NO FREE TOPOLOGY: MARGIN, SLOPE, AND ARCHITECTURAL COST IN DEEP CLASSIFIERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Continuous separability alone does not reveal the architectural cost of reliable classification. We study this cost for scalar piecewise-linear classifiers: which decision distinctions must be preserved, and what resources does their realization require? A confidence-weighted margin deficit links recovered sign alternations to a necessary slope budget. With independent uniform threshold noise after the final score, the deficit of the clipped score equals twice the excess classification risk. For dyadic targets with K boundaries, we obtain the exact cost $a \log_2 K + b\lceil \log_2(1/(2r)) \rceil$ at excess risk $0 < r < 1/8$ in a scalar catalog whose two-piece and affine modules cost $a \geq b > 0$ and each have Lipschitz constant at most two. For irregular regions with unequal confidence, joint piece-slope bounds screen architecture budgets before optimization. Across 15 selection tasks, screening preserves every minimum-cost catalog choice while avoiding 38.0% of dynamic-programming work. A complementary study uses finite diagnostics to choose repairs for information loss, metric compression, or insufficient resolution. On 270 clean synthetic tasks, it selects the same successful heads as cost-ordered search with half as many repair fits; label corruption and combined failures expose its limits. Additional MNIST and Fashion-MNIST experiments probe the distinction between score amplification and input-space separation in learned classifiers. Together, these results give a resource interpretation of metric-topology factorization: preserve task distinctions, then separate them economically. The guarantees depend on the resource catalog and downstream noise model; increasing gain does not improve normalized input margins.

1 INTRODUCTION

Representation learning seeks coordinates in which decisions become simple (Bengio et al., 2013; Cover, 1965). A single threshold in those coordinates can classify many disconnected input regions. However, its simplicity may depend on an encoder that changes rapidly between nearby regions with opposite labels. Evaluating only the readout misses part of the cost. To choose an architecture, we need to understand what the *complete classifier* must supply to separate these regions reliably. We study *how a reliability requirement constrains architectural resources*. Classical separation provides the starting point: Urysohn's lemma guarantees a continuous separator for disjoint closed sets in a normal space, but does not bound its slope or realization cost (Munkres, 2000, Sec. 33). Expressivity results explain how depth encodes many decision regions compactly (Montúfar et al., 2014; Telgarsky, 2016); margin and Lipschitz analyses address the sensitivity of separation (Bartlett et al., 2017; Sanford and Chatziafratis, 2022). The missing link is loss-sensitive resource accounting: which label regions must survive a reliability budget, and what must their implementation supply? Our contribution couples confidence-weighted recovery of decision transitions to a slope obstruction and explicit module costs. The folding construction and pushforward invariance are classical ingredients; the resulting exact catalog cost and irregular-region certificates price reliable realization.

Our central claim is that reliable separation imposes two simultaneous demands: enough decision transitions to retain important label regions, and enough slope to cross between them with adequate score margin. The tent map illustrates how these demands interact (Figure 1). One fold maps each half of the input interval onto the full interval; repeating this operation creates increasingly frequent oscillations. After L folds, the centered score crosses zero 2^L times, separating alternating label regions, and its slope magnitude is 2^L. Consider the gray score band $|f(x)| \leq \delta$, where $0 < \delta < 1/2$:

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

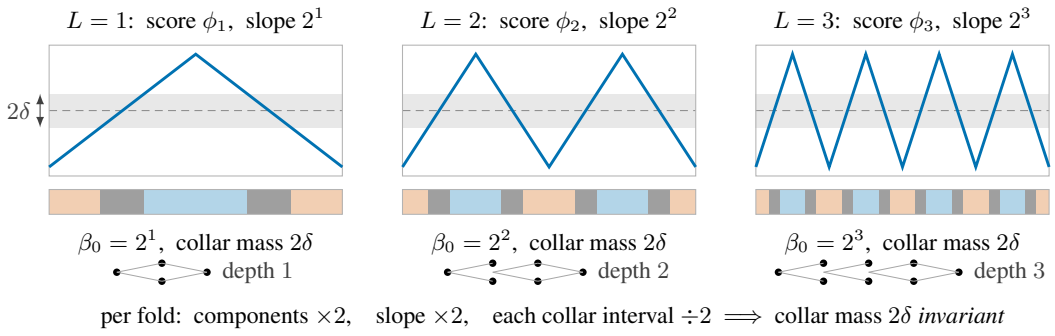


Figure 1: **The central cancellation.** After L folds, 2^L crossings have slope 2^L. The score band $|\phi_L| \leq \delta$ pulls back to intervals of width $2\delta/2^L$, preserving uniform mass 2δ. Network glyphs represent width-two, depth-L compositions: description length grows linearly while each geometric collar narrows exponentially.

these are inputs whose scores lie close to the decision threshold. Around each crossing, the band occupies an input interval of width $2\delta/2^L$. Each additional fold doubles the number of these intervals while halving their individual widths. Under uniform input, their total probability remains 2δ. Depth represents exponentially many decision transitions with linearly many folding operations, while the slope grows exponentially and the probability near the threshold stays unchanged. To give this threshold proximity an operational meaning, we consider fixed-scale threshold noise added after the final score. Correctly signed scores farther from zero are less likely to be flipped by this noise. Amplification can improve readout reliability, but its cost must be included alongside the cost of creating the transitions. With module costs specified, we can ask which architectures achieve a required readout error most economically.

Contributions. For scalar piecewise-linear classifiers, we derive simultaneous constraints on retained decision transitions and slope from a confidence-weighted margin deficit (Section 2). An explicit noisy-readout model converts the deficit into excess classification risk. Matching constructions give an exact minimum cost for dyadic targets in a bounded-gain scalar catalog; for irregular targets, joint piece-slope certificates exclude inadequate budgets (Section 3). We also quantify how score approximation and input-distribution shift perturb collar conservation, and show that sufficiently accurate approximations retain necessary resource demands (Section 4). The certificates support consequential resource choices: across 15 architecture-selection tasks they preserve every minimum-cost catalog choice while avoiding 38.0% of dynamic-programming work. A finite-diagnostic study selects the same successful repairs as cost-ordered search on 270 clean synthetic tasks with half as many repair fits; label corruption and mixed failures expose its limits (Section 5). Image-data depth comparisons extend the geometric diagnostics to MNIST and Fashion-MNIST. We interpret these results through the existing metric-topology factorization framework: preserving task information and realizing reliable separation are distinct design problems. Our contribution is resource accounting for their realization; learning and revising the factorization itself remains open.

2 Decision Structure, Margin, and Slope

Our principal results concern one-dimensional inputs $X = [0, 1]$ and scalar scores. General-domain separation motivates the problem; coarea and a specific tensorized construction provide higher-dimensional extensions in Section D. They do not establish invariant collar mass for arbitrary data manifolds. Let $Y \in \{0, 1\}$ and write $\eta(x) = \mathbb{P}(Y = 1 \mid X = x)$ and $g^*(x) = \mathbf{1}\{\eta(x) \geq 1/2\}$. For a continuous scalar score $f : X \to \mathbb{R}$, define its frontier and score collar by $\Gamma_f := \{x : f(x) = 0\}$, $\mathrm{Col}_\epsilon(f) := \{x : |f(x)| \leq \epsilon\}$. Throughout this paper, continuous piecewise-linear (CPWL) will be used in the usual piecewise-affine sense. We distinguish existence of separation, reliability of its score, and cost of its realization. A Bayes-margin profile below is the distribution of $|\eta(X) - 1/2|$.

From Urysohn separation to transition collars. For nonempty disjoint closed sets A, B in a normal space, Urysohn's lemma provides $u : X \to [0, 1]$ continuous with $u|_A = 0$ and $u|_B = 1$ (Munkres, 2000): a fixed *score* gap, but no metric or architectural cost bound. In a metric space,

every G-Lipschitz score with $f|_A \le -m$ and $f|_B \ge m$, $m > 0$, obeys

$$2m \le G\, d(A, B), \qquad d(A, B) := \inf_{a \in A, b \in B} d(a, b). \tag{1}$$

A fixed score gap across nearby class cores requires a large slope; zero set distance rules out a finite Lipschitz constant. Adjacent Bayes regions can have intersecting closures, so we separate closed cores and charge the intervening transition through density and confidence. For a Urysohn separator, $\mathrm{Col}_\epsilon(u - 1/2)$ is the band $\{|u - 1/2| \le \epsilon\}$ whose probability measures exposure to the decision threshold. The next results explain how transitions, their slopes, and probability mass interact.

Exact conservation under an expanding encoder. Let $T : [0, 1] \to [0, 1]$ be continuous with $k \ge 2$ full branches: on each partition interval's interior, T is a C^1 diffeomorphism onto $(0, 1)$ with $\inf |T'| > 1$. Write $T^{\circ L}$ for its L-fold composition, with $T^{\circ 0} = \mathrm{Id}$. Its transfer operator is

$$(\mathcal{L}_T q)(y) = \sum_{x \in T^{-1}\{y\}} q(x)/|T'(x)| \quad \text{a.e.} \tag{2}$$

The transfer operator describes how the input density changes under iteration. The theorem below first expresses collar probability for an arbitrary density, then establishes exact conservation under the additional assumption of an invariant density p, satisfying $\mathcal{L}_T p = p$.

Theorem 1 (Topology-slope-margin conservation)**.** *For $L \ge 1$ and $c \in (0, 1)$, $f_L = T^{\circ L} - c$ has k^L interior sign-changing zeros. For a density q and $0 < \epsilon < \min(c, 1 - c)$,*

$$P_q(|f_L| \le \epsilon) = \int_{c-\epsilon}^{c+\epsilon} (\mathcal{L}_T^L q)(y)\, dy. \tag{3}$$

If p is invariant, $\mathcal{L}_T p = p$, then

$$P_p(|f_L| \le \epsilon) = \int_{c-\epsilon}^{c+\epsilon} p(y)\, dy, \tag{4}$$

$$\sum_{T^{\circ L}(x) = c} \frac{p(x)}{|(T^{\circ L})'(x)|} = p(c) \quad \textit{for a.e. } c. \tag{5}$$

Under P_p, for any measurable strictly sign-preserving $\psi : [-c, 1 - c] \to [-1/2, 1/2]$ with $\psi(0) = 0$, the regression function $\eta_L = 1/2 + \psi(f_L)$ has the same k^L-point Bayes boundary and an L-independent Bayes-margin distribution.

Full branches supply the crossing count; classical pushforward invariance preserves the collar mass through inverse-slope compensation. The conservation result motivates a broader question: without invariance, what must a reliable classifier still supply? We first fix the reliability interface, then derive separate slope and structural demands.

An operational margin: noisy score readout. Write $y^* = 2g^* - 1$, $w = |2\eta - 1|$, and fix a receiver scale $\tau > 0$. For any score, we define the weighted hinge deficit

$$\mathcal{D}_\tau(f) = \mathbb{E}[w(X)(1 - y^*(X) f(X)/\tau)_+]. \tag{6}$$

Let $Z \sim \mathrm{Unif}[-\tau, \tau]$ be independent of (X, Y) and added *after all score computation and gain*. The receiver predicts $\widehat{g}_Z(X) = \mathbf{1}\{f(X) + Z \ge 0\}$, with excess risk $\mathcal{E}_\tau(f) = \mathbb{P}(\widehat{g}_Z(X) \ne Y) - R^*$, where R^* is the clean Bayes risk. We have the following result regarding the noisy score readout.

Proposition 1 (Readout identity and transfer)**.** *Set $\bar{f} = \mathrm{clip}(f, -\tau, \tau)$. Then*

$$2\mathcal{E}_\tau(f) = \mathcal{D}_\tau(\bar{f}), \qquad \operatorname{sign} \bar{f} = \operatorname{sign} f, \qquad \mathrm{Lip}(\bar{f}) \le \mathrm{Lip}(f). \tag{7}$$

Consequently, all deficit-based slope and recovered-alternation bounds below apply to f at readout-risk budget r by substituting $\varepsilon = 2r$.

Conditioning on X gives a weighted disagreement probability $w\, \mathrm{clip}((1 - y^* f/\tau)/2, 0, 1)$, proving the identity. Note that clipping is a proof device (no extra architectural pieces are charged). Uniform threshold noise makes the relation exact. The noise law is a receiver assumption; we will provide a general symmetric-noise identity in Section A.2. For arbitrary weights $w \in [0, 1]$, the same identity describes weighted noisy disagreement.

A margin requirement imposes a slope cost. For a CPWL score with isolated roots interior to affine pieces, the local collar coefficient is $\sum_j p(r_j)/|f'(r_j)|$. The following result constrains arbitrary scores, including those that omit target boundaries.

Theorem 2 (Margin-slope obstruction). *Let $y^* : [0,1] \to \{-1,1\}$ change sign at distinct interior points $r_1, \dots, r_K$. For each j, suppose there are $\rho_j > 0$ and $\sigma_j \in \{-1,1\}$ such that the intervals $I_j = (r_j - \rho_j, r_j + \rho_j)$ are disjoint and, almost everywhere for $0 < t < \rho_j$, $y^*(r_j + t) = \sigma_j$ and $y^*(r_j - t) = -\sigma_j$. Let P_X have density p, let $w : [0,1] \to [0,1]$, and assume*

$$p(r_j \pm t)w(r_j \pm t) \geq c_j t^{\beta_j} \quad \text{a.e. on } (0, \rho_j), \tag{8}$$

where $c_j > 0$ and $\beta_j \geq 0$. For $\tau > 0$ and a G-Lipschitz score f, $G > 0$, set $b_j = \min\{\rho_j, \tau/G\}$. Then we have

$$\mathcal{D}_\tau(f) \geq \sum_{j=1}^{K} 2c_j \left(\frac{b_j^{\beta_j+1}}{\beta_j + 1} - \frac{G b_j^{\beta_j+2}}{\tau(\beta_j + 2)} \right). \tag{9}$$

In particular, if $\beta_j = \beta$, $c_j \geq c_0$, and $\rho_j \geq \rho_0$, then

$$\mathcal{D}_\tau(f) \geq \frac{2c_0 K}{(\beta+1)(\beta+2)} \min\{\rho_0, \tau/G\}^{\beta+1}. \tag{10}$$

Consequently, if its left side is smaller than $2c_0 K \rho_0^{\beta+1}/((\beta+1)(\beta+2))$, then

$$G \geq \tau \left[\frac{2c_0 K}{(\beta+1)(\beta+2)\mathcal{D}_\tau(f)} \right]^{1/(\beta+1)}. \tag{11}$$

For Bayes labels and $w = |2\eta - 1|$, the deficit upper-bounds clean excess risk; a deficit lower bound transfers only through the noisy-readout identity, applied to $\bar{f}$. The inverse bound requires its displayed separation condition; c_0 and ρ_0 cannot be suppressed when varying K. For the calibrated tent family in Eq. (17), under uniform input, write $u_L = \eta_{L,\delta,\tau} - 1/2$ and $K = 2^L$. Then

$$\mathcal{D}_\tau(u_L) = 2\tau\delta/3, \quad \mathrm{Lip}(u_L) = \tau K/\delta, \tag{12}$$

and Eq. (9) is attained with $c_j = 2\tau K/\delta$, $\beta_j = 1$, and $\rho_j = \delta/K$.

A deficit budget limits the structure that can be missed. The slope bound refers to target boundaries even when the predictor omits them. To obtain an architectural bound, we also need to know how much of the target's alternating structure survives. The relevant quantity is the *confidence mass* of each label region: missing a large, confident region consumes more of the available deficit than missing a small or uncertain one.

Theorem 3 (Confidence-weighted recovery of decision structure). *Let $J_0 < \cdots < J_K$ be disjoint open intervals on which y^* has alternating signs $s_i \in \{-1,1\}$ almost everywhere, and let $v_i = \int_{J_i} pw > 0$. For $\varepsilon \geq 0$, define*

$$A_\varepsilon := \min_{B \subseteq \{0,\dots,K\} : \sum_{i \in B} v_i \leq \varepsilon} \mathrm{alt}(s_i : i \notin B), \tag{13}$$

where alt *counts consecutive sign changes in the retained ordered sequence, with value zero for length at most one. If f is continuous and $\mathcal{D}_\tau(f) \leq \varepsilon$, it has at least A_ε ordered sign alternations. Writing $v_{(i)}$ for the masses in increasing order and $b_\varepsilon = \max\{b : \sum_{i=1}^{b} v_{(i)} \leq \varepsilon\}$,*

$$A_\varepsilon \geq (K - 2b_\varepsilon)_+ \geq \left(K - 2\lfloor \varepsilon / v_{\min} \rfloor\right)_+. \tag{14}$$

Also, with $d_0 = d_K = 1$ and $d_i = 2$ internally ($K \geq 1$),

$$A_\varepsilon \geq \left\lceil K - \varepsilon \max_i (d_i/v_i) \right\rceil_+. \tag{15}$$

If a region contains no correctly signed point, its entire confidence mass is charged to the deficit. Deleting one region removes at most two alternations. The same deficit budget now supplies both a slope demand and a realized-topology demand. It is worth noting that in higher dimensions, coarea replaces crossing count by a weighted level-set integral involving inverse gradient magnitude (Proposition 4) but does not bound arbitrary topological features (Evans and Gariepy, 2015).

3 Architectural Cost of Reliable Decisions

The two demands are simultaneous: sufficient slope does not supply missing sign alternations and additional pieces do not replace required slope. An architecture determines how economically both are realized. The tent family provides a target on which lower and upper costs match. Define the tent map on $[0,1]$ by $\Delta(t) := 1 - |1-2t|$, $\quad \phi_L(t) := \Delta^{\circ L}(t) - \frac{1}{2}$, $\quad t \in [0,1]$. Its zero set is

$$\Gamma_L = \left\{ \frac{2j+1}{2^{L+1}} : j = 0, \dots, 2^L - 1 \right\}. \tag{16}$$

For $0 < \delta < 1/2$ and $0 < \tau \le 1/2$, let's define the continuous noisy regression model

$$\eta_{L,\delta,\tau}(t) := \frac{1}{2} + \tau \ \mathrm{clip}\left(\frac{\phi_L(t)}{\delta}, -1, 1 \right). \tag{17}$$

Its Bayes classifier is $g_L^*(t) = \mathbf{1}\{\phi_L(t) \ge 0\}$ up to the null boundary. The standard fold chain realizes ϕ_L with width two and $O(L)$ nonzeros, whereas a depth-D, width-m ReLU score realizing its committed signs satisfies $(m+1)^D \ge 2^L$. We only discuss the key results regarding architectural cost here. The full depth-width statement and invariant margin profile can be found in Appendix (Theorem 9 and Corollary 5).

From deficit to architectural cost. Piece counts and Lipschitz constants are submultiplicative under scalar CPWL composition. For module bounds $B(q) \ge 1$, $A(q) = \max(1, \mathrm{Lip}(q))$, and costs $c(q) \ge \lambda \log B(q) + \mu \log A(q)$, $\lambda, \mu \ge 0$, the recovery and slope demands give

$$C(f) \ge \lambda \log \max(1, A_\varepsilon) + \mu \log \max(1, S_\varepsilon), \tag{18}$$

where S_ε is Eq. (11) under its low-deficit condition (Theorem 6). A specified catalog admits a match.

Theorem 4 (Architectural cost at prescribed margin deficit)**.** *Fix coordinate units. Allow arbitrary scalar CPWL modules with at most two affine pieces and Lipschitz constant at most 2, costing a, and affine modules with Lipschitz constant at most 2, costing b, where $a \ge b > 0$. Translations and sign reversal are free; every amplification is charged. For uniform input, $K = 2^n$, $n \ge 1$, deterministic labels $y^* = \mathrm{sign}\,\phi_n$, and $w = \tau = 1$, let $C_K^*(\varepsilon)$ be the infimum cost over all finite compositions with $\mathcal{D}_1(f) \le \varepsilon$. For $0 < \varepsilon < 1/4$, set $m = \lceil \log_2(1/\varepsilon) \rceil$. Then we have $C_K^*(\varepsilon) = an + bm$. The construction is $f = 2^m \phi_n$: n tent modules followed by a translation and m gain modules $z \mapsto 2z$; its deficit is exactly 2^{-m}.*

Corollary 1 (Exact cost of noisy readout)**.** *In the catalog and deterministic task of Theorem 4, with unit-scale downstream noise and $0 < r < 1/8$, the minimum cost satisfying $\mathcal{E}_1(f) \le r$ is $C^*_{K,\mathrm{readout}}(r) = an + b\lceil \log_2(1/(2r)) \rceil$.*

Recovery applies to $\bar{f}$ (the clipped score introduced in Proposition 1) while counting alternations of f; clipping need not belong to the catalog. The gain chain attains the bound; unbounded gain at cost b instead gives cost at most $an + b$ for every $r > 0$: the growing reliability cost depends on the stagewise gain limit and fixed downstream noise.

Beyond dyadic targets: allocating spline resources. For arbitrary ordered label cells, a small reliable region can require as many sign changes as a large one, but omitting it consumes less confidence mass. Now we face a concrete allocation problem: which regions should a limited spline resolve and when can additional knots help without more slope? For $P \ge 1$, let

$$M_P = \min_{B:\, \mathrm{alt}(s_i : i \notin B) \le P} \sum_{i \in B} v_i. \tag{19}$$

Proposition 2 (Irregular-region resource certificate)**.** *For Bayes labels and $w = |2\eta - 1|$, under the local assumptions of Theorem 2 and the cell assumptions of Theorem 3, every CPWL score with at most P affine pieces and Lipschitz constant at most G satisfies*

$$\mathcal{E}_\tau(f) \ge \max\left\{ \frac{M_P}{2}, \sum_{j=1}^{K} c_j \left[\frac{b_j^{\beta_j+1}}{\beta_j+1} - \frac{G b_j^{\beta_j+2}}{\tau(\beta_j+2)} \right] \right\}, \quad b_j = \min(\rho_j, \tau/G). \tag{20}$$

The bound applies to irregularly spaced regions without requiring a fold factorization. It also provides a screening rule: any architecture budget whose lower bound exceeds the target risk can be excluded before fitting (Corollary 8 in Section I of Appendix).

4 APPROXIMATION, DISTRIBUTION SHIFT, AND ARCHITECTURAL COST

How much conservation survives? Before proceeding to guide representations, we ask what survives inexact realization. Score error and input-law mismatch perturb conservation separately; approximation also retains resource demands. Distribution mismatch changes a score interval's probability; uniform score error changes its endpoints. We measure distribution shift using total variation distance (Levin and Peres, 2017): $\mathrm{TV}(\mu,\nu) = \sup_A |\mu(A) - \nu(A)|$. To describe the corresponding reference collar, let $H_{p,c}(r) = \int_{c-r}^{c+r} p(y)\,dy$ denote the probability mass within distance r of the readout threshold c under the reference density p. The following theorem characterizes the approximation and invariance defects.

Theorem 5 (Approximation and invariance defects)**.** *Let $T_1,\dots,T_L : [0,1] \to [0,1]$ be measurable, set $F_L = T_L \circ \cdots \circ T_1$, let $\mu(dx) = p(x)dx$, and define $\alpha_j = \mathrm{TV}((T_j)_{\#}\mu, \mu)$ and $d_0 = \mathrm{TV}(\nu,\mu)$. Suppose $\|\widehat{f} - (F_L - c)\|_\infty \le \xi$. With $d_L = \min\{1, d_0 + \sum_j \alpha_j\}$, for $\xi \le \epsilon$ and $\epsilon + \xi < \min\{c, 1-c\}$, the score-collar probability satisfies*

$$H_{p,c}(\epsilon - \xi) - d_L \le P_\nu(|\widehat{f}| \le \epsilon) \le H_{p,c}(\epsilon + \xi) + d_L. \tag{21}$$

If $p \le M$, the absolute difference from $H_{p,c}(\epsilon)$ is at most $2M\xi + d_L$. If $\widehat{f} = \widehat{T}_L \circ \cdots \circ \widehat{T}_1 - c$, all approximate maps are self-maps of $[0,1]$, $\|\widehat{T}_j - T_j\|_\infty \le e_j$, and $\mathrm{Lip}(T_j) \le A_j$, one may take

$$\xi = \sum_{j=1}^{L} e_j \prod_{i=j+1}^{L} A_i. \tag{22}$$

An early layer error is multiplied by all subsequent slopes in Eq. (22): stability requires a controlled perturbation budget even when the reference collar probability is invariant.

Approximate scores retain architectural demands. Uniform error below $1/2$ preserves committed signs and piece demand, while the slope demand weakens continuously with error.

Corollary 2 (Approximate tent realizations retain their cost)**.** *Let $K = 2^L$ and let a continuous score $\widehat{f}$ satisfy $\|\widehat{f} - \phi_L\|_\infty \le \xi < 1/2$ on $[0,1]$. If $\widehat{f}$ is Lipschitz, then $\mathrm{Lip}(\widehat{f}) \ge K(1-2\xi)$. For every $\xi < \delta < 1/2$, its sign agrees with ϕ_L outside $\mathrm{Col}_\delta(\phi_L)$. Consequently a CPWL realization needs at least K affine pieces, and a depth-D, width-m ReLU realization obeys $(m+1)^D \ge K$. Under uniform input, for $\xi \le \epsilon$ and $\epsilon + \xi < 1/2$,*

$$\left|\mathbb{P}(|\widehat{f}(X)| \le \epsilon) - 2\epsilon\right| \le 2\xi. \tag{23}$$

What the comparison implies for MLPs and KANs. ReLU depth and spline resolution supply piece capacity in different ways, under a common derivative demand.

Corollary 3 (MLP-KAN payment schedules on the folded family)**.** *Let $K = 2^L$ and consider the score ϕ_L and committed labels in Theorem 9 (refer to Section D in the Appendix). Then we have:*

- *(i) At depth D, a ReLU MLP realizes ϕ_L with width at most $2^{\lceil L/D \rceil}$, while every depth-D, width-m ReLU realization of the committed labels satisfies $(m+1)^D \ge K$.*
- *(ii) A width-one, depth-L order-1 KAN realizes ϕ_L with one three-knot tent spline per nonlinear layer and $O(L) = O(\log K)$ spline coefficients.*
- *(iii) A shallow scalar order-1 spline with G affine intervals on $[0,1]$ that realizes the committed labels must satisfy $G \ge K$; equivalently, it needs at least $K+1$ distinct grid sites, including the endpoints.*
- *(iv) Every CPWL factorization of the calibrated score $u_L = \eta_{L,\delta,\tau} - 1/2 = \Phi_{D+1} \circ \cdots \circ \Phi_1$ has layer Jacobian budgets $J_\ell := \operatorname{ess\,sup}_z \|D\Phi_\ell(z)\|_2$ satisfying*

$$\prod_{\ell} J_\ell \ge \mathrm{Lip}(u_L) = \frac{\tau K}{\delta}. \tag{24}$$

The payment schedule applies to both ReLU MLP and order-1 KAN realizations (Corollary 4).

The above comparison concerns CPWL realizations only. Smooth-basis KAN variants and hardware implementation costs require separate accounting (we elaborate more in Section L).

5 Experiments: Testing the Resource Account

5.1 Experimental Results

The experiments connect capacity (Figure 2), achieved accuracy (Figure 3), and hybrid allocation (Figure 4) to resource selection. E1–E6 supply exploratory training records and exact constructions; E6-C uses separate validation and test samples. E7–E12 test bounds and resource decisions on synthetic tasks. R1–R4 extend the diagnostics to planar mixtures and image data (Section N). Code and saved records accompany the paper.

How do composition and resolution supply decision structure? Corollary 3 predicts that shallow-spline resolution grows with crossing count, whereas composition generates repeated structure with logarithmically many layers. Figure 2 reports the first tested configuration reaching 95% evaluation accuracy. For folded labels (E1), shallow-spline counts rise from 6 to 130 over $L = 1{:}6$, while successful deep splines use 10–30 parameters. Compositional savings are visible where training succeeds; censoring records a finite search failure, not unrepresentability. On checkerboards (E2), the tested KAN uses 30–294 parameters over $k = 2{:}10$, while the depth-3 MLP uses 2545 at $k = 4$. The tested depth-6 ladder gives no consistent saving, whereas a separate fixed-task sweep does benefit from depth (Table 8). The outcomes depend on the search and training protocol. E1–E5 use the evaluation split for selection, so their performance estimates are exploratory.

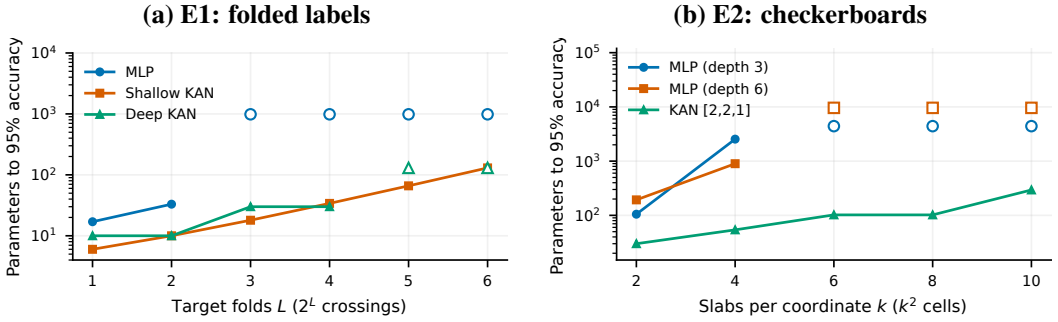


Figure 2: **Composition and resolution have different observed costs.** (a) E1: folded labels. (b) E2: checkerboards. Counts refer to selected tested configurations reaching 95% evaluation accuracy; open markers denote search-cap censoring at the last tested parameter count. No value beyond the tested ladder is inferred. The checkerboard depth sweep is in Table 8.

Why does representational capacity not determine accuracy? The three tested models lose accuracy as spiral winding increases (E3, Figure 3a); their actual budgets are 361, 252, and 162 parameters, so the comparison is not parameter matched. Panel c (E4) contrasts constructed compositional crease counts with trained accuracies: at $L = 5$, ReLU, shallow-spline, and deep-spline models achieve .509, .989, and .890. The corrected constructions realize the same centered L-fold score and have $2^L - 1$ interior creases. Their allocated parameter counts differ; the training arms are separate, larger configurations (Section F). Together these panels distinguish available capacity from the score found by training, without isolating an optimization mechanism. Panel b (E5) isolates label uncertainty: fair independent labels on mass .2 impose Bayes error .10, near observed errors .098 and .101. This explains why Theorem 1's collar mass needs an explicit noise model to imply an error floor. Section 4 guarantees retained resource demands under controlled score approximation and bounds collar changes under distribution shift. E1–E5 do not estimate the approximation or invariance defects required by those guarantees.

Can the same decision structure be realized more economically? E6 uses Corollary 6 to allocate $K = 2^L R$ crossings between L folds and a residual spline. At $K = 64$, cost $J(L) = 3L + 2K2^{-L} + 1$ is minimized at $(L, R) = (5, 2)$ with 20 parameters. Every exact allocation has unit accuracy, 64 crossings, and collar mass .300: representation cost changes while the score and its geometric demands stay fixed. In Figure 4(a), exact fold and hybrid constructions succeed at the original matched budget, while the cold protocol often misses the available solution. Near-exact starts also use a different learning rate, so the comparison does not isolate initialization. Panel b tests $K = 96$:

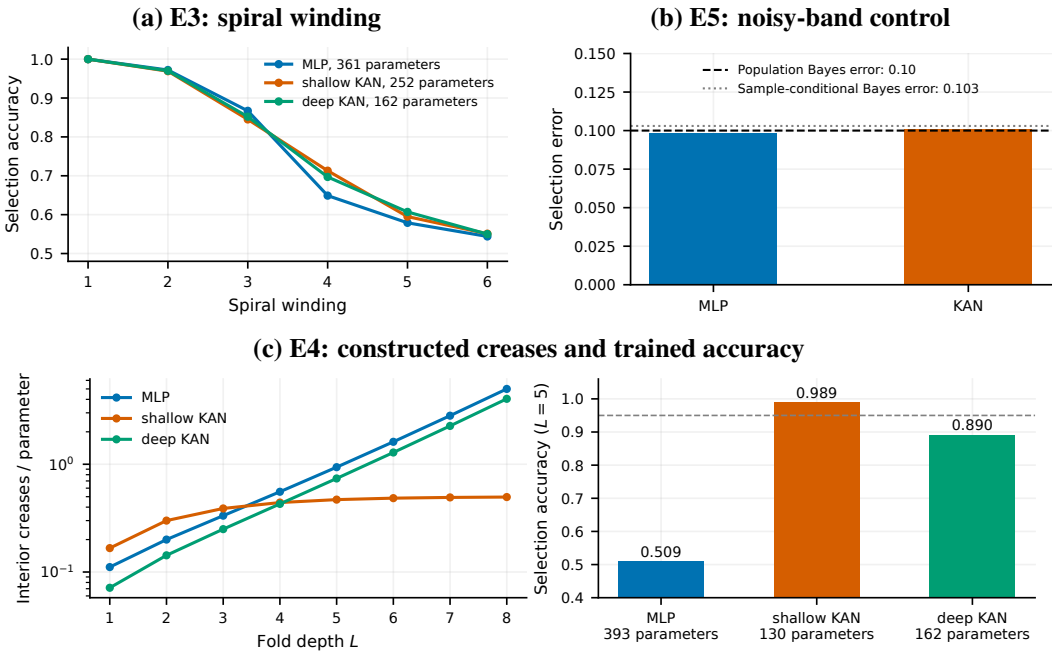


Figure 3: **Capacity, trained accuracy, and label uncertainty.** (a) E3: winding at the actual, unequal parameter counts shown in the legend. (b) E5: error under explicitly randomized labels in a band. (c) E4: corrected same-score constructions (left) and original training outcomes at the displayed parameter counts (right). The panels distinguish expressible complexity from achieved classification and irreducible label error. Accuracy and error use the selection split.

the hybrid uses 22 parameters versus 193 for the specified shallow spline, both with unit accuracy. Dyadic-only schedules achieve .500 under their restricted dictionary. The result supports joint fold-resolution allocation within the declared family. The independent E6-C control crosses both initializations with both learning rates at $L = 3, 5$ over five seeds. Cold starts remain near chance; warm-start accuracy falls from .943–.977 at rate .003 to .626–.646 at .02. Frozen exact folds exceed .986 at either rate. Both initialization and step size matter in the tested regime (Section F.4).

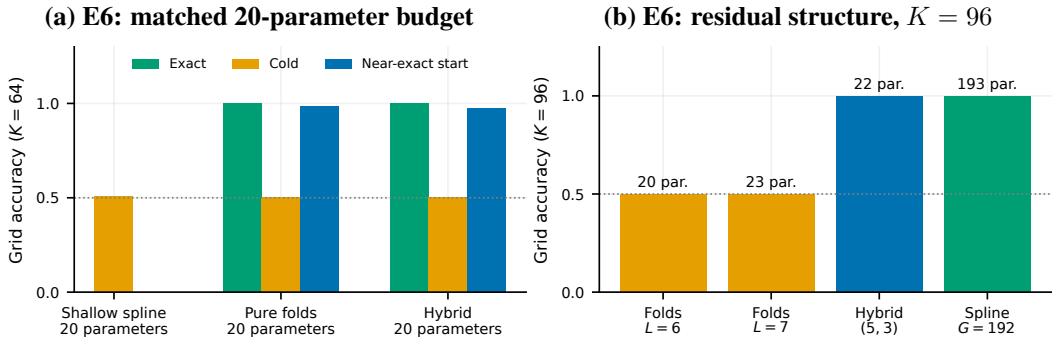


Figure 4: **E6: allocating composition and residual resolution.** (a) $K = 64$: exact (green), cold-start (orange), and near-exact-start (blue) models at the original 20-parameter budget; the shallow spline is knot-truncated. (b) $K = 96$: constructed residual comparison. Panel (a)'s budget differs from the per-depth schedule in Table 1; cold/warm learning rates also differ.

Can the bounds guide consequential resource decisions? E7–E9 check the deficit bounds, recovery accounting, and readout-noise interface; E10 studies irregular, unequal-confidence regions (Sections G and H). E11 applies the joint certificate to screen budgets before exact finite-catalog optimization. Across 15 tasks, screening preserves every minimum-cost feasible choice while avoiding 38.0% of DP work; all work savings come from slope screening (Section I). E12 instead uses finite diagnostics to choose among information restoration, increased gain, and finer resolution.

On 270 clean synthetic tasks, it chooses the same successful heads as cost-ordered search with one rather than two repair fits on average. Including the common initial fit, the counts are two versus three; diagnosis also requires 64 labels. The supplied rule and repair menu fail on mixed mechanisms, and corrupted labels reduce their effectiveness (Section J). These studies establish controlled resource decisions, not end-to-end speedups or autonomous discovery.

Do the geometric diagnostics extend to image data? Experiment R3 trains ReLU classifiers with one to four hidden layers on MNIST 3-versus-5 and 4-versus-9, and Fashion-MNIST T-shirt/top-versus-shirt (LeCun et al., 1998; Xiao et al., 2017). Across three training seeds and approximately comparable sizes (27,343–30,655 parameters), the spectral product grows by 6.7–16.2 times from the shallowest to deepest model. However, the median absolute score divided by a sampled directional slope falls from 2.23 to 1.44, 2.01 to 1.30, and 1.09 to .68 on the three tasks, respectively. Accuracy changes by at most .007 between those endpoints. Greater gain accompanies no improvement in this measured separation proxy (refer to Figure 11 and Table 11 in Sec. N of the Appendix). Since sampled slopes can underestimate the global Lipschitz constant, the proxy is not a certified robustness radius. The experiment tests geometric diagnostics, not certified resource selection on image tasks.

5.2 Implications for Structural Discovery

Learning a useful representation involves two related decisions: which task distinctions to preserve, and how to realize their separation economically. Metric-topology factorization (MTF) (Li, 2026a) organizes these decisions through a structural index, within-state coordinates, and a task readout. If inputs requiring incompatible outputs share the full representation, repair must restore information through the encoder, additional coordinates, or routing. Sharing an index remains compatible with successful classification when within-state coordinates retain the necessary distinction. When information survives, improved reliability may instead require more gain or finer readout resolution. Once the necessary distinctions are retained, compositional reuse can reduce the cost of separating them reliably. From the metric-slingshot perspective (Li, 2026b), contexts can share a metric operation while retaining task-dependent routes or readouts. Such sharing reallocates the work across the architecture: for Lipschitz factors, the slope obstruction still requires $\mathrm{Lip}(r)\,\mathrm{Lip}(z) \geq \mathrm{Lip}(f) \geq S_\varepsilon$ under its assumptions. The benefit of reuse lies in realizing the required structure and sensitivity more economically.

Structural discovery must learn where reuse is possible and when a factorization needs revision. For open-ended agents (Wang et al., 2023), predictive coarse-graining offers a possible consolidation mechanism: merge histories whose differences do not materially affect predictions or task outcomes within a specified task and continuation family. New demands may invalidate an earlier merger, requiring retained evidence or new observations to recover the missing distinction. The objective is an affordable, revisable realization that accounts for structural storage, routing, coordinates, and readout. Our fixed-task bounds constrain realization costs; extending them to structural learning requires principles for consolidating reusable abstractions and revising them when task demands change. To achieve open-ended intelligence (Hughes et al., 2024), representational economy must account for both the cost of present competence and the evidence and resources needed for future adaptation.

6 Conclusion

A simple decision in representation space can conceal substantial work in the encoder. Our scalar results make the cost explicit: under the specified readout model, reliability constrains both the decision distinctions that must survive and the slope needed to separate them. Composition can encode repeated structure economically, but the complete classifier must still meet the geometric demands of reliable separation. The resulting resource account distinguishes preserving task information from realizing reliable separation. Under declared module costs, that distinction supports concrete architectural choices while exposing where additional gain or resolution cannot repair erased information. For open-ended intelligence, the broader challenge is to learn reusable abstractions through predictive coarse-graining while retaining the evidence needed to revise them as task demands expand. The present bounds constrain the cost of reliable realization; learning, consolidating, and revising the structure itself remains open.

Reproducibility Statement

The release includes E1–E12 source code, the original E1–E6 notebooks and selected gradient checks, and the independent E6-C implementation and all 60 runs. Corrected E1–E6 figures are regenerated from saved records, except E4a, whose same-score constructions and crease counts were recomputed and checked without training. R1–R4 retain the supplied notebook's saved outputs; image-model training was not rerun. The original notebooks, plotting data, figure-generation script, and construction checks are included. Original source discrepancies, selection limitations, and incomplete seed-level records are documented in the appendix and release notes. Proofs accompany the mathematical statements.

AI Use Statement

Generative AI assisted with theoretical development and proof checking, experimental code, execution, literature checking, and manuscript preparation. The supplied E1–E12 experiment records were generated with AI assistance; the E6-C control was implemented and executed with AI assistance during revision. AI also assisted with inspecting and integrating the supplied R1–R4 notebook. The authors remain responsible for verifying all results and this disclosure.

References

J.-Y. Audibert and A. B. Tsybakov. Fast learning rates for plug-in classifiers. *The Annals of Statistics*, 35(2):608–633, 2007.

S. Aziznejad, H. Gupta, J. Campos, and M. Unser. Deep neural networks with trainable activations and controlled Lipschitz constant. *IEEE Transactions on Signal Processing*, 68:4688–4699, 2020.

R. Balestriero and R. G. Baraniuk. A spline theory of deep learning. In *International Conference on Machine Learning*, volume 80 of PMLR, pp. 374–383, 2018.

P. L. Bartlett, D. J. Foster, and M. J. Telgarsky. Spectrally-normalized margin bounds for neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Xin Li. Structural Learning Theory: A Metric-Topology Factorization Approach. *arXiv preprint arXiv:2602.07974*, 2026. https://arxiv.org/pdf/2602.07974.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, 2013.

L. Béthune, T. Boissin, M. Serrurier, F. Mamalet, C. Friedrich, and A. González-Sanz. Pay attention to your loss: understanding misconceptions about 1-Lipschitz neural networks. In *Advances in Neural Information Processing Systems*, 2022. https://arxiv.org/abs/2104.05097.

P. Bohra, J. Campos, H. Gupta, S. Aziznejad, and M. Unser. Learning activation functions in deep (spline) neural networks. *IEEE Open Journal of Signal Processing*, 1:295–309, 2020.

V. Chatziafratis, S. G. Nagarajan, I. Panageas, and X. Wang. Depth–width trade-offs for ReLU networks via Sharkovsky's theorem. In *International Conference on Learning Representations*, 2020.

T. M. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, EC-14(3):326–334, 1965.

S. Dirksen, M. Genzel, L. Jacques, and A. Stollenwerk. The separation capacity of random neural networks. *Journal of Machine Learning Research*, 23(309):1–47, 2022.

E. Ergen and M. Grillo. Topological expressivity of ReLU neural networks. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of PMLR, pages 1599–1642, 2024.

L. C. Evans and R. F. Gariepy. *Measure Theory and Fine Properties of Functions*, revised edition. CRC Press, 2015.

B. B. Gaines and J. Bi. Characterizing the discrete geometry of ReLU networks. In *International Conference on Learning Representations*, 2026. Oral presentation.

A. Goujon, A. Etemadi, and M. Unser. On the number of regions of piecewise linear neural networks. *Journal of Computational and Applied Mathematics*, 2024. Preprint: `https://arxiv.org/abs/2206.08615`.

B. Hanin and D. Rolnick. Complexity of linear regions in deep networks. In *International Conference on Machine Learning*, volume 97 of PMLR, pp. 2596–2604, 2019.

C. Keup and M. Helias. Origami in N dimensions: How feed-forward networks manufacture linear separability. arXiv:2203.11355, 2022.

D. A. Levin and Y. Peres. *Markov Chains and Mixing Times*. American Mathematical Society, second edition, 2017.

A. Lasota and M. C. Mackey. *Chaos, Fractals, and Noise: Stochastic Aspects of Dynamics*, second edition. Springer, 1994.

A. Lasota and J. A. Yorke. On the existence of invariant measures for piecewise monotonic transformations. *Transactions of the American Mathematical Society*, 186:481–488, 1973.

Z. Liu, Y. Wang, S. Vaidya, F. Ruehle, J. Halverson, M. Soljačić, T. Y. Hou, and M. Tegmark. KAN: Kolmogorov–Arnold networks. In *International Conference on Learning Representations*, 2025.

E. Mammen and A. B. Tsybakov. Smooth discrimination analysis. *The Annals of Statistics*, 27(6):1808–1829, 1999.

G. Montúfar, R. Pascanu, K. Cho, and Y. Bengio. On the number of linear regions of deep neural networks. In *Advances in Neural Information Processing Systems*, volume 27, 2014.

J. R. Munkres. *Topology*, second edition. Prentice Hall, 2000. Section 33: The Urysohn Lemma.

G. Naitzat, A. Zhitnikov, and L.-H. Lim. Topology of deep neural networks. *Journal of Machine Learning Research*, 21(184):1–40, 2020.

S. Parkinson, G. Ongie, R. Willett, O. Shamir, and N. Srebro. Depth separation in norm-bounded infinite-width neural networks. In *Proceedings of the 37th Conference on Learning Theory*, volume 247 of PMLR, pages 4082–4114, 2024.

N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville. On the spectral bias of neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of PMLR, pages 5301–5310, 2019.

A. Rényi. Representations for real numbers and their ergodic properties. *Acta Mathematica Academiae Scientiarum Hungaricae*, 8:477–493, 1957.

I. Safran, R. Eldan, and O. Shamir. Depth separations in neural networks: What is actually being separated? In *Proceedings of the 32nd Conference on Learning Theory*, volume 99 of PMLR, pages 2664–2666, 2019.

C. H. Sanford and V. Chatziafratis. Expressivity of neural networks via chaotic itineraries beyond Sharkovsky's theorem. In *AISTATS*, volume 151 of PMLR, pp. 9505–9549, 2022. `https://proceedings.mlr.press/v151/sanford22a.html`.

T. Serra, C. Tjandraatmadja, and S. Ramalingam. Bounding and counting linear regions of deep neural networks. In *International Conference on Machine Learning*, volume 80 of PMLR, pp. 4558–4566, 2018.

M. Telgarsky. Benefits of depth in neural networks. In *Proceedings of the 29th Annual Conference on Learning Theory*, volume 49 of PMLR, pages 1517–1539, 2016.

R. Yu, W. Yu, and X. Wang. KAN or MLP: A fairer comparison. arXiv:2407.16674, 2024.

Xin Li. The Metric Slingshot: Navigational Reuse as Width-Optimal Structural Decoupling in Continual Learning. *Neural Computation*, 2026.

Edward Hughes, Michael Dennis, Jack Parker-Holder, Feryal Behbahani, Aditi Mavalankar, Yuge Shi, Tom Schaul, and Tim Rocktäschel. Open-endedness is essential for artificial superhuman intelligence. *arXiv preprint arXiv:2406.04268*, 2024.

Ziyao Li. Kolmogorov-Arnold networks are radial basis function networks. *arXiv preprint arXiv:2405.06721*, 2024. https://arxiv.org/abs/2405.06721.

Blealtan. efficient-kan: An efficient pure-PyTorch implementation of Kolmogorov-Arnold Network (KAN). Software, 2024. https://github.com/Blealtan/efficient-kan. Accessed September 12, 2026.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. https://doi.org/10.1109/5.726791.

H. Xiao, K. Rasul, and R. Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747, 2017. https://arxiv.org/abs/1708.07747.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

GUIDE TO THE APPENDIX

The appendix first supplies the proofs supporting the main paper, then the supplementary constructions and representation facts. Experimental protocols and results follow in E1–E12 order, with each study's figures and tables kept in its own section. Additional diagnostics, related work, and limitations follow. The new R1–R4 supplement is collected in Section N, preserving the existing appendix labels.

Section	Contents
A	Proofs for decision structure, margin, and slope (Section 2)
B	Proofs for architectural cost and safe selection (Section 3)
C	Proofs for stability and approximation (Section 4)
D	Supporting geometry, tent constructions, and higher-dimensional extensions
E	Separation and structural discovery
F–J	E1–E12: protocols, supplementary results, and controls
K	Additional margin and checkerboard diagnostics
L–M	Related work and limitations
N	R1–R4: planar and image-data diagnostics

A PROOFS FOR DECISION STRUCTURE, MARGIN, AND SLOPE

The proofs follow the order of Section 2: exact conservation, the noisy-readout identity, the margin–slope obstruction, and confidence-weighted recovery.

A.1 EXACT CONSERVATION

Proof of Theorem 1. Because every branch of T maps its interior diffeomorphically onto $(0,1)$, the iterate $T^{\circ L}$ has k^L monotone full branches. On each branch it takes the values below and above every $c \in (0,1)$, so the intermediate value theorem and injectivity give one preimage of c. The crossing is transverse because $(T^{\circ L})'$ is nonzero there. This proves the crossing count.

Let $J_1^{(L)}, \dots, J_{k^L}^{(L)}$ be the branch intervals of $T^{\circ L}$, and let $v_a : (0,1) \to (J_a^{(L)})^\circ$ denote the corresponding inverse branches. For every nonnegative measurable h, change of variables on each branch gives

$$\int_0^1 h(T^{\circ L}(x))q(x)\,dx = \int_0^1 h(y) \sum_{a=1}^{k^L} q(v_a(y))|v_a'(y)|\,dy \tag{25}$$

$$= \int_0^1 h(y)(\mathcal{L}_T^L q)(y)\,dy. \tag{26}$$

Taking $h = \mathbf{1}_{[c-\epsilon,c+\epsilon]}$ proves Eq. (3). If p is invariant, then $\mathcal{L}_T^L p = p$ in L^1, which proves Eq. (4). The inverse-function theorem yields

$$(\mathcal{L}_T^L p)(c) = \sum_{x:T^{\circ L}(x)=c} \frac{p(x)}{|(T^{\circ L})'(x)|} \tag{27}$$

at almost every c for compatible representatives. Since $\mathcal{L}_T^L p = p$ almost everywhere, this is Eq. (5). Finally, dividing Eq. (4) by 2ϵ and applying the Lebesgue differentiation theorem gives the local limit $p(c)$ at its Lebesgue points.

For the Bayes-margin statement, sign preservation makes the zero set of $\eta_L - 1/2 = \psi \circ f_L$ equal to $f_L^{-1}(0)$. Under the invariant density, $T^{\circ L}(X)$ again has density p. Applying Eq. (26) with $h(y) = \mathbf{1}\{|\psi(y-c)| \le s\}$ proves the claimed margin distribution. □

A.2 NOISY READOUT AND THE MARGIN DEFICIT

The model is a scalar decision interface with an independent random threshold offset added after its last gain stage. Coordinate units and the noise law are held fixed while the architecture changes. This is an explicit model of readout uncertainty, not evidence that a particular device has uniform noise. It turns the score-margin requirement into an observable classification loss.

Proof of Proposition 1. For a randomized prediction $\widehat{g}$ conditionally independent of Y,

$$\mathbb{P}(\widehat{g} \neq Y \mid x) - \min(\eta(x), 1-\eta(x)) = |2\eta(x) - 1|\mathbb{P}(\widehat{g} \neq g^*(x) \mid x).$$

The identity also holds at $\eta(x) = 1/2$, where the weight is zero. With uniform independent Z, the conditional disagreement probability is $\operatorname{clip}((1 - y^*(x)f(x)/\tau)/2, 0, 1)$. For $z \in \mathbb{R}$, $y \in \{-1, 1\}$,

$$2\operatorname{clip}((1 - yz/\tau)/2, 0, 1) = (1 - y\operatorname{clip}(z, -\tau, \tau)/\tau)_+.$$

Integration proves Eq. (7). Clipping is nonexpansive and sign-preserving, so a slope lower bound for $\bar{f}$ also bounds f, and its ordered sign alternations are those of f. Clipping may introduce extra affine pieces: we count signs of the implemented score, not pieces of an implemented clipping layer. □

Positive gain preserves $|f(x)|/\operatorname{Lip}(f)$ and the clean decision regions. For the upstream-noise receiver $g(\phi(X) + Z)$, the sign is independent of $g > 0$. Scaling downstream noise together with gain also preserves the signal-to-noise ratio. Related Lipschitz-classifier work emphasizes the loss scale and chosen robustness objective (Béthune et al., 2022); strong oscillatory approximation bounds precede this work (Sanford and Chatziafratis, 2022). The readout identity is elementary. Its role is to transfer the coupled recovery and slope accounting to a specified risk.

Other downstream noise laws. Let Z be independent of (X, Y) with a continuous symmetric cumulative distribution function F_Z, and define excess risk $\mathcal{E}_Z$ relative to the clean Bayes risk. Conditioning on X gives

$$\mathcal{E}_Z(f) = \mathbb{E}[w(X)F_Z(-y^*(X)f(X))]\,. \tag{28}$$

Indeed, for $y^* = 1$ disagreement has probability $F_Z(-f)$, and for $y^* = -1$ symmetry gives $1 - F_Z(-f) = F_Z(f)$. Thus Gaussian noise gives $\mathbb{E}[w\Phi(-y^*f/\sigma)]$, rather than the clipped-hinge identity. If Z has density at least $\kappa > 0$ on $[-\tau, \tau]$, then

$$\mathcal{E}_Z(f) \geq \kappa\tau\,\mathcal{D}_\tau(\operatorname{clip}(f, -\tau, \tau)). \tag{29}$$

To verify the inequality, set $a = y^*f$. For $|a| \leq \tau$, integrate the density lower bound over $[-\tau, -a]$ to obtain $F_Z(-a) \geq \kappa(\tau - a)$. For $a < -\tau$, integrate over $[-\tau, \tau]$; for $a > \tau$ the right-hand side is zero. Integration against w proves the claim. A risk budget r permits the conservative substitution $\varepsilon = r/(\kappa\tau)$ in the deficit bounds wherever their assumptions hold. Uniform noise has $\kappa = 1/(2\tau)$ and recovers the exact identity. Other laws need not preserve the exact catalog optimum. No assertion about the distribution of SGD-trained margins, or a measured device noise law, is needed for these conditional statements.

A.3 Margin–slope obstruction and sharpness

Proof of Theorem 2. Fix j and $0 < t < \rho_j$, and write $a = \sigma_j f(r_j)$. Lipschitz continuity gives

$$y^*(r_j + t)f(r_j + t) \leq a + Gt, \qquad y^*(r_j - t)f(r_j - t) \leq -a + Gt.$$

Because $z \mapsto (1 - z/\tau)_+$ is decreasing and convex, the sum of the two hinge terms is at least

$$\left(1 - \frac{a + Gt}{\tau}\right)_+ + \left(1 - \frac{-a + Gt}{\tau}\right)_+ \geq 2\left(1 - \frac{Gt}{\tau}\right)_+.$$

By Eq. (8), the contribution of the paired points to Eq. (6) is at least $2c_j t^{\beta_j}(1 - Gt/\tau)_+$. Integrating from 0 to ρ_j, summing over the disjoint I_j, and evaluating the integral up to $b_j = \min\{\rho_j, \tau/G\}$ proves Eq. (9). For the common parameters, $Gb_j/\tau \leq 1$, so each summand in Eq. (9) is at least $2c_0 b_j^{\beta+1}/((\beta+1)(\beta+2))$, and $b_j \geq \min\{\rho_0, \tau/G\}$. This proves Eq. (10). If the stated strict deficit condition holds, the minimum cannot equal ρ_0; it equals τ/G, and rearrangement proves Eq. (11). □

Sharpness of the margin-deficit obstruction. For the calibrated tent score u_L, the integrand in Eq. (6) vanishes outside the K disjoint intervals of radius δ/K around the roots. Writing t for distance to a root, each side of such an interval has $|u_L| = \tau K t/\delta$ and $w = 2|u_L|$. Hence

$$\mathcal{D}_\tau(u_L) = K \cdot 2\int_0^{\delta/K} 2\frac{\tau K t}{\delta}\left(1 - \frac{Kt}{\delta}\right) dt = \frac{2\tau\delta}{3}. \tag{30}$$

Moreover, Eq. (8) holds with $\beta_j = 1$, $c_j = 2\tau K/\delta$, and $\rho_j = \delta/K$, while $G = \tau K/\delta$. Every inequality in the paired-point proof is an equality, proving Eq. (12) and sharpness of Eq. (9) on this family.

A.4 CONFIDENCE-WEIGHTED RECOVERY

Proof of Theorem 3. Call J_i missed when $s_i f(x) \le 0$ for every $x \in J_i$. Its contribution to $\mathcal{D}_\tau$ is at least v_i, since the hinge is at least one almost everywhere there. If B is the set of missed intervals, disjointness gives $\sum_{i\in B} v_i \le \mathcal{D}_\tau(f) \le \varepsilon$. Every retained interval contains a point with $s_i f > 0$ and, by continuity, a nonempty subinterval of that sign. Their ordered sign sequence has at least A_ε alternations. Deleting one symbol from an alternating sequence removes at most two transitions, even after previous deletions. Thus deleting b intervals leaves at least $(K-2b)_+$ transitions. Any b missed intervals have mass at least the sum of the b smallest masses, so $|B| \le b_\varepsilon \le \lfloor \varepsilon / v_{\min} \rfloor$. For the last bound, view the original alternating sequence as a path with K edges and vertex degrees d_i. Every edge with both endpoints retained still contributes an alternation. The number of removed edges is at most $\sum_{i\in B} d_i \le \max_i(d_i/v_i) \sum_{i\in B} v_i \le \varepsilon \max_i(d_i/v_i)$. Taking the integer ceiling proves Eq. (15). □

Reliability above a specified fraction of the margin. For $0 \le \theta < 1$, call a region missed when $s_i f \le \theta\tau$ throughout it. Its deficit contribution is at least $(1-\theta)v_i$. Replacing the budget in Eq. (13) by $\varepsilon/(1-\theta)$ certifies an ordered sequence of subintervals with $s_i f > \theta\tau$ and at least that many alternations. The main theorem uses $\theta = 0$. The weighted-order certificate need not be attainable by a continuous score at the same deficit: it is a necessary bound, not an exact loss formula.

B PROOFS FOR ARCHITECTURAL COST AND SAFE SELECTION

This section supports Section 3. We first state and prove the module accounting bound used in Eq. (18), then prove the exact catalog costs and the certificates for irregular regions and safe selection.

B.1 GENERAL MODULE ACCOUNTING AND COUPLED RESOURCE DEMANDS

Theorem 6 (Primitive-independent payment bound)**.** *Let $f = q_D \circ \cdots \circ q_0$ be a composition of scalar CPWL modules between prescribed compact intervals, with coordinate scales fixed as part of the module catalog. Suppose module q has at most $B(q) \ge 1$ affine pieces, $A(q) := \max\{1, \mathrm{Lip}(q)\}$, and additive cost $c(q) \ge 0$. Suppose there are $K_{\mathrm{eff}} + 1$ ordered nonempty intervals on which f is nonzero with alternating sign, and suppose $\mathrm{Lip}(f) \ge S$. Then*

$$\prod_{\ell=0}^{D} B(q_\ell) \ge \max\{1, K_{\mathrm{eff}}\}, \qquad \prod_{\ell=0}^{D} A(q_\ell) \ge \max\{1, S\}. \tag{31}$$

For any $\lambda, \mu \ge 0$ such that every allowed primitive satisfies

$$c(q) \ge \lambda \log B(q) + \mu \log A(q), \tag{32}$$

the total architectural cost obeys

$$\sum_{\ell=0}^{D} c(q_\ell) \ge \lambda \log \max\{1, K_{\mathrm{eff}}\} + \mu \log \max\{1, S\}. \tag{33}$$

Proof of Theorem 6. On any affine piece of q_0, the composition $q_1 \circ q_0$ encounters at most $B(q_1)$ affine pieces of q_1. Induction gives

$$N_{\mathrm{aff}}(f) \le \prod_{\ell=0}^{D} B(q_\ell). \tag{34}$$

A continuous CPWL function with K_{eff} ordered sign transitions needs at least K_{eff} affine pieces, including when a zero set contains a degenerate flat interval; hence the first inequality in Eq. (31). Lipschitz constants are submultiplicative, so

$$S \le \mathrm{Lip}(f) \le \prod_{\ell=0}^{D} \mathrm{Lip}(q_\ell) \le \prod_{\ell=0}^{D} A(q_\ell), \tag{35}$$

which proves the second inequality. Finally, sum Eq. (32) over modules and apply the logarithms of the two product bounds. Nonnegativity of λ and μ preserves both inequalities and yields Eq. (33). □

Proof of Eq. (18). Use Theorem 3 to choose $K_{\text{eff}} = A_\varepsilon$ in Theorem 6. If the common-parameter low-deficit threshold of Theorem 2 is larger than ε, then

$$S_\varepsilon = \tau \left[\frac{2c_0 K}{(\beta+1)(\beta+2)\varepsilon} \right]^{1/(\beta+1)}$$

is a necessary slope. Substitution gives the result. All intervals, density, confidence weights and score scale are those of the same classification problem. No separate exact-recovery hypothesis is invoked. □

B.2 Exact dyadic catalog cost

Proof of Theorem 4. All primitive bounds are measured in fixed scalar coordinates on their prescribed domains. Equivalently, use globally defined modules with the stated bounds; all ranges of each finite construction are compact. Let N_F count charged two-piece modules and N_G charged affine modules. Free translations and sign reversals increase neither resource. Composition has at most 2^{N_F} affine pieces and Lipschitz constant at most $2^{N_F+N_G}$. The K tent roots have disjoint radius-$1/(2K)$ collars. With $p = w = \tau = 1$, Theorem 2 implies

$$\mathcal{D}_1(f) \geq K \min\{1/(2K), 1/G\}.$$

Hence $\mathcal{D}_1(f) \leq \varepsilon < 1/2$ forces $G \geq K/\varepsilon$. Constant scores are excluded by the same inequality applied with arbitrarily small positive G. The $K+1$ label cells have masses $1/(2K)$ at the two endpoints and $1/K$ internally. Thus $d_i/v_i = 2K$ for every cell. Eq. (15) certifies at least $\lceil K(1-2\varepsilon) \rceil$ sign alternations. Since $\varepsilon < 1/4$, this is strictly greater than $K/2$, so $N_F \geq n$. The slope demand gives $N_F + N_G \geq \lceil \log_2(K/\varepsilon) \rceil = n + m$. For $a \geq b$, these imply

$$C(f) = aN_F + bN_G = (a-b)N_F + b(N_F + N_G) \geq an + bm. \tag{36}$$

More generally, integer demands $N_F \geq t$ and $N_F + N_G \geq s$ yield $C(f) \geq at + b\max(0, s-t)$. The weaker certificate $k_\varepsilon = (K - 2\lfloor 2K\varepsilon \rfloor)_+$ uses $t = \lceil \log_2 \max(1, k_\varepsilon) \rceil$ and $s = \lceil \log_2(K/\varepsilon) \rceil$; E8 reports this coarser bound too.

For the upper bound, the global two-piece map $q(z) = 2z - 4\,\mathrm{ReLU}(z - 1/2)$ has Lipschitz constant 2 and equals Δ on $[0, 1]$. Compose n copies, translate by $-1/2$, and apply m gains $z \mapsto 2z$. The result $f = 2^m \phi_n$ costs $an + bm$ and has slope $G = 2^m K$. Since $m \geq 3$, the unit-margin transition intervals are disjoint. Integrating the triangular deficit at each root gives

$$\mathcal{D}_1(f) = 2K \int_0^{1/G} (1 - Gt)\, dt = K/G = 2^{-m} \leq \varepsilon.$$

This proves the exact cost characterization. The generic dual uses $\lambda = (a-b)/\log 2$ and $\mu = b/\log 2$; integer rounding strengthens it. The result concerns minimum deficit over the full specified catalog, not only exact realization of the constructed score. No assertion is made for $\varepsilon \geq 1/4$. □

What happens without the primitive slope cap? If the affine catalog additionally admits $z \mapsto Hz$ at cost b for arbitrarily large H, then $H\phi_n$ has deficit $1/H$ for $H \geq 2$. Thus $C_K^*(\varepsilon) \leq an + b$ for every $\varepsilon > 0$ in that modified catalog. This is an upper bound sufficient to refute any universal growing margin-cost term for unrestricted-gain parameter accounting.

B.3 Exact cost under noisy readout

Proof of Corollary 1. A feasible f satisfies $\mathcal{D}_1(\bar{f}) \leq 2r < 1/4$. The local slope argument in Theorem 4, with $\varepsilon = 2r$, yields $\mathrm{Lip}(f) \geq \mathrm{Lip}(\bar{f}) \geq K/(2r)$. Recovery gives at least $\lceil K(1-4r) \rceil > K/2$ sign alternations of f. If its composition has N_F two-piece and N_G affine modules, then $2^{N_F} > K/2$ and $2^{N_F+N_G} \geq K/(2r)$. $N_F \geq n$ and $N_F + N_G \geq n + m$ for $m = \lceil \log_2(1/(2r)) \rceil$. Since $a \geq b$, its cost is at least $an + bm$. Conversely, the admissible $f = 2^m \phi_n$ has $y^* f \geq 0$, so clipping does not change its hinge deficit 2^{-m}. Its readout excess risk is $2^{-m-1} \leq r$, attaining the bound. No clipping module is implemented. If one unrestricted affine gain costs b, gain at least $\max(2, 1/(2r))$ gives the $an + b$ upper bound. □

B.4 IRREGULAR-REGION CERTIFICATE

Proof of Proposition 2. Let B contain the cells having no point with $s_i f(x) > 0$. On each such cell, $\mathcal{D}_\tau(\bar{f})$ pays at least v_i. Choose a correctly signed point in each retained cell. Their signs have at most P alternations, since a continuous function affine on P intervals has at most P ordered sign alternations. Thus B is feasible in Eq. (19), and $2\mathcal{E}_\tau(f) \geq \sum_{i\in B} v_i \geq M_P$. Applying Eq. (9) to $\bar{f}$ and dividing by two proves the slope bound. The maximum is valid because both hold simultaneously; adding them would potentially charge the same loss twice. □

B.5 SAFE SELECTION FROM VALID CERTIFICATES

Proof of Corollary 8. Every score in an architecture class obeys its valid lower bound $L(P, G)$. a rejected class contains no score meeting the target. In the fixed cost and tie order, every class preceding the first feasible one is either correctly rejected by its bound or declared infeasible by its exact oracle. The first feasible class is never screened out and its oracle returns a feasible score. Hence both searches select the same minimum-cost feasible budget. If all classes are exhausted, none is feasible. The statement does not assert that the certificate alone decides feasibility, nor that it gives the global optimum outside the catalog. With approximate training, a feasible score would still certify an upper bound, but failure to find one would not justify an infeasibility declaration or minimum-cost conclusion. □

C PROOFS FOR STABILITY AND APPROXIMATE ARCHITECTURAL COSTS

The proofs follow Section 4. The tent constructions and geometric bounds used here are collected in Section D.

C.1 APPROXIMATION AND INVARIANCE DEFECTS

Proof of Theorem 5. Write $\nu_j = (T_j)_\# \nu_{j-1}$ and $\nu_0 = \nu$. For any measurable map T, preimages of measurable sets show that $\mathrm{TV}(T_\#\lambda, T_\#\kappa) \leq \mathrm{TV}(\lambda, \kappa)$. The triangle inequality gives

$$\mathrm{TV}(\nu_j, \mu) \leq \mathrm{TV}((T_j)_\#\nu_{j-1}, (T_j)_\#\mu) + \mathrm{TV}((T_j)_\#\mu, \mu) \leq \mathrm{TV}(\nu_{j-1}, \mu) + \alpha_j.$$

Induction and the upper bound one for total variation give $\mathrm{TV}((F_L)_\#\nu, \mu) \leq d_L$. Uniform approximation implies the event inclusions

$$\{|F_L - c| \leq \epsilon - \xi\} \subseteq \{|\widehat{f}| \leq \epsilon\} \subseteq \{|F_L - c| \leq \epsilon + \xi\}.$$

The total-variation bound applied to the two intervals proves Eq. (21). The difference between either enlarged or reduced reference mass and $H_{p,c}(\epsilon)$ integrates p over two strips of combined length 2ξ, hence is at most $2M\xi$.

For the composition bound, put $F_0 = \widehat{F}_0 = \mathrm{Id}$ and $E_j = \|\widehat{F}_j - F_j\|_\infty$. All intermediate inputs lie in $[0, 1]$, so

$$\begin{aligned} |\widehat{T}_j(\widehat{F}_{j-1}(x)) - T_j(F_{j-1}(x))| &\leq |\widehat{T}_j(\widehat{F}_{j-1}(x)) - T_j(\widehat{F}_{j-1}(x))| \\ &\quad + |T_j(\widehat{F}_{j-1}(x)) - T_j(F_{j-1}(x))| \\ &\leq e_j + A_j E_{j-1}. \end{aligned}$$

Iterating $E_j \leq e_j + A_j E_{j-1}$ from $E_0 = 0$ proves Eq. (22). No invertibility or absolute continuity of the intermediate pushforwards is required for this argument. □

C.2 APPROXIMATE TENT REALIZATIONS

Proof of Corollary 2. At consecutive grid sites $x_i = i/K$, $i = 0, \ldots, K$, the values of ϕ_L alternate between $-1/2$ and $1/2$. Consequently $|\widehat{f}(x_{i+1}) - \widehat{f}(x_i)| \geq 1 - 2\xi$. The secant quotient over an interval of length $1/K$ proves $\mathrm{Lip}(\widehat{f}) \geq K(1 - 2\xi)$. If $|\phi_L(x)| > \delta > \xi$, the approximation error cannot change its sign. Choose a nonempty closed subinterval in each of the $K + 1$ committed regions; Lemmas 5 and 6 give the piece and network bounds. Finally apply Theorem 5 with every $T_j = \Delta$, $\nu = \mu$ uniform, $c = 1/2$, and $\alpha_j = 0$. Then $H_{p,c}(\epsilon) = 2\epsilon$ and Eq. (23) follows. □

C.3 MLP–KAN PAYMENT SCHEDULES

Proof of Corollary 3. Part (i) is Theorem 9(ii,v). For part (ii), the tent map Δ is the order-1 spline with grid sites $0, 1/2, 1$ and corresponding values $0, 1, 0$. A width-one KAN layer has one scalar edge function, so composing L copies of this spline computes $\Delta^{\circ L}$; an affine readout subtracts $1/2$. The grid size is constant in every layer, hence the number of spline coefficients is $O(L)$.

For part (iii), removing the K collar intervals leaves $K+1$ ordered committed intervals with alternating signs. By Lemma 6, any continuous CPWL score realizing those labels has at least K affine pieces. A scalar order-1 spline with G affine grid intervals has at most G such pieces, so $G \geq K$.

Finally, Corollary 5 has small-margin constant $C = 2\delta/\tau$ under the uniform law, for which $m_0 = 1$. Applying Corollary 4 with $K = 2^L$ gives $\prod_\ell J_\ell \geq 2K/C = \tau K/\delta$. Equality with $\mathrm{Lip}(u_L)$ was established in Theorem 9(iv), proving part (iv). □

D SUPPORTING GEOMETRY AND EXPLICIT CONSTRUCTIONS

These results supply the geometric and constructive ingredients used above. Each supplementary statement is followed by its proof or by the lemmas needed to prove it.

D.1 INVERSE-SLOPE COLLARS AND JACOBIAN BUDGETS

Theorem 7 (Inverse-slope collar law)**.** *Let $f : [0,1] \to \mathbb{R}$ be continuous and piecewise affine, with its complete zero set consisting of distinct points $0 < r_1 < \cdots < r_K < 1$. Assume each r_j lies in the interior of an affine piece on which $f(t) = s_j(t - r_j)$ with $s_j \neq 0$. Then for all sufficiently small $\epsilon > 0$, $\lambda_1\{|f| \leq \epsilon\} = 2\epsilon \sum_{j=1}^{K} \frac{1}{|s_j|}$. If P_X has density bounded above by M, then $P_X(|f(X)| \leq \epsilon) \leq 2M\epsilon \sum_j |s_j|^{-1}$. If the density p is continuous at the roots,*

$$\lim_{\epsilon \downarrow 0} \frac{P_X(|f(X)| \leq \epsilon)}{2\epsilon} = \sum_{j=1}^{K} \frac{p(r_j)}{|s_j|}. \tag{37}$$

Proof of Theorem 7. Because each root lies in the interior of an affine piece, choose disjoint neighborhoods I_j of r_j on which $f(t) = s_j(t - r_j)$. For sufficiently small ϵ, the set $\{|f| \leq \epsilon\}$ is contained in $\bigcup_j I_j$ and intersects I_j in the interval $|t - r_j| \leq \epsilon/|s_j|$. Summing its lengths gives $2\epsilon \sum_j |s_j|^{-1}$. A density upper bound multiplies this Lebesgue measure by at most M. If p is continuous at each root, then

$$\frac{1}{2\epsilon} \int_{r_j - \epsilon/|s_j|}^{r_j + \epsilon/|s_j|} p(t)\, dt \longrightarrow \frac{p(r_j)}{|s_j|}, \tag{38}$$

and summing over finitely many roots proves Eq. (37). □

Theorem 8 (Root-conditioned Lipschitz converse)**.** *Let $u(x) = \eta(x) - 1/2$ be G-Lipschitz and have $K \geq 2$ ordered interior sign-changing zeros $r_1 < \cdots < r_K$. Set $\rho = \min\{r_1, 1 - r_K, \frac{1}{2}\min_{i \neq j} |r_i - r_j|\} > 0$. If P_X has density at least $m_0 > 0$, then for $0 < s \leq G\rho$,*

$$P_X(|\eta(X) - 1/2| \leq s) \geq \frac{2m_0 K s}{G}. \tag{39}$$

Thus an upper profile $P_X(|\eta - 1/2| \leq s) \leq Cs$ implies $G \geq 2m_0 K/C$.

Proof of Theorem 8. For every root r_j, Lipschitz continuity and $u(r_j) = 0$ give $|u(x)| \leq G|x - r_j|$. Hence $[r_j - s/G, r_j + s/G] \subseteq \{|u| \leq s\}$. The definition of ρ and $s \leq G\rho$ make these K intervals disjoint and contained in $[0,1]$. Their total length is $2Ks/G$, and the density lower bound gives Eq. (39). Combining it with an upper profile Cs and cancelling $s > 0$ yields the claimed bound on G. □

Corollary 4 (Architecture-neutral Jacobian budget). *Suppose u in Theorem 8 factors as $u = \Phi_{D+1} \circ \cdots \circ \Phi_1$, where each Φ_ℓ is globally Lipschitz CPWL, and set $J_\ell = \operatorname{ess\,sup}_z \|J\Phi_\ell(z)\|_2$. Assume also that $P_X(|u(X)| \le s) \le Cs$ for all sufficiently small $s > 0$, with $C > 0$. Then*

$$\prod_{\ell=1}^{D+1} J_\ell \ge \mathrm{Lip}(u) \ge \frac{2m_0K}{C}. \tag{40}$$

For a ReLU layer, $J_\ell \le \|W_\ell\|_2$; for an order-1 KAN layer, J_ℓ is the essential supremum of the spectral norm of the matrix of edge-spline derivatives.

Proof of Corollary 4. For a CPWL map Φ_ℓ and any two points a, b, the segment $[a, b]$ crosses finitely many affine cells. Integrating the directional derivative along that segment gives

$$\|\Phi_\ell(a) - \Phi_\ell(b)\|_2 \le J_\ell \|a - b\|_2. \tag{41}$$

Thus $\mathrm{Lip}(\Phi_\ell) \le J_\ell$, and composition yields

$$\mathrm{Lip}(u) \le \prod_{\ell=1}^{D+1} J_\ell. \tag{42}$$

The other inequality in Eq. (40) is Theorem 8 with $G = \mathrm{Lip}(u)$. For a ReLU layer, its Jacobian is $A(z)W_\ell$ with a diagonal matrix $A(z)$ whose entries lie in $\{0, 1\}$ almost everywhere, so $J_\ell \le \|W_\ell\|_2$. For an order-1 KAN layer, differentiating $\Phi_\ell(z)_j = \sum_i \psi_{\ell ji}(z_i)$ away from the knots gives the stated Jacobian entries; taking their essential supremum gives J_ℓ. □

D.2 Tent geometry and the depth-width tradeoff

Theorem 9 (Conservation along a near-tight ReLU depth-width tradeoff). *Let $L \ge 1$, $0 < \delta < 1/2$, and $0 < \tau \le 1/2$. Then:*

(i) The Bayes boundary has $\beta_0(\Gamma_L) = 2^L$.

(ii) The score ϕ_L is realized by a depth-L, width-2 ReLU network with $O(L)$ nonzero parameters, followed by one affine threshold, and width 2 is minimal. More generally, at every depth $D \ge 1$ the same score has an exact realization of width at most $2^{\lceil L/D \rceil}$; for $D \le L$, any partition $L = q_1 + \cdots + q_D$ gives layer widths 2^{q_ℓ}, and for $D > L$ the depth-L construction can be padded by identity layers.

(iii) If P_X has density at most M, then $\mathbb{P}_X(\mathrm{Col}_\delta(\phi_L)) \le 2M\delta$. For uniform P_X the collar mass equals 2δ, and the entire calibrated margin profile is independent of L as stated in Corollary 5. Every classifier $\widehat{g}$ agreeing with g_L^* outside this collar satisfies $R(\widehat{g}) - R^* \le 4M\tau\delta$.

(iv) The conserved probabilistic margin is purchased with geometric and norm cost. Under uniform P_X, with $K = 2^L$,

$$\mathbb{P}\big(\mathrm{dist}(X, \Gamma_L) \le r\big) = 2Kr, \quad 0 \le r \le \frac{1}{2K}, \tag{43}$$

so the median distance to the boundary is $1/(4K) = 2^{-L-2}$. Moreover, for $u_L := \eta_{L,\delta,\tau} - 1/2$,

$\mathrm{Lip}(\phi_L) = K, \mathrm{Lip}(u_L) = \frac{\tau K}{\delta}$.

Consequently, every exact continuous piecewise linear (CPWL) compositional realization of u_L has an effective layer-Jacobian product at least $\tau K/\delta$. In particular, every ReLU realization has product spectral norm at least $\tau K/\delta$, and every order-1 KAN realization obeys the corresponding spline-Jacobian bound.

(v) Let f be a scalar ReLU network with D hidden layers of width at most m. If $\operatorname{sign} f(t) = \operatorname{sign} \phi_L(t)$ for every $t \notin \mathrm{Col}_\delta(\phi_L)$, then $(m+1)^D \ge 2^L, \quad m \ge 2^{L/D} - 1$. Together with part (ii), this gives $2^{L/D} - 1 \le m_{\min} \le 2^{\lceil L/D \rceil}$, in addition to $m_{\min} \ge 2$.

Geometry of the iterated tent map.

Lemma 1 (Zeros, slopes, and collar length)**.** *For $L \geq 1$, the function $\phi_L = \Delta^{\circ L} - 1/2$ has the 2^L zeros in Eq.* (16)*. Between consecutive dyadic breakpoints it is affine with slope magnitude 2^L. For $0 < \delta < 1/2$,*

$$\mathrm{Col}_\delta(\phi_L) = \bigcup_{j=0}^{2^L-1} \left[\frac{2j+1}{2^{L+1}} - \frac{\delta}{2^L}, \frac{2j+1}{2^{L+1}} + \frac{\delta}{2^L} \right], \tag{44}$$

where the intervals are disjoint and contained in $[0,1]$. Consequently,

$$\lambda_1(\mathrm{Col}_\delta(\phi_L)) = 2\delta. \tag{45}$$

Proof. The tent map sends each of the intervals $[0, 1/2]$ and $[1/2, 1]$ linearly onto $[0, 1]$ with slope magnitude 2. Inductively, $\Delta^{\circ L}$ has 2^L monotone branches, each mapping its dyadic interval linearly onto $[0, 1]$ with slope magnitude 2^L. Each branch crosses level $1/2$ once, at its midpoint $(2j + 1)/2^{L+1}$. On either side of such a crossing, $|\phi_L(t)| = 2^L|t - r_j|$. Hence $|\phi_L| \leq \delta$ is precisely the interval of radius $\delta/2^L$ around r_j. Adjacent roots are spaced by 2^{-L}, so $\delta < 1/2$ makes these intervals disjoint. Their total length is $2^L(2\delta/2^L) = 2\delta$. □

Exact deep realization.

Lemma 2 (Minimal constant-width realization)**.** *The map ϕ_L is realized on $[0, 1]$ by a ReLU network with L hidden layers of width 2 and $O(L)$ parameters.*

Proof. On $[0, 1]$, the tent map has the three-hinge representation $\Delta(t) = 2\,\mathrm{ReLU}(t) - 4\,\mathrm{ReLU}(t - 1/2) + 2\,\mathrm{ReLU}(t - 1)$. Outside this interval the right-hand side is the zero-extended tent, not $1 - |1 - 2t|$. On the domain used here the last hinge vanishes, so

$$\Delta(t) = 2\,\mathrm{ReLU}(t) - 4\,\mathrm{ReLU}(t - 1/2), \qquad t \in [0, 1], \tag{46}$$

which is one hidden layer of width 2 followed by an affine combination. Because $\Delta([0, 1]) \subset [0, 1]$, every iterate remains in the domain of validity of Eq. (46), so composing L copies realizes $\Delta^{\circ L}$ on $[0, 1]$ using L width-2 hidden layers. The final affine map subtracts $1/2$. The construction has a constant number of nonzero parameters per layer. □

Lemma 3 (Width 1 is monotone)**.** *Every scalar ReLU network with one unit per hidden layer computes a (weakly) monotone function of its input. Consequently no width-1 network realizes the committed labels of ϕ_L for any $L \geq 1$.*

Proof. Each scalar hidden layer computes $u \mapsto \mathrm{ReLU}(au + b)$, which is weakly monotone, and the scalar affine readout is weakly monotone as well. A composition of weakly monotone scalar maps is weakly monotone. Such a function changes sign at most once, while the committed labels of ϕ_L alternate across $2^L + 1 \geq 3$ intervals, requiring at least two sign changes. □

Lemma 4 (Single-layer dyadic sawtooth and block realization)**.** *For every $q \geq 1$ and $t \in [0, 1]$,*

$$\Delta^{\circ q}(t) = 2^q\,\mathrm{ReLU}(t) + 2^{q+1} \sum_{j=1}^{2^q-1} (-1)^j\,\mathrm{ReLU}\left(t - \frac{j}{2^q}\right), \tag{47}$$

one hidden layer of width 2^q. Consequently, for every partition $L = q_1 + \cdots + q_D$, the score ϕ_L is realized on $[0, 1]$ by a ReLU network of depth D with layer widths $2^{q_1}, \ldots, 2^{q_D}$; a balanced partition ($q_\ell \leq \lceil L/D \rceil$) gives width at most $2^{\lceil L/D \rceil}$. For $D > L$, the depth-L width-2 realization can be padded by $D - L$ identity ReLU layers without increasing its width.

Proof. The right-hand side of Eq. (47) is continuous piecewise affine on $[0, 1]$ with breakpoints at $j/2^q$, value 0 at $t = 0$, and slope on $(j/2^q, (j+1)/2^q)$ equal to $2^q + 2^{q+1}\sum_{i=1}^{j}(-1)^i$, which is $+2^q$ for even j and -2^q for odd j. These are the breakpoints, initial value, and alternating slopes of $\Delta^{\circ q}$ (Lemma 1), and a one-dimensional CPWL function is determined by them. Since $\Delta^{\circ q}([0, 1]) \subseteq [0, 1]$, the blocks compose: $\Delta^{\circ L} = \Delta^{\circ q_D} \circ \cdots \circ \Delta^{\circ q_1}$ with each factor implemented by Eq. (47) on inputs in $[0, 1]$. The final affine map subtracts $1/2$. When $D > L$, every intermediate value lies in $[0, 1]$, so the identity is implemented by $z = \mathrm{ReLU}(z)$ and supplies the claimed padding. □

One-dimensional region bound.

Lemma 5 (Affine pieces of a one-dimensional ReLU network)**.** *Let $f : \mathbb{R} \to \mathbb{R}$ be computed by a ReLU network with D hidden layers of widths $m_1, \dots, m_D$. Then f is continuous piecewise affine with at most*

$$N_D \le \prod_{\ell=1}^{D} (m_\ell + 1) \tag{48}$$

affine intervals. In particular, if every $m_\ell \le m$, then $N_D \le (m+1)^D$.

Proof. Start with the single affine interval of the input line. Suppose all activations up to layer $\ell - 1$ are jointly affine on each cell of a partition with $N_{\ell-1}$ intervals. On one such interval, every preactivation in layer ℓ is affine and has at most one zero unless it is identically zero. Across m_ℓ neurons, at most m_ℓ distinct interior zero locations split the interval into at most $m_\ell + 1$ subintervals. ReLU is affine on each subinterval, so the entire layer is jointly affine there. $N_\ell \le (m_\ell + 1)N_{\ell-1}$. Induction yields the product bound, and the final affine readout creates no new breakpoints. □

Lemma 6 (Alternating committed intervals force affine pieces)**.** *Let $f : [0,1] \to \mathbb{R}$ be continuous piecewise affine. Suppose there are $K+1$ ordered nonempty closed intervals $I_0 < I_1 < \cdots < I_K$ on which f is strictly nonzero and its sign alternates from one interval to the next. Then f has at least K affine pieces whose closures intersect $f^{-1}(0)$; in particular it has at least K affine pieces.*

Proof. For each $j = 1, \dots, K$, continuity and opposite signs on I_{j-1} and I_j force $f^{-1}(0)$ to intersect the open gap between them. These K gaps are disjoint. Any affine piece that is not identically zero contains at most one zero; an identically-zero affine piece lies in only one of the disjoint gaps because the committed intervals themselves have strict sign. Hence the K gaps require at least K affine pieces intersecting the zero set. □

Proof of Theorem 9. Part (i) follows from Lemma 1: the boundary is a finite set of 2^L isolated points, hence its zeroth Betti number is 2^L. For part (ii), the depth-L, width-2 realization is Lemma 2; minimality of width 2 is Lemma 3; and the depth-D realizations at layer widths 2^{q_ℓ} are Lemma 4.

For part (iii), Lemma 1 gives collar Lebesgue measure 2δ. If the density is bounded by M, then $P_X(\mathrm{Col}_\delta) \le 2M\delta$. The standard binary excess-risk identity is

$$R(\widehat{g}) - R^* = \mathbb{E}\left[|2\eta(X) - 1|\, \mathbf{1}\{\widehat{g}(X) \ne g^*(X)\}\right]. \tag{49}$$

By assumption the indicator can be nonzero only on the collar. Under Eq. (17), $|2\eta - 1| \le 2\tau$ everywhere, so

$$R(\widehat{g}) - R^* \le 2\tau P_X(\mathrm{Col}_\delta) \le 4M\tau\delta. \tag{50}$$

For part (iv), put $K = 2^L$. The roots are the midpoints of the K equal subintervals of $[0,1]$. If $0 \le r \le 1/(2K)$, the K radius-r neighborhoods of the roots are contained in $[0,1]$ and overlap only on null endpoints. Their union has length $2Kr$, proving Eq. (43); setting $2Kr = 1/2$ gives the median. The slope calculation in Lemma 1 gives $\mathrm{Lip}(\phi_L) = K$. On every unsaturated collar interval, $u_L = (\tau/\delta)\phi_L$, while clipping cannot increase slope, so $\mathrm{Lip}(u_L) = \tau K/\delta$. For any exact CPWL factorization, Corollary 4 bounds its Lipschitz constant by the product of its effective layer Jacobians. That product is at least $\tau K/\delta$; the ReLU spectral-norm and order-1 KAN spline-Jacobian statements are the two specializations in that corollary.

For part (v), removing the 2^L collar intervals leaves $2^L + 1$ nonempty committed intervals on which $\operatorname{sign}\phi_L$ alternates. Any f agreeing with those signs satisfies the hypotheses of Lemma 6 with $K = 2^L$ and must have at least 2^L affine pieces. By Lemma 5, a depth-D, width-m ReLU network has at most $(m+1)^D$ pieces. Thus $(m+1)^D \ge 2^L$, which is equivalent to $m \ge 2^{L/D} - 1$. □

The tent map is the $k = 2, p \equiv 1$ instance of Theorem 1. Each iterate doubles the number of preimages and the slope, leaving their ratio fixed. Parts (ii) and (v) add the neural statement: composition represents this conserved pullback exponentially more compactly than fixed depth.

Approximation preserves these costs: if $\|\widehat{f} - \phi_L\|_\infty \le \xi < 1/2$, then $\mathrm{Lip}(\widehat{f}) \ge 2^L(1 - 2\xi)$ whenever $\widehat{f}$ is Lipschitz, and every continuous CPWL approximation still needs at least 2^L pieces. Under uniform input the collar-mass error is at most 2ξ when $\xi \le \epsilon$ and $\epsilon + \xi < 1/2$; see Corollary 2.

D.3 CALIBRATED MARGIN PROFILES AND EXCESS RISK

Corollary 5 (Invariant calibrated margin profile)**.** *Under uniform P_X, the regression model in Eq.* (17) *satisfies* $\mathbb{P}\big(|\eta_{L,\delta,\tau}(X) - 1/2| \le s\big) = \begin{cases} (2\delta/\tau)s, & 0 \le s < \tau, \\ 1, & s \ge \tau, \end{cases}$ *independently of L. For $0 < s < \tau$, equality holds in Eq.* (39)*. Under only $p_X \le M$, the probability is at most $(2M\delta/\tau)s$ for $0 \le s < \tau$.*

Proof of Corollary 5. For $0 \le s < \tau$, clipping is unsaturated on the event $|\eta_{L,\delta,\tau} - 1/2| \le s$, so this event is

$$|\phi_L(X)| \le \frac{\delta s}{\tau}. \tag{51}$$

By Lemma 1, its uniform probability is $2\delta s/\tau$. Since ϕ_L has slope magnitude 2^L on every monotone branch, the clipped regression score has Lipschitz constant $G_L = \tau 2^L/\delta$. Substitution into Eq. (39) with $K = 2^L$ and $m_0 = 1$ gives equality. Under a density upper bound M, the same Lebesgue set has probability at most M times its length. At $s = \tau$, every point satisfies $|\eta_{L,\delta,\tau} - 1/2| \le \tau$, and the same is true for all $s > \tau$. □

The jump at $s = \tau$ is the atom created by clipping; the unsaturated linear profile is invariant.

From collar mass to excess risk. The risk part of Theorem 9 is an instance of a general elementary fact. Let $B_\tau := \{x : |\eta(x) - 1/2| \le \tau\}$. We have

Proposition 3 (Boundary-strip excess-risk bound)**.** *For any binary classification problem and any classifier $\widehat{g}$ satisfying $\widehat{g} = g^*$ on $X \setminus B_\tau$, $R(\widehat{g}) - R^* \le 2\tau\, P_X(B_\tau)$.*

Proof of Proposition 3. Use the excess-risk identity Eq. (49). The disagreement event is contained in B_τ, where $|2\eta - 1| \le 2\tau$.

$$R(\widehat{g}) - R^* \le 2\tau P_X(B_\tau). \tag{52}$$

□

D.4 OTHER FOLDING BASES AND HYBRID SCHEDULES

Remark 1 (Other folding bases)**.** For $k \ge 2$, the uniform k-fold CPWL map Λ_k has k full branches, slopes of magnitude k, and invariant Lebesgue density. Its L-fold composition has k^L threshold preimages and the same collar profile by Theorem 1. The dyadic block construction in Theorem 9(ii) uses the identity $\Delta^{\circ q} = \Lambda_{2^q}$. More generally, with $\Lambda_1(t) = t$, subdivision of every branch gives the semigroup identity

$$\Lambda_r \circ \Lambda_s = \Lambda_{rs}, \qquad r, s \in \mathbb{N}. \tag{53}$$

Corollary 6 (Fold-and-refine hybrid schedule)**.** *Let $L \ge 0$, $R \ge 1$, $K = 2^L R$, and define*

$$f_{L,R}(x) := \Lambda_R\big(\Delta^{\circ L}(x)\big) - \frac{1}{2}. \tag{54}$$

Then $f_{L,R} = \Lambda_K - 1/2$, so it has K sign-changing zeros, $\mathrm{Lip}(f_{L,R}) = K$, and, for uniform X and $0 < \epsilon < 1/2$,

$$\mathbb{P}\big(|f_{L,R}(X)| \le \epsilon\big) = 2\epsilon. \tag{55}$$

It is realized by L width-2 ReLU fold blocks followed by one scalar order-1 spline with R affine grid intervals, using $O(L + R)$ nonzero parameters. Conversely, a shallow scalar CPWL score realizing its committed alternating labels outside any sufficiently small disjoint collar of the zeros needs at least K affine pieces.

Suppose a fold block costs $a > 0$, a residual spline branch costs $b > 0$, and fixed overhead costs c. For a fixed K, every feasible exact factorization $R = K/2^L$ has representation cost

$$J_K(L) = aL + bK2^{-L} + c. \tag{56}$$

Let $L_{\max} := \max\{\ell : 2^\ell \text{ divides } K\}$. In the continuous relaxation $0 \le L \le L_{\max}$, the minimizer is

$$L^* = \left\lceil \log_2\!\left(\frac{bK \ln 2}{a}\right) \right\rceil_0^{L_{\max}}, \qquad R^* = K2^{-L^*}, \tag{57}$$

where $[z]_u^v := \min\{v, \max\{u, z\}\}$. *At an interior optimum,* $R^* = a/(b \ln 2)$ *is independent of* K*; the exact discrete optimum is attained at one of the neighboring feasible integer depths.*

Proof of Corollary 6. To verify Eq. (53), partition each of the s full branches of Λ_s by the preimages of the $r-1$ interior breakpoints of Λ_r. This produces rs consecutive equal affine intervals. On them $\Lambda_r \circ \Lambda_s$ takes alternating endpoint values zero and one and has slope magnitude rs, which uniquely identifies Λ_{rs}. Thus, using $\Delta^{\circ L} = \Lambda_{2^L}$,

$$\Lambda_R \circ \Delta^{\circ L} = \Lambda_R \circ \Lambda_{2^L} = \Lambda_{2^L R} = \Lambda_K. \tag{58}$$

The map Λ_K has K full affine branches, each with slope magnitude K and one crossing of level $1/2$. This proves the zero count and Lipschitz claim. Since Λ_K preserves Lebesgue measure, the preimage of $[1/2-\epsilon, 1/2+\epsilon]$ has measure 2ϵ, proving Eq. (55).

Each copy of Δ is the width-2 ReLU block in Lemma 2. The map Λ_R is an order-1 scalar spline with breakpoints j/R, $j = 0, \ldots, R$, and alternating values zero and one, so its readout uses $R+1$ grid values. This gives $O(L+R)$ parameters. After removing sufficiently small disjoint collars around the K zeros, the remaining $K+1$ committed intervals have alternating signs; the shallow lower bound follows from Lemma 6.

Substituting $R = K2^{-L}$ gives Eq. (56). Its derivatives on the continuous relaxation are

$$J_K'(L) = a - bK(\ln 2)2^{-L}, \qquad J_K''(L) = bK(\ln 2)^2 2^{-L} > 0. \tag{59}$$

Hence the unconstrained stationary point is $\log_2(bK \ln 2/a)$, strict convexity makes it the unique unconstrained minimum, and clipping to $[0, L_{\max}]$ proves Eq. (57). At an interior optimum, substitution yields $R^* = K2^{-L^*} = a/(b \ln 2)$. Convexity also implies that the discrete minimizer is one of the neighboring feasible integer depths. □

The hybrid assigns repeated dyadic topology to the ReLU composer and the non-dyadic residual to the spline readout. Eq. (55) shows that changing this payment schedule changes neither the realized score nor its margin profile. Optimization is not included in Eq. (56); E6 compares its exact representation cost with the outcomes of specified training protocols.

D.5 Higher-dimensional coarea

Let $\Omega \subset \mathbb{R}^d$ be open and $f : \Omega \to \mathbb{R}$ be Lipschitz. Let P_X have a fixed Borel representative p of its density, and assume $\|\nabla f\|_2 > 0$ almost everywhere on $\{|f| \le \epsilon_0\}$. For almost every t, define $\mathfrak{C}_p(f;t) := \int_{f^{-1}(t)} \frac{p(x)}{\|\nabla f(x)\|_2}\, d\mathcal{H}^{d-1}(x)$. The integrand is taken to be zero where the approximate gradient is undefined or vanishes. Under the stated assumptions, this convention does not change the quantity for almost every level. The coarea formula (Evans and Gariepy, 2015) yields

Proposition 4 (Probability-weighted coarea complexity)**.** *For* $0 < \epsilon \le \epsilon_0$,

$$P_X(|f(X)| \le \epsilon) = \int_{-\epsilon}^{\epsilon} \mathfrak{C}_p(f;t)\, dt \le 2\epsilon \operatorname*{ess\,sup}_{|t| \le \epsilon} \mathfrak{C}_p(f;t). \tag{60}$$

Proof of Proposition 4. The coarea formula for a Lipschitz scalar map states that for nonnegative measurable g,

$$\int_\Omega g(x)\|\nabla f(x)\|_2\, dx = \int_{\mathbb{R}} \int_{f^{-1}(t)} g(x)\, d\mathcal{H}^{d-1}(x)\, dt. \tag{61}$$

Apply it on $\{|f| \le \epsilon\}$ with $g(x) = p(x)/\|\nabla f(x)\|_2$, using the nonvanishing-gradient assumption almost everywhere in the collar. This gives the equality in Eq. (60); bounding the integrand by its essential supremum over $t \in [-\epsilon, \epsilon]$ gives the inequality. □

D.6 Tensorized boundary compression

The one-dimensional result embeds into any $\mathbb{R}^d$ by depending on one coordinate. To create many Bayes cells, define $B(a,b) := \frac{1}{2}\big(|a+b| - |a-b|\big) = \operatorname{sign}(ab)\min\{|a|, |b|\}$, let B_d be a balanced binary tree of B gates, and set

$$F_{d,L}(x) := B_d\big(\phi_L(x_1), \ldots, \phi_L(x_d)\big). \tag{62}$$

Corollary 7 (Tensorized boundary compression). *On $[0,1]^d$, $F_{d,L}$ is realized by a ReLU network of depth $O(L+\log d)$, width $O(d)$, and $O(dL)$ nonzero parameters. The complement of its zero set has $(2^L+1)^d$ relatively open sign cells. Under uniform measure, $\lambda_d(\mathrm{Col}_\delta(F_{d,L})) = 1-(1-2\delta)^d \le 2d\delta$. Under a density bound $dP_X/d\lambda_d \le M$, the probability is at most $2Md\delta$. Any depth-D, width-m ReLU realization of the committed checkerboard labels obeys $m \ge 2^{L/D}-1$ by one-dimensional restriction. For uniform measure, its coarea complexity satisfies $\mathfrak{C}_1(F_{d,L};0) = d$, independently of L.*

Indeed, the zero set consists, up to lower-dimensional intersections, of $d2^L$ unit-area coordinate hyperplanes, and $\|\nabla F_{d,L}\|_2 = 2^L$ almost everywhere on them. Coarea recovers the same topology-to-slope cancellation.

Proof of Corollary 7. The identity

$$B(a,b) = \frac{1}{2}(|a+b| - |a-b|) = \mathrm{sign}(ab)\min\{|a|,|b|\} \tag{63}$$

is checked by the four sign cases. Since $|z| = \mathrm{ReLU}(z) + \mathrm{ReLU}(-z)$, one B gate has constant ReLU size. A balanced tree of B gates computes

$$F_{d,L}(x) = \left(\prod_{i=1}^{d} \mathrm{sign}\,\phi_L(x_i)\right) \min_i |\phi_L(x_i)| \tag{64}$$

with depth $O(\log d)$ and size $O(d)$ after the coordinatewise folds. The d copies of ϕ_L are computed in parallel using depth L, width $2d$, and size $O(dL)$. This proves the network bounds.

Each coordinate has 2^L zeros and 2^L+1 sign intervals. Their Cartesian products give $(2^L+1)^d$ open cells, and the sign of $F_{d,L}$ is constant on each cell.

Because $|F_{d,L}| = \min_i |\phi_L(x_i)|$,

$$\mathrm{Col}_\delta(F_{d,L}) = \bigcup_{i=1}^{d} \{x : |\phi_L(x_i)| \le \delta\}. \tag{65}$$

Under uniform product measure, the complement requires all d coordinates to avoid a one-dimensional set of measure 2δ, hence its mass is $(1-2\delta)^d$. The collar mass is $1-(1-2\delta)^d \le 2d\delta$ by the union bound. A density bound by M multiplies the Lebesgue-volume bound by M.

Finally choose any line on which $x_2,\dots,x_d$ are fixed strictly inside sign intervals of their respective ϕ_L. Along this line the sign of $F_{d,L}$ is a constant multiple of $\mathrm{sign}\,\phi_L(x_1)$. Any network realizing the committed d-dimensional labels realizes the one-dimensional alternation on that restriction, and Theorem 9(v) applies.

Away from intersections of coordinate hyperplanes, one coordinate attains the minimum in $|F_{d,L}| = \min_i |\phi_L(x_i)|$. There $\|\nabla F_{d,L}\|_2 = 2^L$. The zero set contains $d2^L$ unit-area coordinate hyperplanes, whose intersections have $\mathcal{H}^{d-1}$-measure zero. Hence $\mathfrak{C}_1(F_{d,L};0) = d2^L/2^L = d$. □

E Separation and Structural Discovery

E.1 Touching regions and positive margin

Proposition 3 is a risk *upper bound*. In particular, touching Bayes cells can prevent a continuous encoder from opening a positive metric margin, but they do not prevent the Bayes classifier itself from attaining Bayes risk. We record the distinction explicitly.

Proposition 5 (Touching cells obstruct continuous positive margin, not Bayes risk). *Let $C_0, C_1 \subset X$ satisfy $\mathrm{cl}_X(C_0) \cap \mathrm{cl}_X(C_1) \neq \varnothing$, where the closures are relative to X, and let $\Phi : X \to Z$ be continuous into a metric space. Then $\mathrm{dist}(\Phi(C_0), \Phi(C_1)) = 0$ and an exact positive-margin separation by a continuous encoder is impossible.*

Proof of Proposition 5. Choose $x_n \in C_0$ and $y_n \in C_1$ converging to a common point $x_\star \in \mathrm{cl}_X(C_0) \cap \mathrm{cl}_X(C_1)$. By continuity, $\Phi(x_n) \to \Phi(x_\star)$ and $\Phi(y_n) \to \Phi(x_\star)$. Hence $d_Z(\Phi(x_n), \Phi(y_n)) \to 0$, which forces the set distance between $\Phi(C_0)$ and $\Phi(C_1)$ to be zero. □

E.2 TASK-COMPATIBLE REPRESENTATIONS

This section formalizes the elementary facts behind Section 5.2; they are not claimed as new separation theorems. In particular, sufficiency of a representation and cost of implementing it are separate questions.

Proposition 6 (Fiber compatibility and metric realization)**.** *Let $z : X \to Z$ and let $\{y_t : X \to \{-1, 1\} : t \in \mathcal{T}\}$ be a specified family of deterministic classification tasks. There exist set maps $h_t : z(X) \to \{-1, 1\}$ satisfying $y_t = h_t \circ z$ for all t if and only if*

$$z(x) = z(x') \implies y_t(x) = y_t(x') \quad \text{for every } t \in \mathcal{T}. \tag{66}$$

No continuity, measurability, efficient computation or learnability of the induced maps follows from this set-theoretic statement alone. If X, Z are metric spaces, z is L_z-Lipschitz and a real-valued readout r is L_r-Lipschitz, then $f = r \circ z$ obeys

$$\mathrm{Lip}(f) \le L_r L_z. \tag{67}$$

If also $f(a) \le -m$ and $f(b) \ge m$ for some $m > 0$, then $2m \le L_r d_Z(z(a), z(b)) \le L_r L_z d_X(a, b)$.

Proof. Necessity of Eq. (66) follows because a function has one value at each argument. For sufficiency, assign $h_t(v)$ the common value of y_t on the nonempty fiber $z^{-1}(v)$; the condition makes this well defined. Finally, $|r(z(x)) - r(z(x'))| \le L_r d_Z(z(x), z(x')) \le L_r L_z d_X(x, x')$. The separated-pair inequality follows by inserting the required score gap. □

E.3 INFORMATION PRESERVATION AND METRIC REALIZATION

For the representation in metric-topology factorization, equality of $q(x)$ and $q(x')$ does not imply equality of $z(x)$ and $z(x')$. A metric coordinate may retain task information within a shared chart. Conversely, once the full representation violates Eq. (66), changing its downstream readout cannot repair exact classification. One must refine the encoder, retain another coordinate, add routing information, or accept the error. Two isolated conflicting points do not themselves imply a positive population-risk lower bound. Our confidence-weighted recovery theorem provides the needed mass-sensitive statement on ordered label regions.

E.4 COORDINATE CONTRACTION AND REUSABLE OPERATIONS

Let a normed latent space be rescaled by $z_s = sz$, $s > 0$, and set $r_s(v) = r(v/s)$. The composite score is unchanged, while its factor bounds become sL_z and L_r/s; their product is unchanged. Contracting coordinates does not alone demonstrate a cheaper reliable classifier. A reusable metric operation can still be valuable: several contexts may share one operation while retaining different calibrated readouts or routes. Its benefit must be measured against the cost of those interfaces and at a fixed task and reliability scale. None of the present theorems proves that every contraction reduces neural width, or that score scaling improves normalized input robustness.

E.5 EXPANDING TASK FAMILIES AND THE SCOPE OF CONSERVATION

If $\mathcal{T} \subset \mathcal{T}'$ and a representation satisfies Eq. (66) only for $\mathcal{T}$, a new task can reveal an erased distinction. Refinement is then required for exact sufficiency even if the old classifier was correct. This explains one motivation for structural discovery, but the present results do not establish how to find a witness from finite data or choose an optimal sequence of refinements. On a fixed task, E11 uses known population structure and E12 chooses among supplied repairs from finite diagnostics; neither implements this expanding-task process.

What is and is not conserved. The exact collar identity assumes a full-branch encoder and invariant input law. Fiber compatibility is instead a task-information constraint; it does not preserve connected components or arbitrary homology. The architectural inequality Eq. (18) further needs the stated relationship between module capabilities and prices. Treating these three statements as one unrestricted conservation law would obscure their different assumptions. Together they motivate a research program: discover task-compatible structure, optimize its metric realization, and charge the full encoder–router–readout system. They do not yet constitute a general structural-learning theorem.

F E1–E6: Architecture Comparisons and Hybrid Allocation

This section gives the protocols behind Figures 2 to 4 and the expanded E6 results. These training experiments are exploratory: their evaluation splits participate in model or checkpoint selection.

F.1 E1–E5: model implementations and selection protocol

Models in E1–E5 are implemented in numpy with hand-derived gradients (no deep-learning framework), using explicit order-1 interpolation. This implementation choice does not itself establish the correctness of its gradients. A KAN layer stores, for each edge (j, o), values at $G+1$ uniform knots and computes $y_o = \sum_j \mathrm{interp}(x_j; V_{jo})$; the first layer's grid spans the data range, hidden grids span $[-1.5, 1.5]$ with inputs clipped (clip gradients masked). Hidden splines start from random signed linear profiles plus noise. Training uses binary cross-entropy and Adam with schedule $.02(1 + \cos(\pi s/S))/2 + .0002$, batch size 256, and $S = 2{,}400$ steps for E1–E4 ($S = 1{,}200$ for E5), 8,000/4,000 training/evaluation points, and two seeded restarts per configuration with best-of reporting. The smallest-tested-configuration protocol walks a growing configuration ladder (MLP: depth–width pairs; KAN: knot counts; deep KAN: depths at fixed grid) and stops at the first configuration whose best restart reaches 95% evaluation accuracy; if none does, the point is reported as censored at the ladder cap, never silently dropped. E3 uses 361 MLP, 252 shallow-KAN, and 162 deep-KAN parameters; the revised legend reports those counts without claiming budget parity. The original crease diagnostic counted second differences above a threshold on $8{,}192$ probe points and could count one kink twice. The original deep spline also replaced its last fold by an affine map. Neither diagnostic is used for the revised constructed panel.

Corrected E4a construction check. All three constructions now realize $\phi_L(x_1)$ on the probe $t \mapsto (t, 0.3)$, for $L = 1{:}8$. The MLP uses the width-two tent chain of Lemma 2; the shallow KAN interpolates ϕ_L on the $2^L + 1$ dyadic grid sites. The deep KAN uses L tent-spline layers with seven knots per edge, subtracting $1/2$ from the final tent rather than replacing it by an affine layer. Allocated parameter counts, including zeros, are $6L + 3$, $2(2^L + 1)$, and $7(L + 1)$, respectively. Each score has $2^L - 1$ interior creases and slope magnitude 2^L. We count one slope change per dyadic cell interface and verify scores on $8{,}193$ uniform and $4{,}096$ random points. Closed-interval spline interpolation removes the original implementation's artificial 10^{-9} endpoint offset. All 24 constructions pass the checks; the script and numerical records are included. The E4b training results are unchanged, drawn from the saved per-arm console records at $L = 5$: .509, .989, and .890 for 393, 130, and 162 parameters. These trained architectures are not the minimal constructions plotted in E4a, so the panels are not a matched optimization ablation.

For E5 the training labels are drawn from $\eta = \frac{1}{2}$ inside the band $\{|\phi_2(x_1)| < \rho/2\}$ of empirical mass $\approx \rho = 0.2$ and are deterministic outside it; both arms train at generous budgets (MLP $[2, 16, 16, 16, 1]$; KAN $[2, 4, 1]$ with $G = 32$). The experiment record identifies `mlp_kan_experiments.ipynb` as the implementation. The selection protocol uses the evaluation split for choosing configurations and restarts; the resulting performance values do not constitute independent final-test estimates. They are preserved here as exploratory records. Censoring indicates failure within a finite configuration and optimization budget, not a proof of unrepresentability. The recovered notebooks and release audit supply the source and six selected finite-difference gradient checks, all passing their stated tolerance. These local checks do not prove optimizer correctness or retroactively make the selection split an independent test set. Revised E1–E2 figures place every censored marker at its actual last tested count, including the E1 deep-KAN cap of 130 rather than the original 95. The censor statuses are unchanged. Saved per-arm console records are used for plotting; separately rounded summary lines can differ in their last digit.

F.2 E6: fold-and-refine methods

E6 uses the target labels $y = \mathbf{1}\{\Lambda_K(x) > 1/2\}$ with $x \sim \mathrm{Unif}[0, 1]$. Each fold is $u \mapsto \alpha u - \beta\,\mathrm{ReLU}(u - \theta)$ with three trainable parameters, and the readout linearly interpolates $G + 1$ trainable values on a uniform grid. Constructed models use exact tent parameters $(2, 4, 1/2)$ and exact zigzag grid values. Their logit is $4f_{L,R}$, so the reported .300 collar is $\{|4f_{L,R}| \le 4\delta\}$ with $\delta = .15$. Trained models use $G = \max\{4, 2R\}$, $8{,}000$ training samples, batches of $1{,}024$, binary cross-entropy, $4{,}000$ Adam steps with cosine decay, and seeds $\{0, 1, 2\}$. Cold initialization uses near-

identity folds and a nearly flat spline at learning rate .02; warm initialization adds independent standard-deviation .01 perturbations to the exact construction and uses nominal learning rate .003. The original schedule is $\mathrm{lr}(1 + \cos(\pi s/S))/2 + .001$, so the cold and warm runs start at .021 and .004 and approach .001. Accuracy, crossing count, and slope diagnostics are evaluated on a fixed 16,384-point grid. The $K = 96$ comparison is constructed. The experiment record identifies `hybrid_fold_refine.ipynb` as the implementation. The reported means are not accompanied by additional seed-level statistics. The constructed readout uses $G = 2R$; the trained readout uses $G = \max\{4, 2R\}$. Thus $L = 6, R = 1$ has 21 constructed and 23 trained parameters. The separate dyadic-only baselines in the $K = 96$ table retain their reported 20 and 23 parameter counts. The recovered notebook uses $G = 1$ for those pure-fold baselines, explaining their 20 and 23 parameter counts; they are not used to compute the hybrid optimum. The main comparison uses the fully specified spline-readout family.

F.3 E6: EXPANDED RESULTS AND INTERPRETATION

The expanded E6 sweep in Figure 5 and Table 1 uses the per-depth parameter convention specified in the preceding protocol. It supplements the main comparison in Figure 4.

(a) $K = 64$: exact schedule and optimization

L	0	1	2	3	4	5	6
$R = 64/2^L$	64	32	16	8	4	2	1
constructed parameters	129	68	39	26	21	20	21
trained parameters	129	68	39	26	21	20	23
cold accuracy	.990	.823	.528	.489	.516	.502	.501
cold crossings	64.0	59.7	17.0	8.0	5.7	1.0	.3
warm accuracy	.987	.986	.982	.983	.985	.976	.990
warm crossings	64	64	64	64	64	64	64

(b) $K = 96$: constructed residual test

schedule	crossings	params	accuracy
binary folds, $L = 6$	64	20	.500
binary folds, $L = 7$	128	23	.500
hybrid, $(L, R) = (5, 3)$	96	22	1.000
shallow spline, $G = 192$	96	193	1.000

Table 1: Fold-and-refine payment schedules. In (a), all constructed rows have 64 crossings, accuracy 1.000, and collar mass .300; trained rows are means over three seeds. The trained $L = 6$ readout uses the protocol's minimum four-interval grid, hence 23 parameters. In (b), all rows are constructed and the spline rows use the common two-interval-per-branch grid. Binary-fold aliasing is a statement about these uniform tent schedules, not an impossibility result for arbitrary deep networks.

Exact representation cost. For $K = 64$, the hybrid uses L constrained fold units $u \mapsto \alpha u - \beta\,\mathrm{ReLU}(u - \theta)$ and a spline with $G = 2R$ intervals. Counting three free scalars per fold and $G + 1$ grid values gives $J(L) = 3L + 2K2^{-L} + 1$. Its continuous optimum is $L^* = 4.89$; the discrete minimum is 20 parameters at $(L, R) = (5, 2)$. All exact factorizations have 64 crossings and unit accuracy, with collar mass .300 at $\delta = .15$. These counts concern a constrained fold parameterization and fixed grid sites, not arbitrary dense-layer weights.

Training exposes a protocol gap, not an initialization-only effect. Across three seeds, cold-start Adam reaches .990 grid accuracy at $L = 0$ and approximately chance for $L \geq 2$, with at most 17 crossings in those deeper configurations. Near-exact warm starts retain all 64 crossings and .976–.990 accuracy (Figure 5 and Table 1). Cold and warm runs use learning rates .02 and .003, respectively; these observations compare complete protocols. They show that the tested cold protocol fails to find available solutions, while the warm protocol maintains them. They do not isolate initialization or establish an optimizer-independent barrier. The independent control below crosses these factors and freezes exact folds to distinguish their effects within a new, fully specified implementation.

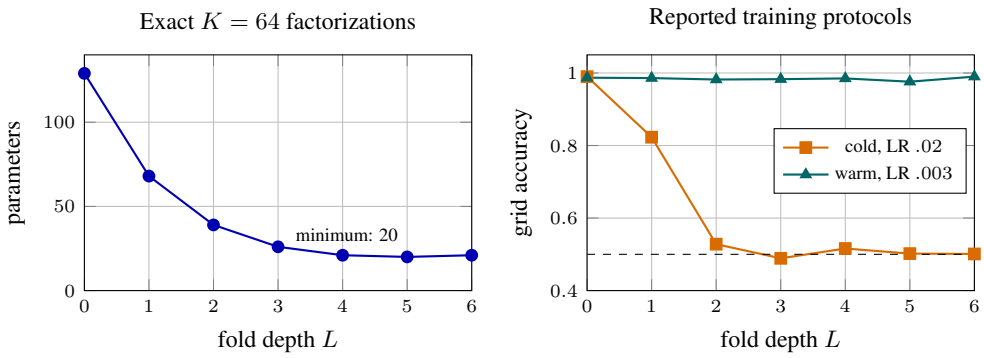


Figure 5: **E6: exact cost and protocol-dependent training outcomes.** Left: parameter counts for exact factorizations of the same 64-crossing score. Right: reported mean accuracy over three seeds; initialization and learning rate both differ. Warm starts are perturbations of the known exact factorization. The trained $L = 6$ model uses $G = 4$ and 23 parameters, whereas its constructed counterpart uses $G = 2$ and 21 parameters.

Residual structure and comparison scope. For $K = 96 = 2^5 \cdot 3$, the constructed hybrid uses 22 parameters versus 193 for the shallow spline with the common two-interval-per-branch grid; minimal grids give 19 versus 97. The prescribed dyadic-only schedules with 64 and 128 crossings achieve .500 accuracy on this target. This is a restriction of that fold dictionary: a three-branch ReLU fold is available under Remark 1. The hybrid optimum is conditional on the allowed primitives and costs, not a universal advantage over ReLU networks.

F.4 E6-C: INDEPENDENT INITIALIZATION AND LEARNING-RATE CONTROL

The new implementation tests the $K = 64$ target at $L \in \{3, 5\}$, with $R = 64/2^L$, $G = \max\{4, 2R\}$, and 26 or 20 stored parameters. Cold starts use identity folds $(1, 0, 1/2)$ and independent $N(0, .02^2)$ readout ordinates. Warm starts perturb every parameter of the exact fold-and-readout construction by independent $N(0, .01^2)$ noise. A third arm fixes exact folds and trains a readout initialized as in the cold arm; its trainable parameter counts are 17 and 5, respectively, while all fold parameters remain part of the stored model. Each arm is crossed with initial learning rates .003 and .02. Five seeds give $2 \times 3 \times 2 \times 5 = 60$ runs. The folds have the form specified for E6; only the final spline input is clipped to $[0, 1]$, with zero derivative outside that interval. These trainable folds do not impose the exact-cost theorem's slope cap.

Each seed supplies independent uniform training, validation, and final-test samples of sizes 8,000, 4,000, and 16,000. All six arms share those samples and the same minibatch sequence, with 2,000 Adam updates, batch size 256, binary cross-entropy, and cosine learning-rate decay. Within each initialization, both rates start from identical parameters. Checkpoints are evaluated every 100 steps, and the one minimizing validation cross-entropy is tested once. Crossings are measured afterward on a fixed 16,384-point midpoint grid. No test-based tuning is performed. The protocol differs from original E6; its results remain separate from the subsequently recovered notebook and its implementation audit.

Cold starts remain near chance at either rate. Warm starts perform much better at .003 but lose accuracy and crossings at .02. Frozen exact folds exceed .986 mean accuracy at both rates, showing that the supplied coordinate makes residual fitting tractable in the tested budget. Thus the observed gap involves both the starting representation and its stability under updates. Warm starts use privileged knowledge of the target; the control establishes neither an optimizer-independent obstruction nor an algorithm for discovering the folds. Five seeds characterize variability within the prescribed setup, not robustness across tasks or optimizers.

Central finite differences with step 10^{-7} check every parameter at $(L, G) = (0, 7), (1, 8), (3, 16), (5, 4)$, using random inputs away from knots and including clipped inputs. The largest absolute discrepancy is 3.59×10^{-9}; the largest relative L2 discrepancy (normalizing by $\max\{1, \|g_{\text{numerical}}\|_2\}$) is 1.35×10^{-9}. The script `experiments/e6_controlled.py` runs

initialization	initial LR	$L = 3$ test accuracy	$L = 3$ crossings	$L = 5$ test accuracy	$L = 5$ crossings
Cold	0.003	0.494 ± 0.017	8.6	0.498 ± 0.004	1.8
Cold	0.02	0.508 ± 0.011	7.8	0.503 ± 0.008	1.0
Warm	0.003	0.977 ± 0.005	63.2	0.943 ± 0.018	61.6
Warm	0.02	0.626 ± 0.067	26.0	0.646 ± 0.131	26.8
Frozen folds	0.003	0.987 ± 0.003	64.0	0.996 ± 0.003	64.0
Frozen folds	0.02	0.995 ± 0.001	64.0	0.998 ± 0.002	64.0

Table 2: Independent E6-C control: accuracy mean $\pm$ sample standard deviation and mean crossing count over five seeds. Exact constructions have accuracy 1 and 64 crossings. The frozen arm trains only the readout.

these checks before training. The release includes all seed-level results, validation trajectories, selected parameters, and runtime versions. The gradient checks cover the new implementation only; original E1–E6 optimization claims remain qualified. An independent E1–E5 replication with final-test splits and a broader optimizer search is still needed.

G E7–E8: Numerical Checks of Recovery and Resource Bounds

Can resource bounds guide architecture selection? E7–E9 numerically check the margin-deficit and recovery bounds, the predicted architectural costs, and the role of noise location in reliable readout (Sections G and H). Building on these checks, E11 uses resource certificates to rule out inadequate architecture budgets before optimization. Across 15 selection tasks, screening preserves every minimum-cost catalog choice while avoiding 38.0% of dynamic-programming work. The benefit is reduced search effort while achieving the same model cost and risk. These controlled results demonstrate how necessary resource bounds can guide architectural investment when task structure is known. E12 further tests finite-observation repair choices within a supplied menu. Protocols, full results, and limitations for E7–E12 are in Sections G to J.

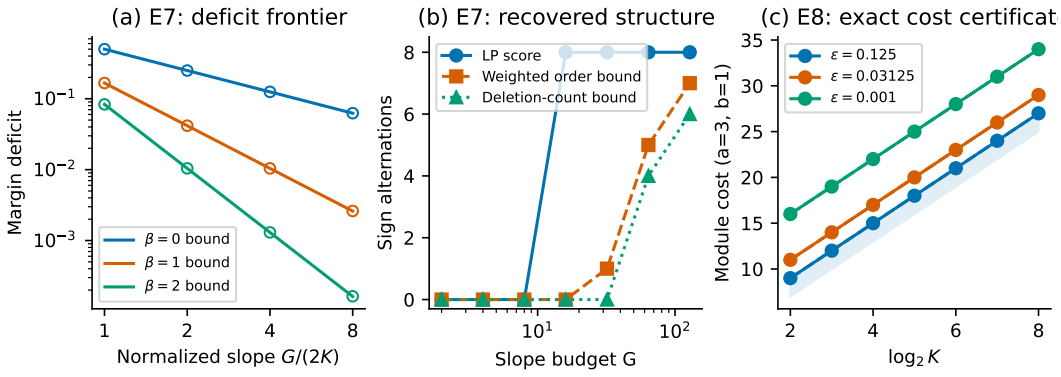


Figure 6: **E7–E8: direct numerical tests.** (a) Lines are the analytic margin-deficit bound; circles are independently integrated convex solutions ($K = 16$; all K in the release). (b) Unequal-confidence cells: observed alternations and the two recovery lower bounds. (c) Circles are feasible construction costs and dashed lines are matching weighted lower bounds. Shading shows the gap from the coarser deletion-count bound, at most 2 for $(a, b) = (3, 1)$.

E7 tests the margin–slope and recovery bounds in Theorems 2 and 3; E8 checks the catalog cost in Theorem 4. These controlled numerical studies complement the exploratory training protocols in Section F. The script `experiments/validate_resources.py` regenerates all CSV files, `results/summary.json`, and Figure 6. The recorded environment is Python 3.12.14, NumPy 2.3.5, and SciPy 1.17.0.

G.1 E7A: INDEPENDENTLY OPTIMIZED MARGIN-DEFICIT FRONTIER

Let $y^* = \text{sign}(\Lambda_K - 1/2)$, $p = 1$, $\tau = 1$, and $w(x) = (2|\Lambda_K(x) - 1/2|)^\beta$, with $\beta \in \{0, 1, 2\}$. This can also be interpreted as Bayes confidence for $\eta(x) = (1 + y^*(x)w(x))/2$. Around every root, $\rho = 1/(2K)$ and $pw = \rho^{-\beta}t^\beta$. We use $K \in \{8, 16, 32\}$ and $G/(2K) \in \{1, 2, 4, 8\}$. For each setting, SciPy's HiGHS solver minimizes midpoint-quadrature hinge loss over scalar values $f_i \in [-1, 1]$, with constraints $|f_{i+1} - f_i| \leq G/N$ on $N = 128K$ and $N = 256K$ midpoint nodes. Hinge slack variables make this a linear program. Clipping to $[-1, 1]$ cannot increase hinge loss or Lipschitz constant, so this range constraint introduces no restriction on the continuous optimum.

The solution is extended by linear interpolation and constant endpoint segments, giving an actual G-Lipschitz score. Its deficit is then evaluated independently of the optimization objective: partition at interpolation knots, target/confidence breakpoints, and unit-margin crossings, and apply three-point Gauss quadrature on each interval. For these integer confidence exponents, the integrands on each interval are polynomials of degree at most three, so the evaluation is exact up to floating-point error. The continuous lower bound is Eq. (9). In this regime it is attained by clipping G times the signed distance to the roots to $[-1, 1]$, providing an analytic optimum independent of the LP implementation.

All 72 programs solved successfully. At the finer resolution, the largest relative excess above the continuous optimum is 0.5861%; the largest absolute change between resolutions is 1.832×10^{-4}. The maximum recorded positive violation of the interpolant's slope budget is zero at reported precision. These errors concern discretization, not statistical uncertainty. The data are in `results/E7_frontier.csv`.

G.2 E7B: LOSS PERMITS DELETION ACCORDING TO CONFIDENCE MASS

For $K = 8$ there are nine alternating cells, of widths $(1/2, 1, 1, 1, 1, 1, 1, 1, 1/2)/8$. We use cell-constant confidence profiles $(1, \ldots, 1)$, $(.12, 1, .12, 1, .12, 1, .12, 1, .12)$, and $(1, 1, .08, .08, .08, .08, .08, 1, 1)$, and slope budgets $G \in \{2, 4, 8, 16, 32, 64, 128\}$. Each of these 21 programs uses $128K$ nodes and the same independent integration procedure. The weights v_i are computed exactly from cell widths and confidence. We evaluate Eq. (13) by enumerating all 2^9 deletion subsets, and also evaluate its simpler deletion-count lower bound. The incidence certificate in Eq. (15) is checked and included in the CSV as well. Neither of the two plotted certificates exceeds the observed ordered nonzero-sign alternations in any case. Numerical zeros are removed using tolerance 10^{-7}; constraints are checked to 10^{-6}. Full cases, including vacuous certificates, are retained in `results/E7_recovery.csv`. For the alternating weak-cell profile, $G = 8$ produces zero alternations, whereas $G = 16$ produces eight; at $G = 64$ the weighted-order and count certificates are five and four, respectively. This demonstrates both the confidence sensitivity and the possible looseness of the necessary bounds.

G.3 E8: ARCHITECTURAL COST WITH AN EXPLICIT SLOPE CAP

We fix $(a, b) = (3, 1)$ and use $K = 2^n$, $n = 2{:}8$, and $\varepsilon \in \{.125, .0625, .03125, .01, .001\}$, totaling 35 cases. For each, the script evaluates the integer lower bound Eq. (36), independently enumerates nonnegative (N_F, N_G) satisfying the two resource inequalities, and directly constructs n tent modules followed by $m = \lceil \log_2(1/\varepsilon) \rceil$ doubling gains. Gauss integration split at roots and margin crossings verifies $\mathcal{D}_1 = 2^{-m} \leq \varepsilon$. Enumeration is a check of the resource relaxation, not an exhaustive search over network functions; the theorem supplies the universal lower bound and the evaluated construction supplies the feasible upper bound.

The weighted integer certificate equals the constructed cost in all 35 cases. The coarser deletion-count certificate is exact in 28 cases and has gap $2 = a - b$ in the other seven. For $K = 64$, the exact costs are 21, 25, and 28 at $\varepsilon = .125, .01, .001$, respectively; the first coarser bracket is $[19, 21]$. The release also records a controlled removal of the gain cap: replacing m gains by one 2^m-gain module reduces construction cost to $3n + 1$, independent of the deficit target. This is a changed catalog, not a like-for-like gain under the original constraint. All cases and gain magnitudes appear in `results/E8_architecture.csv`.

Scope of this validation. E7 is convex population-surrogate approximation with deterministic numerical integration. E8 is a theorem-backed construction and integer-certificate check. Neither

experiment measures SGD trainability, statistical generalization, hardware energy, or costs of vector-valued layers. The finite-resolution frontier gap and the looseness of the coarse recovery certificate are reported separately from the exact architectural theorem.

H E9–E10: Noisy Readout and Irregular-Region Allocation

E9 tests the noise-location distinction underlying Proposition 1. E10 studies allocation over irregular, unequal-confidence regions using the certificate in Proposition 2 and an exact finite-catalog optimizer.

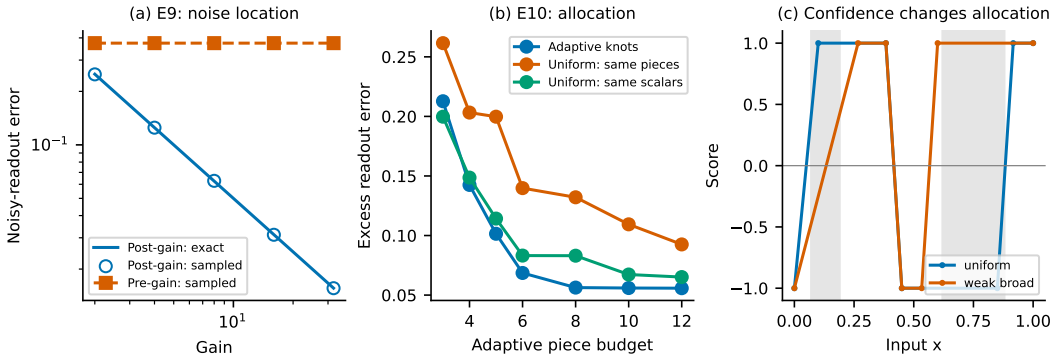


Figure 7: **E9–E10: operational reliability and allocation.** (a) Gain reduces downstream noise error but not upstream noise error. (b) Irregular regions with weak narrow cells, $G = 32$: allocation at matched piece and stored-scalar budgets. (c) At $G = 32, P = 6$, weakening the broad shaded cells changes the optimal allocation relative to uniform confidence. The finite catalog and all comparisons are specified in Section H.

Table 3: **E10: excess readout risk at $G = 32$, $P = 6$.** Adaptive splines store at most $2P = 12$ scalars (ordinates and interior knots). Uniform grids use either $P = 6$ pieces or $2P - 1 = 11$ pieces at the same 12-scalar budget. The certificate applies to all splines with the adaptive (P, G) budgets.

Confidence	Certificate	Adaptive	Uniform, 6	Uniform, 11	Blind/refit
Uniform	.10408	.15317	.22440	.16776	.15317
Weak narrow	.05031	.06857	.13980	.08316	.06857
Weak broad	.04782	.09666	.14288	.08796	.14118

H.1 E9: Noise-Location Control

Let Λ_K be the zigzag with K equal-length branches mapping alternately from zero to one and one to zero. Under uniform input, $\phi = \Lambda_K - 1/2$ is uniform on $[-1/2, 1/2]$. For $y^* = \operatorname{sign} \phi$ and unit-scale downstream noise, $g \geq 2$ gives

$$\mathcal{E}_1(g\phi) = \int_0^{1/g} (1 - gu)\, du = 1/(2g).$$

For noise before gain, the error is $\mathbb{E}[(1 - |\phi|)/2] = 3/8$. E9 uses $K \in \{7, 13\}$, $g \in \{2, 4, 8, 16, 32\}$, and 400,000 independent input/noise pairs (NumPy seed 1729), reused across conditions. Recorded outputs include analytic and sampled downstream error, Monte Carlo standard errors, upstream error, and mean normalized input margin.

H.2 E10: Finite-Catalog Population Allocation

Tasks and budgets. Input is uniform, with endpoints $(t_0, \ldots, t_8) = (0, .07, .19, .23, .41, .57, .62, .88, 1)$. Cell $J_i = (t_i, t_{i+1})$ has sign $s_i = (-1)^{i+1}$ and confidence w_i; $\eta = (1 + w_i s_i)/2$ defines a valid classification model. The uniform profile has all $w_i = 1$;

weak-narrow sets $w_2 = w_5 = .06$; weak-broad sets $w_1 = w_6 = .06$. All other confidences equal one. The oracle knows the target geometry and confidence; they are not estimated from training data.

Scores lie in $[-1, 1]$, with $\tau = 1$, at most $P \in \{3, 4, 5, 6, 8, 10, 12\}$ segments, and slope at most $G \in \{16, 32, 64\}$. The 72 candidate knot locations are the deduplicated union of $\{j/60 : 0 \le j \le 60\}$, $\{j/8 : 0 \le j \le 8\}$, and the target endpoints. The 17 ordinate levels are $\{-1, -7/8, \dots, 1\}$. Endpoints are compulsory; their ordinates are optimized. Including target boundaries makes this an oracle dictionary, not a data-independent approximation result.

Exact objective and dynamic program. For a segment from (a, u) to (b, v), define

$$W_{ab} = \int_a^b w, \quad A_{ab} = \int_a^b w y^* \frac{b-x}{b-a}\, dx, \quad B_{ab} = \int_a^b w y^* \frac{x-a}{b-a}\, dx.$$

Its exact excess risk is $\ell_{ab}(u, v) = (W_{ab} - uA_{ab} - vB_{ab})/2$, and it is feasible when $|v-u| \le G(b-a)$. Integrals are analytic across label cells. For candidate positions x_j and levels h_v, the recurrence is

$$V_{p,j,v} = \min_{i<j,\ u:\ |h_v - h_u| \le G(x_j - x_i)} \{V_{p-1,i,u} + \ell_{x_i x_j}(h_u, h_v)\}, \qquad V_{0,0,v} = 0,$$

with other initial states infinite. The optimum is $\min_{1 \le p \le P, v} V_{p,N-1,v}$. Every feasible spline corresponds to a path through increasing knot indices; its segment risks add. Induction on path length proves optimality in this finite catalog. This standard shortest-path computation is not a new algorithm, nor an optimum over real knots or unquantized ordinates. Independent integration on the union of spline and label breakpoints verifies all objectives.

Baselines and accounting. The piece-matched uniform grid has P intervals. Its ordinates are optimized under the same quantization and slope constraint. An adaptive P-segment spline stores $P + 1$ ordinates and $P - 1$ interior positions: $2P$ scalar slots. The storage-matched uniform grid has $2P - 1$ intervals and $2P$ ordinates, with implicit positions. If an adaptive optimum uses fewer than P segments, $2P$ is its allowed storage budget, not a minimal description length. This accounting is distinct from the fold/gain catalog; it does not measure runtime or bit complexity. The confidence-blind baseline selects knots using the uniform-confidence oracle at the same (P, G), then refits ordinates optimally for the actual confidence. Gains over it isolate location choice instead of withholding confidence from its score fit.

Certificates and verification. Enumerating the 2^8 cell-deletion sets computes both recovery and M_P. For boundary t_j, use $\rho_j = \min(t_j - t_{j-1}, t_{j+1} - t_j)/2$, $c_j = \min(w_{j-1}, w_j)$, and $\beta_j = 0$ in Eq. (20). These symmetric collars are disjoint; the chosen radii can make the bound conservative. For uniform confidence and $G = 32$, the slope floor is .104075, irrespective of P, ruling out readout excess risk .10 by adding knots alone. With weak narrow cells the floor is .050307, and six segments attain .068567. This changes the task's confidence, not just the architecture: boundary count alone does not measure the decision structure that must be reliable.

All 63 adaptive solutions satisfy the slope and recovery certificates. Independent Monte Carlo evaluation averages $w(X)\mathbf{1}\{y^*(X)(f(X) + Z) < 0\}$ on 250,000 new input/noise pairs (seed 271828). It estimates excess readout risk without sampling labels. Every discrepancy from exact integration is within 2.38 estimated standard errors. Conditions share random numbers, so these are correlated checks, not independent significance tests. Nested refinement of the declared representative case (weak-narrow, $G = 32, P = 6$), adding all $j/120$ locations and using 33 levels, reduces risk from .0685667 to .0679417. This single sensitivity check is not a convergence proof.

Results and limits. All configurations, risks, lower bounds, Monte Carlo estimates and selected knots are released. Adaptive knots beat uniform grids in 63/63 piece-matched and 50/63 scalar-matched comparisons. Confidence-aware locations improve 29/42 unequal-confidence cases; the 21 uniform-confidence cases coincide by definition. Other outcomes are retained, including scalar-matched uniform wins. Panel (c) of Figure 7 uses weak-broad confidence to show an observed allocation change; panel (b) uses weak-narrow confidence for budget sensitivity. The feasibility bounds are often loose and do not select an optimum themselves. They reject impossible budgets, while the finite-catalog oracle supplies feasible designs and measures allocation benefits. Wider architectures, finite-sample confidence estimation, and nonuniform or input-dependent receiver noise remain untested.

I E11: Architecture Selection from Resource Certificates

Let $L(P, G)$ denote the lower bound in Eq. (20). We have the following corollary from Proposition 2.

Corollary 8 (Safe selection from resource certificates)**.** *Fix a finite catalog of architecture budgets (P, G), their costs and a tie order, with an exact feasibility oracle for each associated score class. Visit budgets in nondecreasing cost, reject those with $L(P, G) > r$, and query the remaining oracles until finding risk at most r. The selected budget is the same minimum-cost feasible budget as without screening; if none is found, the catalog has no feasible score.*

E11 applies Corollary 8 to the population tasks and spline catalog defined in Section H. It compares screened and unscreened search using the same optimization oracle and computation reuse.

I.1 Decision problem

The three E10 population tasks, their known label-cell geometry and confidence, and the same 72 candidate positions and 17 final-score ordinate levels are retained. A candidate architecture budget is (P, G) with $P = 1, \dots, 16$ and $G = 16 \cdot 2^q$, $q = 0, \dots, 3$. Each budget defines a finite class of continuous splines with at most P pieces, score range $[-1, 1]$, and slope at most G. Its declared cost is $C(P, q) = 2P + 4q$. The first term counts allowed ordinate and interior-knot storage slots; the second assigns a price to increased slope capacity. It is a chosen price schedule, not a claim about parameter counts, energy, or latency. It is also distinct from the fold/gain module catalog of Theorem 4. The problem is to select the lowest-cost budget admitting excess readout risk at most r, for $r \in \{.05, .075, .10, .125, .15\}$. Thus E11 comprises 15 decisions over the same three geometries/confidence profiles, not 15 independent datasets.

I.2 Selection rule and computation accounting

Order all 64 budgets lexicographically by (C, P, G). Compute the resource certificate using cell masses and local boundary collars, without fitting any scores. In screened search, a budget with $L(P, G) > r$ is rejected before querying its feasibility oracle. Otherwise solve its finite spline problem by dynamic programming and stop at the first risk at most r. The baseline has the identical order, oracle and stopping rule, but no certificate. Numerically, rejection uses $L(P, G) > r + 10^{-10}$; feasibility uses risk $\le r + 10^{-10}$. All reported selections are comfortably below their target.

The safe-selection guarantee is proved in Section B; the present experiment measures the computation saved by applying it.

Fair computation accounting. For each slope budget, both methods construct segment costs only when first needed, cache the Bellman table, and extend it from its current piece budget to a larger requested one. Neither repeats previously computed prefixes. Skipping a small P query need not save work: a later larger P query can require those same intermediate states. The work measure counts all candidate predecessor/ordinate entries considered by Bellman minimizations, including infeasible entries represented by infinity. It is a deterministic implementation-level measure, not the number of architecture-budget queries. Certificate evaluation is vectorized over all 2^8 deletion subsets and the four slope budgets. Its time is included in the end-to-end measurement.

Each decision starts with empty caches; no cross-task or cross-search optimizer results are reused. Four methods are evaluated: no screening, piece-only, slope-only, and joint screening. Timing uses three repetitions, with method order randomized using seed 314159 after an initial numerical-kernel warmup. The reported speed ratio divides the sum of unscreened median times by the sum of screened median times. These small CPU timings are descriptive and machine-dependent; deterministic work counts are the primary efficiency measure. An implementation sharing computations differently may have different savings.

I.3 Selected architecture and resource ablations

For uniform confidence at $r = .10$, the slope floor is $.104075$ at $G = 32$, and $.16125$ at $G = 16$. Both rows are rejected for every P, before score optimization. The selected budget is $P = 10, G = 64$, with declared cost 28 and exact risk $.0886667$. Unscreened search makes 42 budget queries and

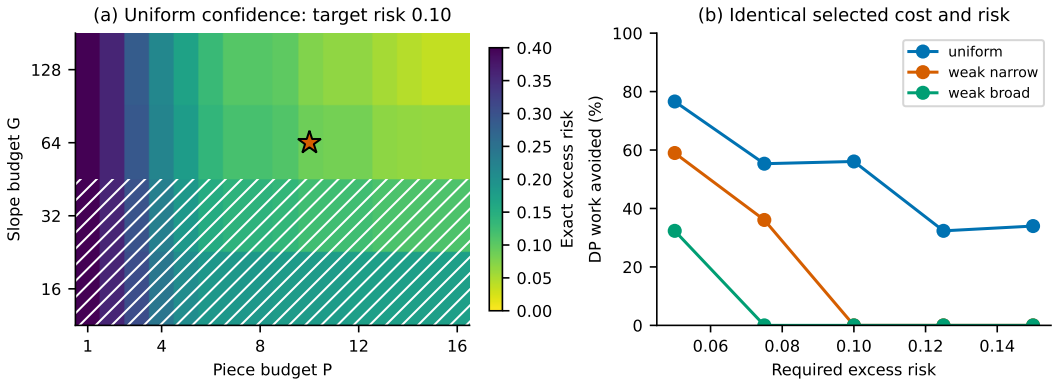


Figure 8: **E11: architecture selection driven by the resource bound.** (a) Uniform confidence, target risk .10: the certificate rejects both hatched slope rows before fitting, for any piece count. The star marks the selected $(10, 64)$ budget; colors are a subsequent exhaustive audit, unavailable to the selector. (b) Work avoided relative to cost-ordered unscreened search with identical DP reuse. Both select the same architecture in every case; zero-savings cases are retained.

considers 26,979,017 Bellman candidate entries. Joint screening makes 16 queries, rejects 26 budgets, and considers 11,849,289 entries: 56.08% less work. The number of slope families instantiated falls from four to two. Median end-to-end time is .10680 seconds unscreened and .04382 seconds screened, including approximately .00021 seconds for the joint certificate. These observations demonstrate a benefit over direct constrained spline optimization: the theory rejects whole resource families before invoking it.

Two one-resource policies further identify the decision's content. Keeping $G = 32$ and increasing P through 16 fails, with best catalog risk .112; the certificate strengthens this to impossibility below .104075 for any number of pieces. Keeping $P = 6$ and increasing G through 128 also fails, with best catalog risk .129. Only the first impossibility is established for unrestricted piece counts; the latter is an observed finite-catalog failure. At the same target risk, weak-narrow and weak-broad confidence each select $(6, 32)$ at cost 16, with risks .06857 and .09666. Both have three score alternations, versus five in the uniform-confidence selected score. These comparisons change the target confidence and illustrate why a fixed raw boundary count does not determine the required resources.

Table 4: **E11 ablation at uniform confidence and $r = .10$.** Every row selects $P = 10, G = 64$, cost 28, risk .08867. Work counts include reused-prefix accounting. Extra query rejections need not save DP work.

Screen	Queries	Rejected	DP entries	Work avoided
None	42	0	26,979,017	0%
Piece only	38	4	26,979,017	0%
Slope only	18	24	11,849,289	56.1%
Joint	16	26	11,849,289	56.1%

I.4 Full sweep and limitations

All 15 decision problems are feasible and both methods choose the same budget and risk as an independently executed exhaustive audit. Across decisions, unscreened DP considers 253,122,095 entries and joint screening considers 156,853,305: a 38.03% reduction. Work savings range from zero to 76.61%. The ratio of summed median runtimes is $1.54\times$, with all certificate costs included. These savings do not improve the final model's risk or declared cost relative to exact unscreened selection; they improve the computation needed to obtain that same choice. Piece-only screening reduces 385 budget queries to 341 but saves no DP work. Slope-only screening uses 231 queries and joint screening 205, with identical total DP work. The computational benefit in this catalog comes entirely from the slope certificate. At weaker confidence and loose risk targets, the bounds can screen nothing. Those zero-benefit cases and the screening overhead remain in the results.

I.5 Independent audit and released records

Only after online searches finish, the script optimizes all $3 \times 64 = 192$ profile/budget combinations to audit every selection and rejected candidate. No rejected candidate is actually feasible. Every score's risk is checked by independent piecewise integration and its slope by its segment differences. Selected scores also receive a fresh Monte Carlo readout evaluation on 500,000 input/noise pairs (seed 161803); it does not influence selection. For the representative choice the estimate is $.088412$ with standard error $.0004015$, agreeing with the exact risk. The release includes every search trace, all audit risks, one-resource policy outcomes, selected score knots and ordinates, timing summaries, and code. The example establishes an operational use of the resource framework in a known population task. Estimating valid certificates from finite data and using them to guide trained vector-valued architectures remain open.

J E12: Finite Diagnostics for Choosing a Repair

E11 assumes known population structure. E12 asks a different question: can a small diagnostic sample select an effective intervention among supplied alternatives? The diagnostic is hand-specified, and the synthetic task families make its signatures interpretable. This is a controlled mechanism study, not a learned discovery policy or a finite-sample-safe screening theorem.

J.1 Tasks and accessible information

Let the raw vector $r \sim \mathrm{Unif}([-1,1]^8)$, choose a hidden coordinate j and orientation $s \in \{-1,1\}$, and put $x = r_j$. The learner receives raw examples, observed labels and a frozen scalar encoding; neither the task family nor j, s is supplied to the selector. We use three mechanisms:

$$\begin{array}{lll} \text{information loss:} & z = x^2, & y = s\,\mathrm{sign}(x), \\ \text{metric compression:} & z = \alpha x, & y = s\,\mathrm{sign}(x), \\ \text{resolution shortage:} & z = x, & y = s\,\mathrm{sign}(\sin(k\pi x)). \end{array} \tag{68}$$

Here $\alpha \in \{.004, .006, .009\}$ and $k \in \{2,3,4\}$. The information family has three independent task draws per seed, not three degrees of information loss. Thirty seed clusters, three mechanisms and three levels give 270 primary instances. The encoder is treated as a black box; exposing its internal coordinate selector would make refinement easier. Its output interval is supplied as fitting metadata, but not to the diagnostic rule.

For each instance, independent streams supply 192 training examples, 64 diagnostic examples, 256 validation examples and 8,192 test examples. Diagnostics consist of 32 pairs $(r, -r)$, reflecting every raw coordinate. Controlled symmetry probes are an explicit privilege of this environment. No test examples or labels enter the decision; no population boundary or confidence map is supplied. Test performance averages $\mathrm{clip}((1-yf)/2, 0, 1)$ over independent inputs, integrating uniform downstream threshold noise analytically. The required risk is .12.

J.2 Interventions and convex fitting

The initial model is a two-piece scalar spline on three equally spaced knots spanning the encoder's range, with slope cap 64. The supplied repairs are: *gain*, raising this cap to 1024 while retaining two pieces; *resolution*, raising the piece count to 32 while retaining cap 64; and *refine*, replacing the head with $b + \langle w, r\rangle$, $\|w\|_1 \le 64$. Refinement learns an affine combination of all raw coordinates; it is not given the informative coordinate. It uses a different head class and bypasses, but does not remove, the resident frozen encoder. This supplied raw-input path is part of the repair menu.

For a design matrix B (linear-spline interpolation or affine raw features), all candidates minimize empirical hinge loss plus $10^{-7}\|\theta\|_1$, using nonnegative hinge slacks u_i and absolute-value auxiliaries:

$$\min_{\theta,u} \ \frac{1}{192}\sum_i u_i + 10^{-7}\|\theta\|_1, \qquad u_i \ge 1 - y_i(B\theta)_i, \quad u_i \ge 0. \tag{69}$$

For a spline, consecutive ordinates obey $|\theta_{l+1} - \theta_l| \le G(t_{l+1} - t_l)$; for refinement the L1 bound is imposed on the eight non-intercept coefficients. The objective's small L1 penalty includes the

intercept; only the hard slope constraint excludes it. SciPy's HiGHS interface solves these linear programs. The finite-sample hinge fit is not a population oracle for threshold-noise risk. No neural optimization or quotient synthesis is performed.

J.3 The diagnostic and comparison policies

The rule first keeps the initial model if its diagnostic risk is at most .12. Otherwise, it chooses refinement if at least 40% of the reflected pairs have equal encodings (absolute tolerance 10^{-10}) and opposite observed labels. Failing that check, it chooses resolution if labels sorted by encoded coordinate exhibit more than two adjacent sign changes; otherwise it chooses gain. These finite diagnostics are heuristic, particularly under label noise; no confidence-certified inference of topology is claimed. The one-shot rule, fixed gain, fixed refinement, fixed resolution and a uniform random repair each permit at most one repair fit and share the same no-op check. They do not select by validation risk. Search policies permit at most three repair fits: cost order (gain, refine, resolution), diagnostic-first followed by the remaining cost order, or exhaustive validation selection. Search stops when a candidate's observed validation risk is at most .12; if none passes, it returns the lowest-validation-risk fitted candidate, including the initial head. Exhaustive selection chooses the cheapest passing head, otherwise the lowest-risk one. Identical training data and candidate fits support a paired audit; each policy sees only the requested candidates. The 900 task instances (including controls below) yield 3,600 fitted heads and 12,960 policy records, including diagnostic-size sensitivity.

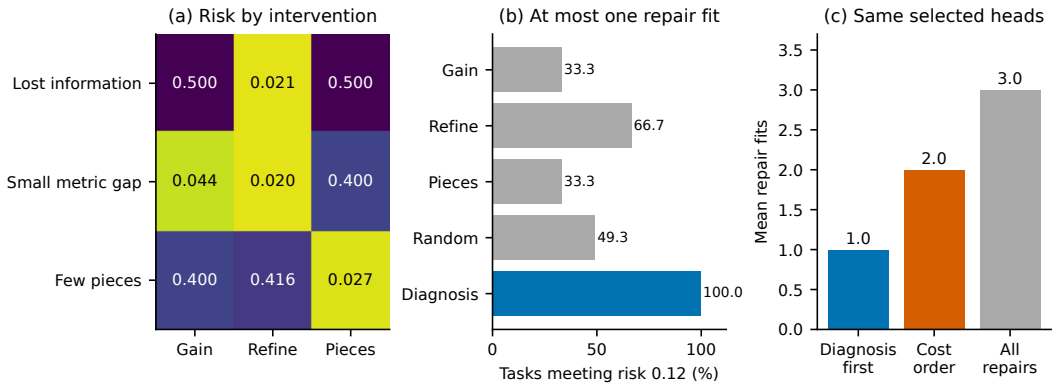


Figure 9: **E12: choosing a repair from finite diagnostics.** (a) Mean test threshold-noise risk over 90 instances of each clean mechanism; refinement also repairs metric compression, at a larger allocated head cost. (b) Success at risk .12 with at most one repair fit, over all 270 primary instances. (c) Diagnosis-first, cost-ordered and exhaustive validation search select identical heads on these instances but request different numbers of repair fits. Counts exclude the common initial fit; this is not an online latency or deployment-cost improvement.

Policy	Risk	Success (%)	Repairs	Slots
fixed gain	0.3144	33.3	1.00	17.00
fixed refine	0.1523	66.7	1.00	19.00
fixed resolution	0.3090	33.3	1.00	43.00
random one	0.2380	49.3	1.00	25.36
diagnostic	0.0306	100.0	1.00	26.33
cost search	0.0306	100.0	2.00	26.33
diagnostic search	0.0306	100.0	1.00	26.33
exhaustive	0.0306	100.0	3.00	26.33

Table 5: **E12 primary results.** Expected test risk averages downstream noise analytically. Success is the fraction of 270 tasks with risk at most .12. Slots include ten common input/encoder slots and the selected head; equal repair-count budgets are not equal FLOP or deployment budgets.

J.4 Results and uncertainty

The diagnostic selects refinement, gain and resolution on all 90 information, metric and resolution instances, respectively. It meets the target in all 270 cases, with mean risk .03056 and a 95% seed-cluster bootstrap interval $[.02983, .03132]$ (4,000 resamples, mechanisms and levels averaged within each of the 30 seed clusters). Cost-ordered search selects the same head in every case. Mean repair fits fall from two to one; including the common initial fit, total fits fall from three to two. Instantiated LP nonzeros fall from 2,780 to 1,728 per task, a 37.84% reduction. Selected storage remains 26.33 slots on average. Observed 100% success is a finite-study outcome, not a guarantee of zero failure probability. Full intervals and paired differences for risk, success, fits, storage and acquisition are released.

J.5 Cost and information accounting

Allocated head slots are 3 initially, 7 for gain, 9 for refinement, and 33 for resolution. Gain charges four bounded-gain doublings; its numerical fit uses the equivalent effective slope cap. Refinement charges all nine allocated coefficients, including fitted zeros. Eight raw-input slots and two resident encoder slots are common costs. These declared counts do not measure hardware cost or unrestricted sparse-model optimality. Raw inputs remain available to every method; no free routing oracle or ground-truth coordinate is supplied. All one-shot policies request 256 labels (192 training plus 64 diagnosis). Both search policies request the same 512 labels on the primary tasks, but search that omits diagnosis requests 448. Thus diagnosis buys one avoided repair fit with 64 additional labels relative to that baseline; whether this is worthwhile depends on acquisition and fitting costs. We report those resources separately instead of choosing a favorable weighted price. LP variable/nonzero counts are work proxies, not FLOPs. Candidate fitting and diagnostic times are recorded, while validation score evaluations are counted separately; reconstructed fit-plus-diagnostic times are not measured online latencies. No end-to-end speedup is claimed.

J.6 Controls that limit the interpretation

We add 90 healthy instances ($z = x$, sign label), 90 compatible folds ($z = x^2$, label $s\,\mathrm{sign}(x^2 - c)$ for $c \in \{.2, .4, .6\}$), and 90 instances of each mixed failure: information loss plus oscillations, and metric compression plus oscillations. Healthy and compatible-fold tasks all pass without repair; the rule does not undo every noninjective encoder. Both mixed groups have zero success under every tested policy at the target: the supplied single-repair menu is inadequate. Search uses all three repairs on these failures; the runs are retained. We also repeat the 270 primary instances with independent 10% label flips in training, diagnosis and validation, retaining the clean test targets and downstream noise. All 90 metric instances are then diagnosed as needing resolution because label flips create spurious alternations. One-shot success falls to 64.81% and mean test risk rises to $.1840$. Diagnosis-first validation search reaches 97.41% success and risk .0791, matching cost-ordered search but requesting 2.893 versus 2.993 repair fits. The validation labels are corrupted in this stress test, so its stopping risk is not an estimate of the clean target risk without further correction. With 16 diagnostic labels, clean one-shot success is 99.26%; 64 and 256 give 100%. These are sensitivity measurements, not sample-complexity bounds. The fixed protocol, all fitted coefficients, per-task choices, raw metrics, five independent mathematical/data-access checks and notebook are released. All LPs passed their slope/L1 checks; maximum primal residual was below 10^{-12}.

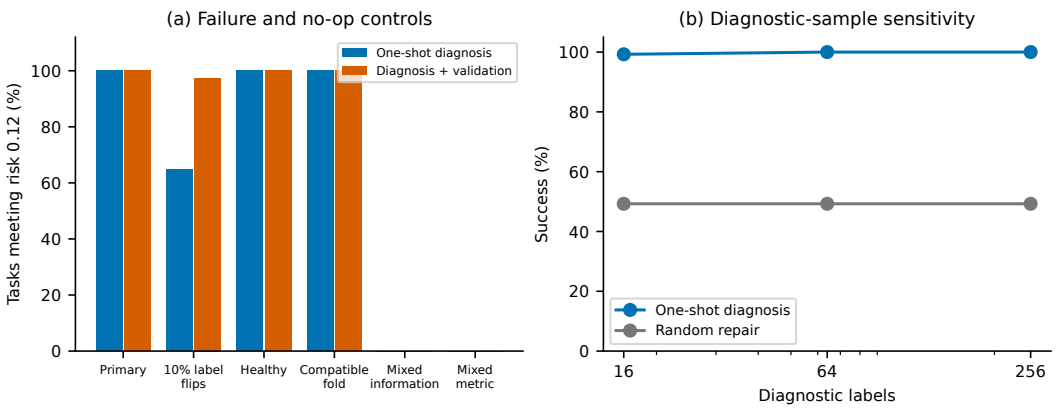


Figure 10: **E12: negative controls and diagnostic size.** (a) Label corruption reduces one-shot success; validation recovers most successes at greater fitting effort. Healthy and compatible-fold tasks need no repair, whereas mixed failures defeat the supplied single-repair menu. (b) Primary-task success versus the number of controlled diagnostic labels. The clean outcome relies on interpretable synthetic mechanisms, symmetry probes and supplied interventions.

K Additional Margin and Checkerboard Diagnostics

These auxiliary studies complement E1–E12. The direct margin-profile check evaluates a known regression function; the trained checkerboard probes and depth–width sweep assess learned scores under their own protocols.

K.1 Monte Carlo check of the invariant margin profile

We draw 300,000 points from $\mathrm{Unif}[0,1]$ with random seed 2026, set $\delta = .15$ and $\tau = .35$, and evaluate the known regression function Eq. (17) for $L = 1, \ldots, 8$. At $s \in \{.035, .07, .14, .21, .28\}$ we estimate the empirical margin mass and fit a zero-intercept least-squares line. The exact theoretical slope is $2\delta/\tau = .857142857$. Across all eight depths the fitted slopes range from .8543 to .8616 and the largest absolute deviation of any reported mass from theory is 1.32×10^{-3}.

L	boundary components 2^L	fitted slope	max. abs. error
1	2	.8554	6.9×10^{-4}
2	4	.8587	8.1×10^{-4}
4	16	.8616	1.3×10^{-3}
6	64	.8543	8.3×10^{-4}
8	256	.8553	5.3×10^{-4}

Table 6: Monte Carlo verification of the invariant margin profile. The exact slope is .8571 for every L, even though boundary topology grows by $128\times$.

Table 6 compares the sampled profile with Corollary 5 as the known boundary count grows from 2 to 256. The agreement checks implementation of the analytic regression function; it does not establish invariance of a learned representation.

K.2 Layerwise gradient and collar diagnostics

For each seed in $\{3, 11, 23\}$ we draw 6,500 i.i.d. uniform points from $[0,1]^2$ for the 4×4 checkerboard, stratify a $70/30$ training/evaluation split, and train a fully connected ReLU MLP with five hidden layers of width 64. We use binary cross entropy, Adam with learning rate 3×10^{-3}, and 1,000

full-batch steps. The checkpoint with highest held-out accuracy among evaluations every 50 steps is used for this diagnostic experiment.

At the input and each hidden layer, coordinates are standardized on the training split and a linear SVM with $C = 3$ is fitted. For probe score $s(z) = w^\top z + b$, we normalize $q = s/\|w\|_2$. Pullback boundary length is the number of horizontal/vertical sign-changing adjacencies on a 121×121 input grid divided by 120. Input gradients $\nabla_x q_\ell(x)$ are computed by automatic differentiation through the prefix of the trained network. We report the median gradient norm among evaluation points with $|q_\ell| \le .1$ (using the closest 10% if fewer than 20 points fall in the collar) and finite-band density $P_n(|q_\ell| \le .1)/.2$.

Table 7 reports the available mean and standard deviation across three seeds beside each gradient and density statistic. Standard deviations for accuracy and boundary length are unavailable in the retained record. Large relative variability limits the strength of any layerwise trend.

repr.	probe acc.	boundary length	boundary grad.	finite-band density
input	.485	.93	$3.46 \pm .02$	$.221 \pm .192$
hidden 1	.794	7.40	$.69 \pm .19$	$4.548 \pm .088$
hidden 2	.958	6.61	$1.79 \pm .56$	$2.590 \pm .715$
hidden 3	.982	6.39	7.36 ± 6.81	$1.058 \pm .827$
hidden 4	.989	6.43	34.88 ± 17.48	$.162 \pm .086$
hidden 5	.989	6.43	68.11 ± 18.09	$.074 \pm .022$

Table 7: Layerwise probes, means over three seeds. Early layers manufacture the multi-cell pullback frontier; after its length stabilizes near the target, later probes have larger near-boundary gradients and lower finite-band density under their respective normalizations. Gradient and density columns include standard deviations; the other columns are means only.

The exact checkerboard frontier has grid-length proxy 6.05. Table 7 associates a roughly stable boundary-length proxy with larger gradients and lower finite-band density in later layers. Because each probe is fitted and standardized separately, this is a qualitative diagnostic, not conservation of one calibrated score across layers. The length proxy is orientation-dependent, and a median gradient does not yield the length-weighted reciprocal-gradient integral in Proposition 4. Selected end-to-end evaluation accuracy is $.9896 \pm .0003$.

K.3 Checkerboard depth–width sweep

For depths $D \in \{1, 2, 3, 5\}$ and widths $m \in \{16, 32, 64, 128\}$ we train on $5{,}000$ independently sampled checkerboard points, with a $70/30$ stratified split, Adam learning rate $.003$, and 500 full-batch steps. Seeds are $\{5, 17\}$ and the best held-out accuracy among checkpoints every 50 steps is reported. This sweep is intentionally presented as an optimization experiment: the one-dimensional theorem is a representability lower bound and does not imply that gradient descent finds every structurally feasible oscillatory solution.

	$m = 16$	$m = 32$	$m = 64$	$m = 128$
$D = 1$	.579	.620	.588	.689
$D = 2$	.689	.772	.893	.961
$D = 3$	.793	.888	.963	.981
$D = 5$	.799	.860	.977	.986

Table 8: Depth-width sweep on the 4×4 checkerboard (mean over two seeds). Depth substantially reduces the width required for high accuracy on this fixed multi-cell task.

In Table 8, one-hidden-layer networks remain below $.70$ even at width 128, whereas depth-2 and depth-3 networks reach $.961$ and $.981$, respectively, at the same width. Depth also reduces the width required to cross $.95$ accuracy, from 128 at depth 2 to 64 at depths 3 and 5. This sweep should not be read as a direct numerical verification of the exact one-dimensional lower bound. Its outcomes

combine representational capacity with the ability of Adam to discover a suitable folding, and the nonmonotonic entries at small widths reflect that optimization variability. Within this exploratory sweep, additional depth makes a multi-cell boundary attainable at substantially smaller width, while increasing width alone is comparatively inefficient. This sweep and E2 use different architecture ladders and training budgets. Their results should be compared within each protocol, not combined into an unrestricted claim that additional depth helps or fails to help.

L Relation to Prior Work

Existence versus resource-bounded separation. Urysohn's lemma separates disjoint closed sets in normal spaces by a continuous scalar function (Munkres, 2000). It supplies a score gap on the specified sets, but no bound on realization cost or input-space sensitivity. Our motivating inequality Eq. (1) is an elementary Lipschitz consequence, not a refinement of the lemma's topological hypotheses. The additional questions are distribution-sensitive: which alternating regions must be represented at a given loss budget, and what combination of piece capacity and slope can support them? The distinction matters because the closures of adjoining Bayes regions need not be disjoint. The main obstruction uses local density and confidence instead of assuming globally separated class supports.

Topology and folding. Ergen and Grillo (2024) prove exponential advantages of depth in Betti complexity, while Naitzat et al. (2020) empirically track topological simplification across trained layers. Keup and Helias (2022) interpret feed-forward networks as progressively folding data into linearly separable configurations. We ask a complementary quantitative question: what probability and Jacobian cost accompanies a topologically complicated pullback boundary? The conservation identity in Theorem 1 describes an exact mechanism, while Theorems 2 and 3 constrain imperfect predictors without assuming that every target boundary survives. The MTF interpretation in Section 5.2 does not assert preservation of input Betti numbers: a task-compatible noninjective representation may legitimately merge same-label components.

Readout capacity versus pullback complexity. For a fixed representation, Cover's theorem prices affine-readout dichotomies through the affine dimension of the represented sample (Cover, 1965). It does not price the encoder class that produced that representation. Here the represented score is scalar and the readout is always one threshold, while the reachable pullback boundary has exponentially many components. The relevant capacity is architectural reachability of pullback frontiers, not the dimension of the final readout.

Depth separation. Region-count bounds and the gap between maximal and typical complexity are developed by Montúfar et al. (2014); Serra et al. (2018); Hanin and Rolnick (2019). The tent-map mechanism is closely related to Telgarsky (2016), and Chatziafratis et al. (2020) connect periodic points to depth–width tradeoffs for ReLU networks. Safran et al. (2019) show why the Lipschitz scale can change the interpretation of oscillatory depth separations. We do not claim a new oscillatory hard function. We place the mechanism inside a calibrated Bayes model and identify the conserved probability-weighted inverse-Jacobian quantity, its distribution-mismatch correction, a margin-deficit obstruction that does not assume all roots are recovered, and a joint topology–slope payment bound for compositional modules.

MLPs versus KANs. Spline operator views and learned spline activations predate KANs (Balestriero and Baraniuk, 2018; Bohra et al., 2020). KANs replace scalar weights by trainable univariate edge functions, commonly represented by splines, and were proposed as an alternative to node-activated MLPs (Liu et al., 2025). Subsequent controlled comparisons emphasize that rankings depend on matching parameters, computation, and the task's functional structure (Yu et al., 2024). Our result does not supply another aggregate architecture ranking. It isolates three currencies that such comparisons should report separately: the number of affine boundary pieces reachable by composition or knots, the effective Jacobian needed to support their margin profile, and the optimization cost of finding the compact realization. Corollary 3 makes the first two exact on the folded family: deep MLPs and deep KANs both achieve logarithmic description cost, a shallow spline pays linearly in pieces, and all of them obey the same Jacobian lower bound.

Basis functions and implementation cost. FastKAN replaces cubic B-splines with Gaussian radial basis functions (Li, 2024); a nonconstant Gaussian basis is not CPWL, so the finite affine-piece bounds and the exact two-piece-module cost do not transfer directly. A derivative-based obstruction still applies to a Lipschitz composite on the domain where its layer bounds hold, with constants computed for the chosen bases and normalization. EfficientKAN instead reorganizes B-spline evaluation into shared basis expansion and matrix multiplication (Blealtan, 2024). Such an implementation changes memory and arithmetic costs; it does not turn higher-order splines into order-1 CPWL functions. Applying our piece bounds requires a CPWL variant, including its base activation. Applying the exact cost theorem additionally requires the declared scalar module catalog and slope cap. A hardware-aware comparison would price basis evaluation, routing, memory traffic, and dense arithmetic explicitly. Parameter parity alone cannot establish a speed or memory advantage.

Transfer operators, margins, and norms. Transfer operators and invariant densities are classical tools for expanding maps (Rényi, 1957; Lasota and Mackey, 1994); margin conditions and coarea are also classical (Mammen and Tsybakov, 1999; Audibert and Tsybakov, 2007; Evans and Gariepy, 2015). Our contribution is to couple these tools: the transfer operator gives exact global conservation, the paired-boundary argument yields a loss-sensitive obstruction, and the module-capability dual gives architecture-independent accounting. The resulting lower bounds complement spline Lipschitz control (Aziznejad et al., 2020) and spectral-margin generalization analysis (Bartlett et al., 2017). It also complements norm-controlled depth separation, where representation cost and sample complexity replace width as the principal resource (Parkinson et al., 2024). Random-feature separation instead follows a stochastic, width-driven route (Dirksen et al., 2022).

Geometry of trained networks. Recent work characterizes polyhedral complexes formed by ReLU networks and validates their geometry empirically (Gaines and Bi, 2026). Our probes target a different object: the pullback frontier of a linear readout, together with its gradient and collar density.

M Discussion and Limitations

The phrase *no free topology* denotes conditional resource accounting, not a universal topological conservation principle. Urysohn separation, metric sensitivity, probability of a transition region, and architecture cost are distinct objects. A task can tolerate deletion of low-confidence regions; a representation can merge components with identical task demands; and a catalog with unbounded gain removes the growing noisy-readout cost. The lower bounds acquire content only after fixing the task, reliability interface, coordinate units, admissible modules and their prices.

The conservation theorem requires a full-branch expanding map and an invariant input density. These assumptions make the mechanism exact, not universal. Under distribution mismatch, Eq. (3) replaces invariance by the pushed-forward density $\mathcal{L}_T^L q$, which may converge, oscillate, or concentrate near the readout threshold. Nor does a fixed Bayes-margin profile imply fixed sample complexity. Smoothness, network norm, optimization, and hypothesis-class complexity may all change; our geometric margin and norm results show explicitly that two of them do.

The architectural cost theorem is restricted to scalar modules with an explicit slope cap. The boundary-count lower bound is one-dimensional. The tensorized corollary creates exponentially many complement cells, but its lower bound is inherited from a one-dimensional restriction. A natural next problem is to derive width or size lower bounds directly from higher-dimensional Betti vectors or the adjacency complex of the pullback frontier. Coarea relates collar mass to a density-weighted reciprocal-gradient surface integral; boundary multiplicity alone does not force exact cancellation. No result here asserts such cancellation for arbitrary non-oscillatory manifolds.

Analytic tasks allow direct checks of population quantities; the added MNIST and Fashion-MNIST experiments instead measure learned scores and finite-sample geometric proxies (Section N). R1 retains synthetic mixtures with known densities, R2 uses prescribed one-dimensional paths, and R3 does not identify an image-data Bayes frontier. R4 engineers failures after learning an image score and has additional protocol limitations. These studies broaden the empirical setting without validating an ambient-dimensional conservation law or a deployment-noise model.

The MLP-KAN interpretation is likewise a mechanism-level statement, not a universal performance ranking. It is exact for scalar order-1 splines and CPWL ReLU networks on the folded family. Higher-

order splines, multivariate feature learning, numerical conditioning, and optimizer bias can change the parameter and training costs, although they cannot invalidate the architecture-neutral Lipschitz converse for the realized score. E6 supplies a controlled scalar stress test by matching factorizations of the same score and varying only the fold–knot allocation. Its cold–warm gap describes two protocols with different learning rates. E6-C crosses both rates with both initializations and adds frozen folds, but its two depths and five seeds still do not establish how initialization, curricula, or hybrid routing behave on natural tasks. A broader test should match induced piece count and effective Jacobian product, and then compare optimization time and generalization.

Finally, score-collar mass alone is neither an irreducible clean excess-risk floor nor an input-space robustness radius. The noisy-readout results add a specific independent downstream threshold channel. Its fixed scale gives margin an operational meaning; upstream or gain-proportional noise need not benefit from amplification. E10 extends resource accounting to irregular regions, but its oracle knows the population labels and confidence, its optimality is confined to a finite catalog, and the certificates can be loose. E11 shows a use for the population certificate in architecture selection; its savings come from the slope bound, and its cost is an illustrative price on scalar storage and slope capacity. The remaining empirical question is whether estimated decision structure can guide selection in learned architectures under a measured deployment noise model.

N R1–R4: Planar and Image-Data Diagnostics

The additional experiments ask which parts of the resource account remain informative outside the analytic folded family. R1 examines a coarea-based collar approximation on noisy planar mixtures; R2 restricts trained scores to prescribed paths; R3 compares depth, gain, and a separation proxy on image classification tasks; R4 tests a supplied repair rule on engineered failures of learned MNIST scores. Only the image experiments address the absence of nonsynthetic data. None assumes that a learned image encoder is an expanding map with an invariant density. The source is `real_manifolds.ipynb`, with its saved full-run outputs (`SMALL=False`). Tables reproduce the printed values; the notebook was inspected but its training was not rerun during integration. The release preserves the notebook unchanged, with extracted CSV files and a script regenerating Figure 11. Per-seed R3 values and per-path R2 records were not serialized, so their dispersions cannot be recovered.

N.1 R1: Collar geometry on noisy planar mixtures

Each class is an equal mixture of 400 isotropic Gaussians centered on an arc or circle, with equal class priors. The moons have standard deviation $.12$; the concentric circles have radii 1 and $.55$ and standard deviation $.10$. Their analytic densities give both $p(x)$ and $\eta(x)$. Models use $6{,}000$ training and $4{,}000$ evaluation samples, binary cross-entropy, and $4{,}000$ Adam updates with batch size 256 and cosine-decayed initial learning rate $.002$. Hidden-depth/width pairs $(1, 64), (2, 24), (3, 18), (4, 15)$ have $257, 697, 757, 781$ parameters, respectively; they are not parameter matched. There is one training run per architecture and dataset. For a regular score level, coarea suggests the small-collar approximation

$$\Pr_p(|f(X)| \le \epsilon,\ X \in B) \approx 2\epsilon C_f(0; B), \qquad C_f(0; B) = \int_{\{f=0\}\cap B} \frac{p(x)}{\|\nabla f(x)\|}\, d\mathcal{H}^1(x). \tag{70}$$

The exact expression integrates the level-dependent coefficient over $[-\epsilon, \epsilon]$; replacing it by its value at zero requires suitable local regularity. The notebook estimates both sides on a 401×401 grid in $B = [-1.8, 2.8] \times [-1.6, 1.8]$, using marching squares for zero segments, central differences for gradients, and midpoint quadrature along segments. For $\epsilon = .02, .05, .10$, the 24 mass-to-approximation ratios range from $.835$ to 1.160. These are finite-grid consistency measurements, not exact equalities or convergence estimates.

Table 9 illustrates how probability-weighted inverse gradients enter collar measurements beyond the dyadic example. It does not establish conservation across trained depths: the coefficient itself changes. The notebook reports one extracted boundary component intersecting its density threshold at every depth. Components are connected before filtering for intersection with $p > 10^{-3} \max_B p$, so the count is not the number of components of the density-restricted zero set. Grid resolution, box

Task	Hidden layers	Accuracy	Mean boundary gradient	$C_f(0; B)$	Ratio range
Moons	1	.998	19.5	.0035	.964–1.160
Moons	2	.998	34.7	.0019	.868–.981
Moons	3	.999	40.0	.0016	1.002–1.054
Moons	4	.998	51.5	.0017	.952–1.006
Circles	1	.988	32.5	.0102	.898–1.056
Circles	2	.987	45.1	.0075	.988–1.017
Circles	3	.987	46.0	.0074	.962–1.067
Circles	4	.987	48.0	.0072	.835–.984

Table 9: R1: learned-boundary measurements on synthetic mixtures. Ratio ranges cover three collar widths. The gradient column is an unweighted mean over extracted segment midpoints, not a global Lipschitz constant.

truncation, and ambiguous marching-squares cells limit topological interpretation; the count does not certify the Bayes topology.

N.2 R2: One-dimensional diagnostics along sampled paths

A path joins alternating-label anchors by straight segments and is parameterized by normalized arclength $t \in [0, 1]$. The restricted score $f \circ \gamma$ is sampled at $4{,}001$ equally spaced values of t. For planar mixtures, the task uses analytic Bayes labels and confidence $|2\eta - 1|$ along the path. Each dataset supplies 200 paths for a strong model (two width-24 hidden layers, $4{,}000$ updates) and a weak model (one width-6 hidden layer, 666 updates), with three to six anchors. The slope calculation uses the $\beta = 1$ form of the deficit obstruction. For MNIST 3-versus-5, strong and weak networks have architectures $[784, 64, 32, 1]$ and $[784, 8, 1]$, trained for $2{,}500$ and 208 updates, respectively, on up to $6{,}000$ official training images. Each is measured on 150 paths through four to eight anchors from the official test split. Between anchors, labels are assigned by the nearest anchor in t, with unit confidence and threshold scale $\tau = 1$. This defines an artificial one-dimensional task with uniform path measure; it is not the MNIST Bayes labeling of interpolated images. Slope is measured with respect to normalized path position, not ambient Euclidean distance.

Task	Model	Paths	Active checks	Median bound/$\widehat{G}$	Violations
Moons	Strong	200	173	.18	0 / 0
Moons	Weak	200	0	—	0 / 0
Circles	Strong	200	196	.39	0 / 0
Circles	Weak	200	1	.53	0 / 0
MNIST 3 vs. 5	Strong	150	21	.12	0 / 0
MNIST 3 vs. 5	Weak	150	15	.64	0 / 0

Table 10: R2: sampled slope and recovery diagnostics. Active checks have a positive computed slope lower bound; median ratios use only active checks. The final column lists numerical slope/recovery violations.

Across $1{,}100$ paths, the notebook records no numerical slope or recovery violations; 406 slope checks are active (Table 10). The connection to Theorems 2 and 3 is through the restricted scalar problem. These measurements do not certify every hypothesis: finite differences can miss slope peaks and crossings, and the planar confidence coefficient is estimated at selected locations. Zero observed violations are a diagnostic consistency result, not $1{,}100$ certified applications of the bounds.

N.3 R3: Depth, gain, and separation on image tasks

The tasks are MNIST 3-versus-5 and 4-versus-9 (LeCun et al., 1998), and Fashion-MNIST class 0 (T-shirt/top) versus class 6 (shirt) (Xiao et al., 2017). Images are flattened to 784 coordinates and scaled to $[0, 1]$. A fixed seeded permutation selects up to $8{,}000$ examples from the official training split and up to $2{,}000$ from the official test split after class filtering. The depth ladder is $[784, 39, 1]$, $[784, 36, 36, 1]$, $[784, 33, 33, 33, 1]$, and $[784, 31, 31, 31, 31, 1]$. Each architecture is trained with three seeds, binary cross-entropy, $2{,}500$ Adam updates, batch size 256, and a cosine schedule starting

at .001. The final iterate is evaluated; there is no test-based checkpoint selection in R3. For each trained score, the notebook computes the spectral product $U = \prod_\ell \|W_\ell\|_2$, an upper bound on its global Lipschitz constant. It also samples 200 straight segments joining opposite-class test images, with 801 points per segment. The largest absolute adjacent score increment divided by Euclidean distance defines $\widehat{G}$. In exact arithmetic these sampled secants satisfy $\widehat{G} \le \mathrm{Lip}(f) \le U$. The reported separation proxy is

$$M_{\mathrm{probe}}(f) = \frac{\mathrm{median}_{x \in \mathrm{test}} |f(x)|}{\widehat{G}}. \tag{71}$$

Because its denominator can underestimate the true Lipschitz constant, M_{probe} is not a certified input margin. It includes both correctly and incorrectly classified examples, and interpolating segments need not remain on the data distribution. Each table entry is a mean over three training seeds; the last column averages the three per-seed ratios.

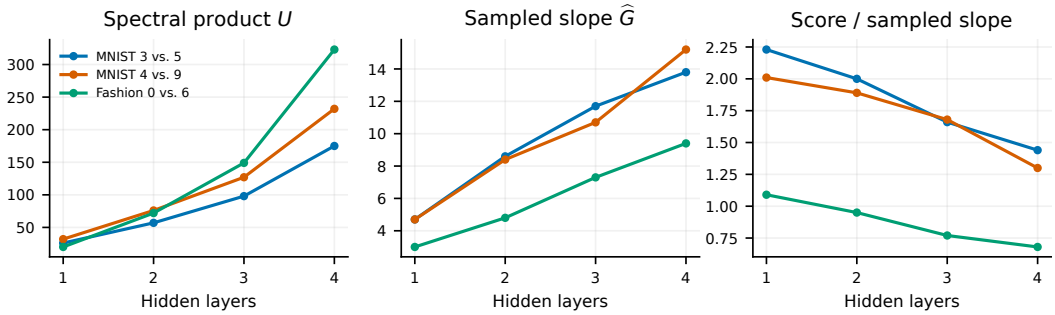


Figure 11: **R3: greater network gain need not accompany greater measured separation.** Spectral products and sampled directional slopes increase over the depth ladder, while the score-to-sampled-slope proxy decreases on all three image tasks. Values are means over three training seeds, reconstructed from rounded console outputs. Seed dispersions were not saved; no error bars or significance claims are inferred.

Task	Depth	Parameters	Accuracy	U	$\widehat{G}$	M_{probe}
MNIST 3 vs. 5	1	30,655	0.988	26	4.7	2.23
MNIST 3 vs. 5	2	29,629	0.992	57	8.6	2.00
MNIST 3 vs. 5	3	28,183	0.990	98	11.7	1.66
MNIST 3 vs. 5	4	27,343	0.989	175	13.8	1.44
MNIST 4 vs. 9	1	30,655	0.984	32	4.7	2.01
MNIST 4 vs. 9	2	29,629	0.984	76	8.4	1.89
MNIST 4 vs. 9	3	28,183	0.986	127	10.7	1.68
MNIST 4 vs. 9	4	27,343	0.984	232	15.2	1.30
Fashion 0 vs. 6	1	30,655	0.860	20	3.0	1.09
Fashion 0 vs. 6	2	29,629	0.857	72	4.8	0.95
Fashion 0 vs. 6	3	28,183	0.854	149	7.3	0.77
Fashion 0 vs. 6	4	27,343	0.853	323	9.4	0.68

Table 11: R3: complete saved depth-ladder means. Approximately comparable model sizes are reported explicitly; neither parameter count nor accuracy is held exactly fixed. Fashion denotes T-shirt/top versus shirt.

From depth one to four, spectral products increase by approximately 6.7, 7.3, and 16.2 times, while M_{probe} decreases by approximately 35–38%. Accuracy changes from .988 to .989, .984 to .984, and .860 to .853, respectively. The results support reporting score scale together with sensitivity: larger network gain alone does not establish better input-space separation. They complement the paper's distinction between readout amplification and geometric margin, but different trained networks are not a controlled rescaling of one score. R3 does not test pure-gain invariance causally or establish a general depth law. An auxiliary layerwise probe uses 600 test images and classwise 8-nearest neighbor graphs on row-normalized hidden representations. The sum of the two class graph-component counts is two at every layer. Affine probes fitted and evaluated on those same representations have increasing

median $|f|/\|w\|_2$, from $.682$ at the input to 5.265 at the last hidden layer. These are in-sample observations. Connectivity at one graph scale does not determine the topology of the class supports, so the result cannot establish that all encoder work is metric.

N.4 R4: Exploratory repairs of learned MNIST scores

R4 learns one $[784, 16, 1]$ encoder for each of MNIST 3-versus-5, 4-versus-9, and 0-versus-8, using up to $6{,}000$ training and $1{,}500$ test images and $1{,}500$ Adam updates. It engineers three scalar tasks: folding the score about its median, compressing its centered value by 10^{-3}, and assigning alternating labels to its four quantile cells. The last task replaces digit labels with synthetic labels; the first two retain them. The study concerns controlled interventions on image-derived representations, not naturally occurring failure categories. A fixed rule uses 64 labeled training examples. It chooses information restoration when opposite labels frequently occur at nearby codes; otherwise, it chooses finer resolution when smoothed labels along sorted codes alternate more than twice, and greater gain otherwise. Candidates are a three-knot spline with slope cap 64, the same head with cap 10^5, a 33-knot spline with cap 64, and an affine classifier on the original 784 pixels. Spline heads use $1{,}500$ Adam updates at learning rate $.05$, with adjacent ordinate differences clipped to enforce the slope cap; the raw-pixel head uses $1{,}200$ updates at $.005$. The latter restores access to information outside the scalar code and has a different parameter cost. The metric is deterministic classification error, not the threshold-noise excess risk used in E12.

Digits	Engineered task	Base	Gain	Resolution	Raw pixels
3 vs. 5	Folded score	0.305	0.305	0.311	**0.035**
3 vs. 5	Compressed score	0.475	**0.027**	0.475	0.035
3 vs. 5	Alternating relabeling	0.387	0.387	**0.033**	0.331
4 vs. 9	Folded score	0.238	0.238	0.245	**0.030**
4 vs. 9	Compressed score	0.521	**0.027**	0.521	0.030
4 vs. 9	Alternating relabeling	0.323	0.323	**0.027**	0.323
0 vs. 8	Folded score	0.369	0.369	0.373	**0.009**
0 vs. 8	Compressed score	0.499	**0.007**	0.499	0.009
0 vs. 8	Alternating relabeling	0.345	0.345	**0.055**	0.255

Table 12: R4: errors on nine fixed image-derived tasks. Bold entries are the heads chosen for all three sampled diagnostic subsets. The subset seeds do not change the encoder or fitted-head errors; they are not independent training replicates.

All 27 decisions match the supplied mechanism-to-repair rule, and selected errors range from $.007$ to $.055$, below the $.15$ target (Table 12). This is 9 fixed tasks evaluated with 3 diagnostic subsets each. Alternative heads can also succeed: the raw-pixel classifier meets the target for compressed scores as well. All candidate heads were fitted, so R4 provides no measured search-time saving or minimum-cost selection guarantee. Two protocol details limit stronger interpretation. First, centering, medians, and quantiles are recomputed separately on training and test scores. Evaluation does not use a single transformation fitted on training data and frozen at deployment. In particular, the quantile task's test labels depend on the test-score distribution. Second, diagnostic subsets come from encoder-training examples. A confirmatory study should freeze transformations using training data, reserve diagnostic examples, and repeat independently trained encoders before making held-out repair claims. The corruption cell prints $.70$ selection accuracy with a 20% label-flip probability. It also changes encoder training from $1{,}500$ to $1{,}250$ updates and changes diagnostic-subset seeds, so it is not a matched label-noise ablation. Its lower selection rate cannot be attributed to label corruption alone. The optional CIFAR-10 cell was skipped after its download failed; no CIFAR-10 result is included.

What the extension establishes. R3 adds ordinary held-out image classification measurements to the empirical study. R1 and R2 examine geometric quantities when population densities are available or paths are prescribed. R4 suggests a way to study repairs of learned image scores, while leaving a controlled, cost-matched validation to future work. Together, they broaden the evidence for the resource perspective without claiming exact conservation, certified robustness, or autonomous structural discovery on natural image distributions.