient removes this null mode. More generally, one suitable pair query per independent null direction gives the claimed minimum repair count. □

**Corollary 16** (Exact conditioning of path and odd-cycle interaction recovery)**.** *Let $N_G$ be the unsigned vertex–edge incidence matrix used to recover the pair coefficients in Theorem 6. For the path and the odd cycle on $n$ vertices,*

$$\sigma^+_{\min}(N_G) = 2 \sin\left(\frac{\pi}{2n}\right).$$

*Consequently,*

$$\left\| N_G^\dagger \right\|_{\mathrm{op}} = \frac{1}{2 \sin(\pi/(2n))} \sim \frac{n}{\pi}.$$

*Thus, exact identifiability does not imply uniform numerical conditioning: the sensitivity of interaction recovery to noise grows linearly with the graph size.*

*Proof.* For the path, $N_G N_G^\top$ is diagonally similar to the ordinary path Laplacian, whose nonzero eigenvalues are

$$4 \sin^2\left(\frac{k\pi}{2n}\right), \qquad k = 1, \ldots, n-1.$$

Hence

$$\sigma^+_{\min}(N_G) = 2 \sin\left(\frac{\pi}{2n}\right).$$

For the odd cycle, the eigenvalues of the signless Laplacian are

$$2 + 2 \cos\left(\frac{2\pi k}{n}\right), \qquad k = 0, \ldots, n-1.$$

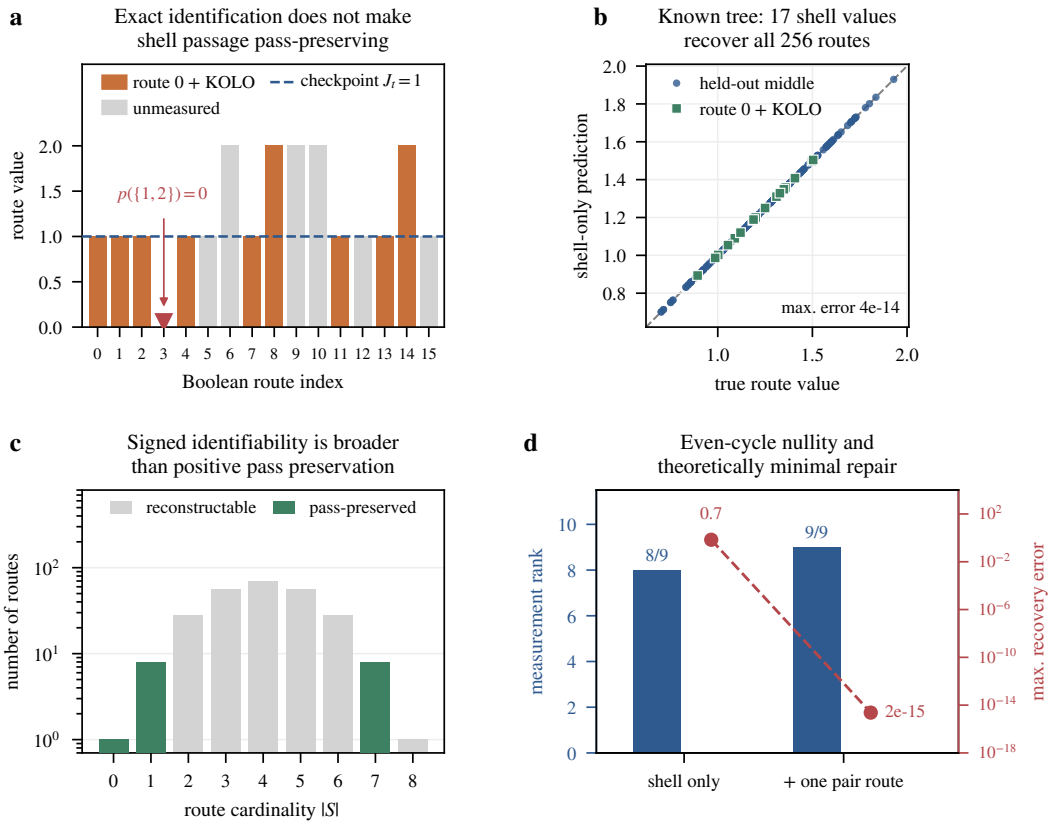



**Figure 2: What the shell can and cannot establish.** (a) The hidden-middle path example passes every measured shell route but fails at an unmeasured pair, proving that identification is not pass preservation. (b) A predeclared path-pair generator is reconstructed from route zero plus KOLO only; all other vertices are held-out truth. (c) In a smooth coupled compiler, the higher-order error and its analytic envelope scale cubically with route radius. (d) A four-cycle has one invisible interaction mode; one rank-revealing pair query removes it. Every panel is generated from the supplied scripts, and oracle pair queries are labeled separately from KOLO reconstruction.

Since $n$ is odd, the smallest eigenvalue is

$$4\sin^2\left(\frac{\pi}{2n}\right),$$

which yields the same value of $\sigma^+_{\min}(N_G)$. Therefore,

$$\left\|N_G^\dagger\right\|_{\mathrm{op}} = \frac{1}{\sigma^+_{\min}(N_G)} = \frac{1}{2\sin(\pi/(2n))} \sim \frac{n}{\pi}.$$

□

## H Representation Semantics

*Proof of Theorem 7.* Add and subtract $K_i$ and $J_i^\star$:

$$J_t - J^\star = (J_t - K_i) + (K_i - J_i^\star) + (J_i^\star - J^\star).$$

Fallback, candidate feasibility, and constraint inclusion make the three terms nonnegative. If $L_0 \le J^\star$, then

$$K_i - L_0 = (K_i - J_i^\star) + (J_i^\star - J^\star) + (J^\star - L_0),$$

which proves Equation (17). □

**Log loss.** Bayes completeness from $\mathsf{X}$ and $\mathsf{Z}_i$ gives optimal risks $H(\mathsf{Y} \mid \mathsf{X})$ and $H(\mathsf{Y} \mid \mathsf{Z}_i)$. Since $\mathsf{Z}_i$ is measurable with respect to $\mathsf{X}$, the chain rule gives

$$H(\mathsf{Y} \mid \mathsf{Z}_i) - H(\mathsf{Y} \mid \mathsf{X}) = I(\mathsf{Y}; \mathsf{X} \mid \mathsf{Z}_i).$$

For a particular complete candidate representing conditional distribution $q_i(\cdot \mid \mathsf{Z}_i)$ that dominates the true conditional law,

$$K_i - H(\mathsf{Y} \mid \mathsf{Z}_i) = \mathbb{E}\mathrm{KL}(P_{\mathsf{Y}|\mathsf{Z}_i} \| q_i),$$

possibly $+\infty$ if domination or integrability fails.

**Squared loss.** Orthogonal projection in $L_2$ yields

$$\mathbb{E}\,\|\mathsf{Y} - g(\mathsf{Z}_i)\|^2 = \mathbb{E}\,\|\mathsf{Y} - \mathbb{E}[\mathsf{Y} \mid \mathsf{Z}_i]\|^2 + \mathbb{E}\,\|g(\mathsf{Z}_i) - \mathbb{E}[\mathsf{Y} \mid \mathsf{Z}_i]\|^2 .$$

Applying the same identity to $\mathsf{X}$ and using the tower property gives

$$D_i^{\mathrm{rep}} = \mathbb{E}\,\|\mathbb{E}[\mathsf{Y} \mid \mathsf{X}] - \mathbb{E}[\mathsf{Y} \mid \mathsf{Z}_i]\|^2 .$$

If either prediction class is restricted, its Bayes approximation term must be written separately and need not have a favorable sign without a declared class inclusion.

# I EXACT AND FINITE KO/LO GEOMETRY

*Proof of Proposition 8.* If $i \in S$, every parameter satisfying all equalities in $S$ also satisfies the equality at $i$, so $\mathcal{C}_t(S) \subseteq \mathcal{C}_t(\{i\})$. Taking infima proves the value inequality. If all singleton values are at least $J_t - \tau$, every nonempty anchored set contains some singleton and inherits the same lower bound. Every co-singleton is nonempty for $P \geq 2$. □

**Proposition 17** (Downstream shielding)**.** *Suppose a serial model admits the complete-cut factorization Equation* (29) *at a retained downstream cut $j$, and the objective is independent of released prefix parameters conditional on the retained state. Releasing any earlier hard equality while retaining the cut-$j$ state leaves the exact constrained value unchanged.*

*Proof.* Every feasible suffix sees the same retained cut state, and the checkpoint prefix is a feasible realization of that state. Enlarging the prefix fiber therefore changes neither the attainable suffix predictions nor the objective infimum. □

*Proof of the reversal criterion.* Subtract the aggregate definitions:

$$\widehat{V}_B^{\mathrm{LO}} - \widehat{V}_B^{\mathrm{KO}} = (V^{\mathrm{LO}} - V^{\mathrm{KO}}) + (\varepsilon_B^{\mathrm{LO}} - \varepsilon_B^{\mathrm{KO}}).$$

Exact dominance makes the first difference nonnegative. Rearrangement gives Equation (19). □

*Proof of the matched-budget family.* At checkpoint $(0, 0, 0)$, $\mathrm{KO}_x$ fixes $x = 0$ and has optimum $1/2$. $\mathrm{KO}_y$ fixes $y = 0$ and sets $x = 1, z = M^{-1}$, giving $1/(2M)$; $\mathrm{KO}_z$ is symmetric. Hence $V^{\mathrm{KO}} = 1/(2M)$. $\mathrm{LO}_x$ fixes $y = z = 0$ and sets $x = 1$, giving $1/M$. The other LO routes have value at least $1/2 + 1/(2M)$, so $V^{\mathrm{LO}} = 1/M$.

On the winning LO affine space, only the unit-curvature $x$ coordinate is free, so exact-line-search steepest descent reaches its optimum in one step. On the best KO affine space, the free quadratic has eigenvalues 1 and $M$. From the checkpoint, its objective error contracts by

$$r_M^2 = \left(\frac{M-1}{M+1}\right)^2$$

per iteration. Therefore

$$\widehat{V}_B^{\mathrm{KO}} = \frac{1}{2M} + \frac{1}{2}(1 + M^{-1}) r_M^{2B}, \qquad \widehat{V}_B^{\mathrm{LO}} = \frac{1}{M}.$$

With $M = n^2$ and $B = n$, Bernoulli's inequality gives $1 - r_M^{2n} \leq 2n(1 - r_M) = 4n/(M+1)$, hence

$$J_t - \widehat{V}_B^{\mathrm{KO}} = \frac{1}{2}(1 + M^{-1})(1 - r_M^{2n}) \leq \frac{2n}{M} = \frac{2}{n}.$$

The LO gain is exactly $1/2$. □

*Simulation and cross-credit statements.* Apply the assumed simulation to the best budget-$B$ LO transcript to obtain $\widehat{V}_{cB}^{\mathrm{KO}} \le \widehat{V}_B^{\mathrm{LO}} + \delta$. For cross-crediting, if $S \subseteq T$, the endpoint set satisfying $T$ is contained in the endpoint set satisfying $S$, so the cross-credited minimum is monotone. □

**Corollary 18** (Exact budget at which native KO overtakes LO)**.** *In the matched quadratic family of Theorem 9, let*

$$r_M = \frac{M-1}{M+1}, \qquad M > 1.$$

*The best finite values after $B$ exact-line-search iterations are*

$$\widehat{V}_B^{\mathrm{KO}} = \frac{1}{2M} + \frac{1}{2}\left(1 + \frac{1}{M}\right) r_M^{2B}, \qquad \widehat{V}_B^{\mathrm{LO}} = \frac{1}{M}.$$

*Hence LO is strictly better exactly when*

$$B < B_\times(M) := -\frac{\log(M+1)}{2\log((M-1)/(M+1))}.$$

*As $M \to \infty$,*

$$B_\times(M) = \frac{M\log M}{4} + \frac{1}{4} - \frac{2\log M + 3}{24M} + O\left(\frac{\log M}{M^2}\right).$$

*Proof.* The displayed finite values follow from the two-eigenvalue exact-line-search contraction. The inequality $\widehat{V}_B^{\mathrm{LO}} < \widehat{V}_B^{\mathrm{KO}}$ is equivalent to

$$1 < (M+1) r_M^{2B}.$$

Since $0 < r_M < 1$, taking logarithms gives precisely $B < B_\times(M)$. The expansion follows from

$$\log r_M = -2\left(M^{-1} + \frac{1}{3M^3} + \frac{1}{5M^5} + \cdots\right).$$

□

## J LOCAL COLLECTIVE ESCAPE

*Proof of Theorem 10.* For (i), each $T_i$ is a linear subspace. Applying the hypothesis to both $v$ and $-v$ gives $g^\top v = 0$ on $T_i$, hence $g \in T_i^\perp = \mathrm{range}(A_i^\top)$ for every $i$.

For (ii), if $v^\top H v < 0$ and $A_i v = 0$ for some $i$, then $v \in T_i$, contradicting restricted nonnegative curvature. Let $N$ be the negative eigenspace of $H$. Since $N \cap \ker A_i = \{0\}$, the restriction $A_i|_N$ is injective and $\dim N \le \mathrm{rank}(A_i)$ for every $i$.

For sharpness, let $r = \min_i \mathrm{rank}(A_i)$. A generic $r$-dimensional subspace $N$ is disjoint from every $T_i$, because $r + \dim T_i \le d$ and the family is finite. For each $i$, compactness of the unit sphere in $T_i$ and $N \cap T_i = \{0\}$ give a constant

$$\gamma_i = \inf_{v \in T_i,\ \|v\|=1,\ P_N v \neq 0} \frac{\|P_{N^\perp} v\|^2}{\|P_N v\|^2} > 0,$$

with the convention that vectors with $P_N v = 0$ impose no restriction. Choose $0 < \epsilon \le \min_i \gamma_i$ and set

$$H = P_{N^\perp} - \epsilon P_N.$$

Then $H$ has exactly $r$ negative eigenvalues, while for every $v \in T_i$, $v^\top H v = \|P_{N^\perp} v\|^2 - \epsilon \|P_N v\|^2 \ge 0$.

For (iii), an LO$_i$ tangent vector lies in $\bigcap_{j \neq i} T_j$, hence in every retained KO tangent. Define

$$\chi(\mathcal{T}) = \max_{\pi \in \Delta(\mathcal{T})} \lambda_{\min}\left(\sum_{T \in \mathcal{T}} \pi_T P_T\right).$$

If $T_{-i} \subseteq T_j$, then $P_{T_{-i}} \preceq P_{T_j}$. Any KOLO mixture can move the weight of an LO projector to one retained KO projector without decreasing the positive-semidefinite sum. Thus the KOLO optimum is no larger than the KO optimum; the reverse inequality follows because KO is a subfamily. □

**Corollary 19** (Quantitative visibility of a hidden mode)**.** *Assume* $\|H\|_{\mathrm{op}} \le L$*. Let* $\|u\| = 1$ *satisfy* $u^T H u \le -\gamma$*, and suppose* $v^T H v \ge -\epsilon_i \|v\|^2$ *on* $T_i = \ker A_i$*, with* $\gamma > \epsilon_i$*. Define*

$$d_i = \frac{\gamma - \epsilon_i}{2L}, \qquad \psi(d) = d\sqrt{1 - \frac{d^2}{4}}.$$

*Then*

$$\operatorname{dist}(u, T_i) \ge \psi(d_i).$$

*If* $\sigma_i^+$ *is the smallest positive singular value of* $A_i$*, then*

$$\|A_i u\| \ge \sigma_i^+ \psi(d_i).$$

*Thus a sufficiently negative hidden mode must be quantitatively visible at every audited cut, not merely nonzero there.*

*Proof of Corollary 19.* Fix $i$ and write $T = T_i$. If $P_T u = 0$, then $\operatorname{dist}(u, T) = 1$ and the result is immediate. Otherwise set $v = P_T u / \|P_T u\|$, so $v \in T$ and $\|v\| = 1$. The Rayleigh quotient is $2L$-Lipschitz on the unit sphere:

$$|v^T H v - u^T H u| \le \|v - u\| \|Hv\| + \|Hu\| \|v - u\| \le 2L \|v - u\|.$$

Since $v^T H v \ge -\epsilon_i$ and $u^T H u \le -\gamma$,

$$\|u - v\| \ge d_i := \frac{\gamma - \epsilon_i}{2L}.$$

Let $b = \operatorname{dist}(u, T)$. Because $v$ is the normalized projection,

$$\|u - v\|^2 = 2\left(1 - \sqrt{1 - b^2}\right),$$

or equivalently

$$b = \delta\sqrt{1 - \delta^2/4}, \qquad \delta = \|u - v\|.$$

The right-hand side is increasing for $0 \le \delta \le \sqrt{2}$; moreover $d_i \le 1/2$ because $\gamma \le L$. Hence $b \ge \psi(d_i)$. Finally, $A_i u = A_i P_{T^\perp} u$, and the smallest positive singular value of $A_i$ gives

$$\|A_i u\| \ge \sigma_i^+ \|P_{T^\perp} u\| = \sigma_i^+ \operatorname{dist}(u, T).$$

□

**Approximate first order.** For weights $\pi \in \Delta_P$, let $S_\pi = \sum_i \pi_i P_{T_i}$ and assume $\chi = \lambda_{\min}(S_\pi) > 0$. If $\|P_{T_i} g\| \le \epsilon_i$, then

$$\|g\|^2 \le \frac{\sum_i \pi_i \epsilon_i^2}{\chi}.$$

**Nonlinear equality manifolds.** If $h_i^{-1}(0)$ is a smooth manifold and $Dh_i$ has constant rank, first-order feasibility is $v \in \ker A_i$. At a constrained stationary point with multiplier $\lambda_i$, the second-order condition uses

$$v^\top \nabla^2_{\theta\theta}(J + \lambda_i^\top h_i) v \ge 0 \quad \forall v \in \ker A_i.$$

The main theorem uses the raw Hessian only for affine constraints or for the explicitly stated linearized quadratic model.

## K CORRECT ROUTE-TO-GLOBAL TRANSFER

*Proof of Theorem 11.* The routewise interval gives

$$F_{\mathrm{comp}}(S) \ge \widehat{F}(S) - E_{\mathrm{rec}}(S) = \underline{F}(S) \qquad \forall S.$$

Hence

$$F^\star_{\mathrm{comp}} = \min_S F_{\mathrm{comp}}(S) \ge \min_S \underline{F}(S),$$

and therefore

$$J_t - F^\star_{\mathrm{comp}} \le \Big[J_t - \min_S \underline{F}(S)\Big]_+ = \tau_{\mathrm{LCB}}.$$

Now telescope the directed minima:

$$\begin{aligned} J_t - J^\star = (J_t - F^\star_{\mathrm{comp}}) + (F^\star_{\mathrm{comp}} - V^\star_{\mathrm{route}}) \\ + (V^\star_{\mathrm{route}} - V^\star_{\mathrm{target}}) + (V^\star_{\mathrm{target}} - J^\star). \end{aligned}$$

Apply Equations (23) to (25).

For comparison with the simpler uniform-error expression, let $E_\star = \max_S E_{\mathrm{rec}}(S)$. Since

$$\min_S(\widehat{F}(S) - E_{\mathrm{rec}}(S)) \ge \min_S \widehat{F}(S) - E_\star,$$

we have

$$\tau_{\mathrm{LCB}} \le \tau_{\mathrm{rec}} + E_\star.$$

Substituting the routewise bound from Theorem 4 gives Equation (28). □

The counterexample in the main text shows why replacing $\tau_{\mathrm{rec}}$ by raw shell passage is invalid under signed interpolation. Theorem 5 gives the exact condition under which that replacement is permitted for a linear generator.

## L Deep-linear prediction images

Let $X \in \mathbb{R}^{d_0 \times n}$ and consider predictions $W_L \cdots W_1 X$. Define

$$\rho = \min\{\mathrm{rank}(X), d_1, \ldots, d_{L-1}, d_L\}.$$

If $H = W_i^t \cdots W_1^t X$ is retained and the remaining widths impose rank limit $s$, define $\mathcal{P}_X^\rho = \{Q : \mathrm{row}(Q) \subseteq \mathrm{row}(X), \mathrm{rank}(Q) \le \rho\}$ and $\mathcal{P}_H^s = \{Q : \mathrm{row}(Q) \subseteq \mathrm{row}(H), \mathrm{rank}(Q) \le s\}$.

**Lemma 20** (Prediction images)**.** *The full deep-linear network realizes exactly $\mathcal{P}_X^\rho$, and the exact KO class at $H$ realizes exactly $\mathcal{P}_H^s$.*

*Proof.* Every product prediction has row space contained in the row space of its input and rank bounded by the narrowest factor. Conversely, if $Q$ satisfies the corresponding row-space and rank conditions, write $Q = AX$ (or $Q = BH$) and take a rank factorization of $A$ (or $B$). The factors can be embedded through the declared widths by zero padding and identity blocks. □

*Proof of Theorem 12.* If $\mathcal{P}_H^s = \mathcal{P}_X^\rho$, the earliest-KO and full prediction infima coincide for every output-only objective. Exact passage means the checkpoint attains the KO infimum and therefore the full infimum. Conversely, if the images differ, choose a prediction $Q_0 \in \mathcal{P}_X^\rho \setminus \mathcal{P}_H^s$ and an output-only objective uniquely minimized at $Q_0$, for example squared distance to $Q_0$. The full class attains zero while the KO class cannot, proving target-uniform necessity. □

*Proof of the squared-loss statement in Theorem 12.* The Eckart–Young–Mirsky theorem gives

$$J^\star = \tfrac{1}{2}\left(\|Y\|_F^2 - E_X\right), \qquad J_H^\star = \tfrac{1}{2}\left(\|Y\|_F^2 - E_H\right).$$

Here $\|Y\|_F$ means $\|Y\|_F$; the source uses the latter notation in displayed results. Exact KO passage gives $J_t = J_H^\star$, and subtraction proves the displayed defect formula in Theorem 12. Tied singular values cause no ambiguity because $E_X$ and $E_H$ are sums of optimal spectral energies; target-specific completeness is exactly equality of these values. □

Factor-dependent regularization invalidates prediction-image reduction because two factorizations of the same output can have different objective values. We make no unrestricted deep-affine claim in this paper.

**Table 2: Deterministic theorem regressions.** These checks validate implementations of the stated constructions; they are not architecture-facing evidence.

| Check | Result |
|---|---|
| Hidden-middle tree | All route-zero/KO/LO measurements are $\geq 1$ while the held-out pair $\{1, 2\}$ has value $0$. |
| Known-path shell inversion | Route zero plus KOLO recovers the full declared pairwise path route table; non-shell vertices are held out during fitting. |
| Even-cycle repair | Shell rank $8/9$; one rank-revealing pair query gives rank $9/9$ and machine-precision recovery. |
| Matched-budget reversal | At $n = 128$, finite LO gain is $0.5$ while finite KO gain is approximately $0.0154$, although exact KO is better. |
| Contractive cube | The observed higher-order tail follows the predicted cubic small-radius scaling and remains below the independently computed derivative envelope. |

## M TRAINER-RELATIVE DEPTH SEPARATION

Consider

$$J(w) = \frac{1}{2}\left(\prod_{r=1}^{K} w_r - 1\right)^2, \qquad w \in [0, \alpha^{-1}]^K, \qquad 0 < \alpha \leq \frac{1}{4}, \qquad K > 2.$$

Balanced projected gradient descent remains balanced, $w_r^{(n)} = s_n$, and before the coordinates reach $2\alpha$ the projection is inactive:

$$s_{n+1} - s_n = \eta(1 - s_n^K)s_n^{K-1}.$$

Let $\varphi(s) = s^{2-K}$. By the mean-value theorem, for some $\xi_n \in [s_n, s_{n+1}]$,

$$\varphi(s_n) - \varphi(s_{n+1}) = (K-2)\xi_n^{1-K}(s_{n+1} - s_n) \leq (K-2)\eta(1 - s_n^K) \leq (K-2)\eta(1 - \alpha^K).$$

Summing until the first iterate at or above $2\alpha$ gives the strengthened lower bound

$$N_{2\alpha} \geq \frac{1 - 2^{2-K}}{(K-2)\eta(1-\alpha^K)}\alpha^{2-K}.$$

This improves the simpler bound by the factor $(1 - \alpha^K)^{-1}$.

For the decisive challenge, let $m = \lfloor K/2 \rfloor$ and retain the middle prefix product $Z_m = s^m$. Keep the prefix weights at $s$ and set each of the $K - m$ suffix weights to

$$s^{-m/(K-m)}.$$

The complete product is one. Since $m/(K-m) \leq 1$, $s \geq \alpha$, and $s \leq 2\alpha \leq 1$, each suffix weight is at most $\alpha^{-1}$. Thus the construction is feasible for every integer $K > 2$, not only even depths. Writing and validating the complete parameter vector costs $\Theta(K)$.

The comparison is relative to the declared trainer, initialization, projection box, and exact conditional construction. It is not a lower bound for all optimization algorithms and does not imply linear total work for proving passage of the full KOLO shell.

## N CONTROLLED THEOREM REGRESSIONS

The controlled checks are intended to falsify implementation mistakes and to reproduce exact constructions; they are not evidence of performance on a trained neural system. Figure 4 visualizes these checks.

## O ARCHITECTURE-FACING EXPERIMENT: COMPLETE RESULTS

The completed experiment uses the predeclared configuration in Table 3. All final values are computed on the held-out audit split after materializing a complete network. Route-native values preserve their

**Table 3:** Predeclared architecture-facing experiment configuration.

| Item | Value |
|---|---|
| Dataset | CIFAR10 |
| Audited boundaries | 8 |
| Base training epochs | 80 |
| Seeds | 0, 1, 2 |
| Checkpoint fractions | 0.25, 0.6, 1.0 |
| Run profile | paper |
| Full route lattice | $2^8 = 256$ routes |
| Compilers | task-shaped and state-faithful |
| Accelerator | NVIDIA A100-SXM4-40GB |

**Table 4:** Complete suite comparison across three seeds and three checkpoint stages. Headroom is normalized by the exhaustive 256-route optimum.

| Compiler | Suite | Routes | Mean | S.d. | Median | Exact |
|---|---|---|---|---|---|---|
| Task | KO | 8 | 88.5 | 18.3 | 100.0 | 5/9 |
| Task | LO | 8 | 11.1 | 33.3 | 0.0 | 1/9 |
| Task | KOLO | 17 | 95.6 | 11.9 | 100.0 | 7/9 |
| Task | $\|S\| \leq 2$ | 37 | 100.0 | 0.0 | 100.0 | 9/9 |
| Task | Full lattice | 256 | 100.0 | 0.0 | 100.0 | 9/9 |
| State | KO | 8 | 63.6 | 32.0 | 79.4 | 1/9 |
| State | LO | 8 | 33.7 | 33.9 | 42.2 | 0/9 |
| State | KOLO | 17 | 72.3 | 28.7 | 85.5 | 1/9 |
| State | $\|S\| \leq 2$ | 37 | 92.0 | 8.9 | 94.9 | 2/9 |
| State | Full lattice | 256 | 100.0 | 0.0 | 100.0 | 9/9 |

discovering compiler; the cross-credited archive separately reuses an endpoint in every hard-anchor class it satisfies. The full lattice is used only to evaluate sparse-suite strength and reconstruction accuracy, never to fit the KOLO shell model.

The compact experiment archive contains the per-run table, complete route JSON files, coupled-cube route values, native and cross-credited summaries, anchor residuals, wall-clock records, environment metadata, and a SHA–256 manifest. The main text reports means and paired behavior; the artifact preserves every route-level observation needed to recompute them.

## P ARCHITECTURE-FACING STUDY PROTOCOL

The companion Colab notebook records the predeclared experiment that produced the architecture-facing results in Section O. It mounts Google Drive and resumes after each training epoch, checkpoint cache, interval solve, route, coupled-cube vertex, and analysis condition. It uses held-out audit data for final route values and stores every complete endpoint, anchor residual, native provenance, cross-credit relation, wall-clock cost, seed, configuration, environment, and hash.

The completed paper profile uses a smooth spectrally constrained residual classifier with $P = 8$ on CIFAR–10. At early, middle, and late checkpoints, it evaluates route zero, KO, LO, KOLO, all $|S| \leq 2$ routes, and all $256$ routes. A separate frozen additive coupled compiler is fit from route zero plus shell observations only; all unexecuted routes are held-out truth. The artifact reports reconstruction error, routewise leverage, shell condition number, positive-interpolation feasibility, challenge headroom per route, and per-checkpoint results.

The reported run used an NVIDIA A100-SXM4-40GB accelerator. Configuration, software versions, hardware metadata, route-level outputs, and timing logs are included in the compact archive.

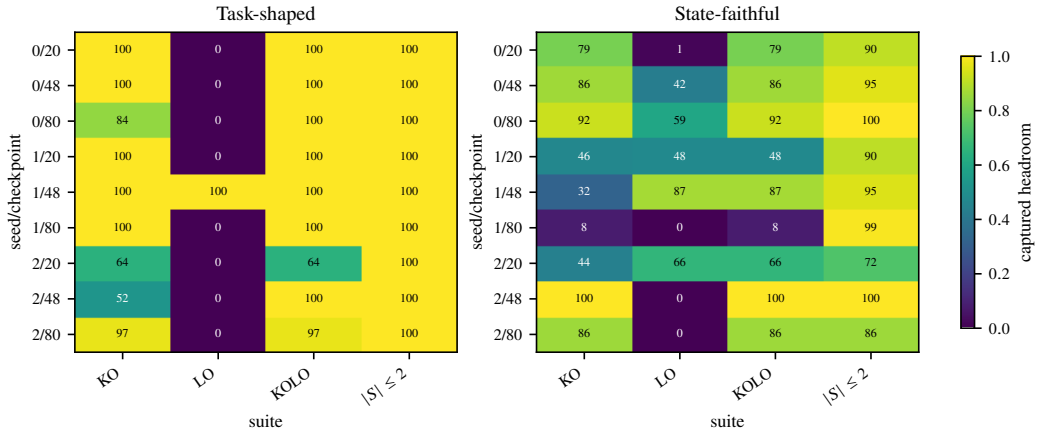



**Figure 3: The sparse-suite conclusion persists across seeds and training stages.** Each cell is the fraction of full-lattice headroom captured at one seed–checkpoint pair; numbers are percentages. The task-shaped compiler is nearly saturated by KOLO in most runs, whereas the state-faithful compiler exhibits genuine middle-route failures. This per-run view rules out an aggregate result driven by a single favorable checkpoint.

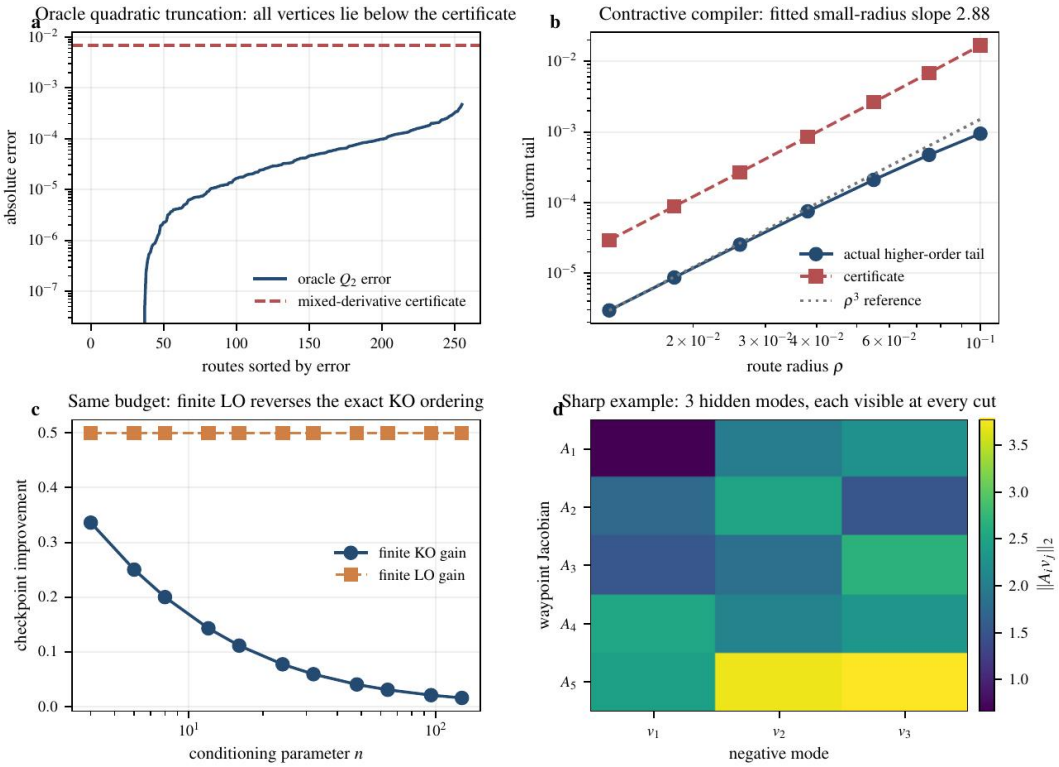



**Figure 4: Controlled regressions for KO/LO semantics.** The panels verify the KO value decomposition, exact-versus-finite reversal, monotone cross-crediting, and the principal-angle sharpness construction for collective escape. These are theorem checks, not substitutes for the trained-network experiment in Section 6.