SMOOTHING DOES NOT COMPRESS HARD STRUCTURED INFERENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Differentiable structured-inference layers replace hard $\min$ recurrences by regularized operators. We ask whether this smoothing can reduce the representation cost of the underlying hard value function. For a broad class of finite DAGs, the answer is no in the sublinear-ray regime. We pair a finite-temperature computation with a same-wiring tropical shadow: bounded min/max deformations are replaced by hard extrema, positive scalar gates are retained, and finite constants are removed. A recursive DAG budget $E(C)$ then bounds the discrepancy on every score ray. Consequently, any $o(t)$-accurate realization has the target's exact shadow and inherits its min-plus or STC lower bound without a size loss. The same budget gives finite-scale open-cone slope and relative-support floor certificates, while annealing changes only the deformation term. For q-ary linear-code inference, we prove $q^\rho \le S_{\min} \le O(nq + nq^{2\rho})$; for the $k \times k$ grid cycle code, $\rho = k - 1 = \Theta(\sqrt{n})$, giving a $2^{\Omega(\sqrt{n})}$ barrier. CUDA double-precision experiments recover the exact width threshold and code-distance slopes, and separate deformation from representation error under annealing. Signed affine preprocessing is covered by a signed shadow with an explicit affine budget.

1 INTRODUCTION

Structured-inference layers map local scores to the best globally feasible structure. For a feasible relation $\mathcal{R} \subseteq \Sigma^X$, the hard value function is

$$\Phi_{\mathcal{R}}(u) = \min_{r \in \mathcal{R}} \sum_{i \in X} u_{i,r_i}. \tag{1}$$

Viterbi decoding, constrained sequence prediction, and many dynamic programs are instances of (1). Differentiable dynamic programming replaces the hard extremum by a regularized operator (Mensch & Blondel, 2018). We study the representation cost of this replacement as a function of the unary scores, allowing arbitrary DAG sharing.

The central construction is a same-wiring tropical shadow. Replace every gate that is within bounded additive distance of $\min$ by $\min$, preserve positive scalar gates, and set finite constants to zero. A DAG recursion gives

$$|C(tu) - t\widetilde{C}(u)| \le E(C), \qquad \lim_{t \to \infty} C(tu)/t = \widetilde{C}(u), \tag{2}$$

uniformly in the score direction. Thus any $o(t)$-accurate realization has $\widetilde{C} = \Phi_{\mathcal{R}}$, and inherits the exact hard lower bound

$$|C| \ge |\widetilde{C}| \ge S_{\min}(\mathcal{R}). \tag{3}$$

The error budget is finite and computable on the actual DAG, including repeated use of shared nodes.

The framework has three consequences. First, the recession proof extends to positive scalar gates, giving an STC transfer theorem. For a compact parameter family of finite-temperature networks, the paired shadows also converge in Hausdorff distance on every compact score set, with the corresponding Kuratowski limit under parameter continuity. Second, signed affine forms have a signed tropical shadow and a valid recession bridge, with their affine costs kept explicit in Proposition 8. Third, annealing changes the deformation budget while the wiring fixes the shadow.

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.

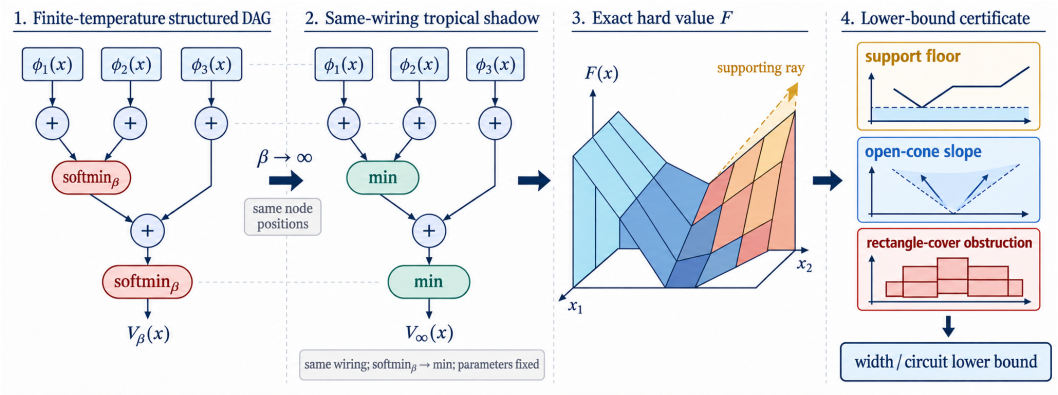

Figure 1: **From finite temperature to a lower-bound certificate.** The first two panels use the same score functions, parameters, node positions, and DAG wiring. They differ only in the value operator: a bounded softmin is replaced by its hard minimum in the tropical shadow. Under sublinear approximation on score rays, the recession argument identifies this shadow with the exact hard target. The last panel displays the quantities used for the lower-bound certificates: the support floor, the open-cone slope, and the rectangle-cover obstruction. Positive scalar gates remain in the shadow; signed affine preprocessing is charged separately in Proposition 8.

For a q-ary linear code, we prove the arbitrary-DAG bound

$$q^{\rho(\mathcal{C})} \le S_{\min}(\mathcal{C}) \le O\big(nq + nq^{2\rho(\mathcal{C})}\big), \tag{4}$$

where ρ is the minimum branch-tree interface dimension. The lower bound uses right-language rectangles and a scope-invariance lemma for shared gates. For the binary cycle code of a $k \times k$ grid, $\rho = k - 1 = \Theta(\sqrt{n})$, yielding a $2^{\Omega(\sqrt{n})}$ barrier. A deficient wiring therefore has linear error on an open cone, and the relative-support theorem gives a scale-free floor whenever the score distribution has relative interior in its support.

The GPU experiments use CUDA double precision and independent combinatorial checks. On the 5×5 grid, the measured slope vanishes exactly at width $16 = 2^\rho$, while the width-15 floor remains positive. Hamming, BCH, and Golay sweeps reproduce their code distances and interface widths; annealing reduces deformation error by 96.8% but leaves the deficient shadow slope unchanged. Training one temperature and one constant per layer reduces deformation error by $34.7\times$, with the representation slope unchanged. These observations align the finite-temperature experiments with the same shadow and floor quantities proved above.

2 Hard Inference Layers

Let X be finite, Σ a finite alphabet with $|\Sigma| = q \ge 2$, and $\mathcal{R} \subseteq \Sigma^X$ a nonempty feasible relation. Its hard inference value is

$$\Phi_{\mathcal{R}}(u) := \min_{r \in \mathcal{R}} \sum_{i \in X} u_{i,r_i}. \tag{5}$$

A pure min-plus network is a finite DAG with unary-score literals $u_{i,a}$, finite real constants, and binary $\min$ and $+$ gates. Literals and operation gates count toward size; constants are free. Sharing is unrestricted. For $\mathsf{M} = \mathbb{R}^{q|X|}$ or $\mathbb{R}^{q|X|}_{\ge 0}$, write

$$S^{\mathsf{M}}_{\min}(\mathcal{R}) := \min\{|C| : C \text{ computes } \Phi_{\mathcal{R}} \text{ on } \mathsf{M}\}. \tag{6}$$

For the scalar-extension audit, an STC is a homogeneous DAG with score literals, positive scalar gates $v = \lambda a$, $\lambda > 0$, and $+$ and $\min$ gates, with no nonzero constants. After sign reversal this is the $+/\max$ form. Under the standard convention scalar gates are free and size counts $+$ and extremum gates:

$$S_{\mathrm{STC}}(\mathcal{R}) := \min\{|C|_{\mathrm{STC}} : C \text{ is an exact STC for } \Phi_{\mathcal{R}}\}. \tag{7}$$

Theorem 7 transfers any independently established exact STC lower bound. The code theorem below concerns the scalar-free model; positive scalars create fractional intermediate traces and require the STC lower-bound input to be stated separately.

3 Bounded Deformations and the Tropical Shadow

3.1 Bounded deformations of min

Definition 1. *An operator $\psi : \mathbb{R}^d \to \mathbb{R}$ is an* ε-deformation of $\min$ *if*

$$\sup_{x \in \mathbb{R}^d} |\psi(x) - \min(x)| \le \varepsilon < \infty. \tag{8}$$

Proposition 2.

1. $\operatorname{softmin}_\beta(a,b) = -\beta \log(e^{-a/\beta} + e^{-b/\beta})$ *satisfies* $\min(a,b) - \beta \log 2 \le \operatorname{softmin}_\beta(a,b) \le \min(a,b)$.
2. *If* $\max_\Omega(x) = \max_{p \in \Delta^d}\{\langle p, x\rangle - \Omega(p)\}$ *(Mensch & Blondel, 2018) and* $L_{\Omega,d} \le \Omega \le U_{\Omega,d}$ *on* Δ^d*, then* $\min_\Omega(x) := -\max_\Omega(-x)$ *is an* ε*-deformation with* $\varepsilon = \max\{|L_{\Omega,d}|, |U_{\Omega,d}|\}$.

A *deformed inference network* is a finite DAG with unary-score literals, finite constants, $+$ gates, and binary ε_g-deformations ψ_g of $\min$. Parameters may vary by gate and may be learned. The scalar extension adds $v = \lambda a$, $\lambda > 0$, with scalar gates free under the STC size convention.

3.2 The tropical shadow

Definition 3. *The* tropical shadow $\widetilde{C}$ *of* C *replaces every* ψ_g *by* $\min$*, preserves positive scalar gates, and replaces every constant leaf by* 0*, retaining the complete DAG wiring. Without scalar gates it is a pure min-plus network; with them it is an STC.*

For the scalar-free class, define the node budget

$$E(v) = \begin{cases} 0, & v \text{ a literal leaf}, \\ |c|, & v \text{ a constant leaf } c, \\ E(a) + E(b), & v = a + b, \\ \varepsilon_v + \max\{E(a), E(b)\}, & v = \psi_v(a,b). \end{cases} \tag{9}$$

For a scalar gate $v = \lambda a$, use $E(v) = \lambda E(a)$.

Lemma 4. *For every node* v*, score vector* u*, and* $t > 0$*,*

$$|C_v(tu) - t\widetilde{C}_v(u)| \le E(v). \tag{10}$$

At the output,

$$\lim_{t \to \infty} C(tu)/t = \widetilde{C}(u) \qquad \text{for every } u. \tag{11}$$

Remark 5. *The recursive budget is DAG-aware: shared values can be added repeatedly, so local gate errors are amplified by the actual reuse pattern. On the trellis in Section 5 it gives* $n\varepsilon$*, rather than the per-gate sum* $nm\varepsilon$*.*

3.3 Transfer of exact lower bounds

Theorem 6. *If a deformed network* C *satisfies*

$$|C(tu) - \Phi_{\mathcal{R}}(tu)| = o(t) \qquad (t \to \infty) \tag{12}$$

for every admissible direction u*, then* $|C| \ge S_{\min}(\mathcal{R})$.

Proof. Positive homogeneity of $\Phi_{\mathcal{R}}$ and (12) give $C(tu)/t \to \Phi_{\mathcal{R}}(u)$. Lemma 4 gives the same limit $\widetilde{C}(u)$. Uniqueness of limits yields $\widetilde{C} = \Phi_{\mathcal{R}}$, and $|C| \ge |\widetilde{C}| \ge S_{\min}(\mathcal{R})$. □

Theorem 7. *Let C be scalar-deformed with shadow $\widetilde{C}$. Under (12) for every admissible direction, $\widetilde{C} = \Phi_{\mathcal{R}}$ and*

$$|C|_{\mathrm{STC}} \geq S_{\mathrm{STC}}(\mathcal{R}), \tag{13}$$

where both sizes count $+$ and extremum gates and treat positive scalar gates consistently.

Proof. The scalar budget in (9) proves (10) by the same topological induction. Divide by t and use positive homogeneity as in the proof of Theorem 6. □

Concrete STC consequence. After sign reversal, Theorem 7 combines with the exact STC lower bounds of Hertrich & Stargalla (2026): bounded-deformation STCs approximating the Birkhoff or directed-spanning-tree support functions with $o(t)$ error on every ray obey $|C|_{\mathrm{STC}} \geq 2^{\Omega(N)}$. The exact STC bounds are the input; the finite-temperature transfer is the new consequence.

Proposition 8. *Consider real affine gates $v = \sum_j \lambda_j a_j + c$, $\lambda_j \in \mathbb{R}$, together with bounded deformations of* min *or* max. *Define*

$$E(v) = \sum_j |\lambda_j| E(a_j) + |c| \tag{14}$$

at an affine gate and $E(v) = \varepsilon_v + \max_j E(a_j)$ at a deformed extremum. Then

$$|C(tu) - t\widetilde{C}(u)| \leq E(C). \tag{15}$$

Consequently, $o(t)$ approximation to a positively homogeneous target forces the signed tropical shadow to compute it exactly.

Proposition 9. *Let $\mathcal{F}_\beta = \{C_{\beta,\theta} : \theta \in \Theta\}$ and $\mathcal{F}_0 = \{\widetilde{C}_\theta : \theta \in \Theta\}$ satisfy*

$$\|C_{\beta,\theta} - \widetilde{C}_\theta\|_\infty \leq \eta_\beta \qquad (\theta \in \Theta), \qquad \eta_\beta \to 0. \tag{16}$$

On compact K, viewed as subsets of $C(K)$,

$$d_H(\mathcal{F}_\beta, \mathcal{F}_0) \leq \eta_\beta, \qquad \operatorname{dist}(f, \mathcal{F}_\beta) \geq \operatorname{dist}(f, \mathcal{F}_0) - \eta_\beta. \tag{17}$$

If Θ is compact and $\theta \mapsto \widetilde{C}_\theta$ is continuous into $C(K)$, then $\mathcal{F}_\beta$ converges to $\mathcal{F}_0$ in the Kuratowski sense. For one m-term log-sum-exp gate, $\eta_\beta = \log(m)/\beta$; for a DAG it is the recursive budget.

Proof. The paired parameter gives the upper and lower Hausdorff bounds in (17). For the Kuratowski lower limit use the same parameter θ. For the upper limit, choose $C_{\beta_j,\theta_j} \to f$; compactness gives $\theta_{j_\ell} \to \theta$, continuity gives $\widetilde{C}_{\theta_{j_\ell}} \to \widetilde{C}_\theta$, and the uniform bound gives $f = \widetilde{C}_\theta$. □

For compact Θ and K, the max-affine family

$$f_{\beta,\theta}(x) = \beta^{-1} \log \sum_{i=1}^{m} \exp(\beta(w_i \cdot x + b_i)), \qquad f_{0,\theta}(x) = \max_i (w_i \cdot x + b_i)$$

has $d_H \leq \log(m)/\beta$ and Kuratowski limit $\{f_{0,\theta} : \theta \in \Theta\}$. This paired class is the exact finite-parameter limit; active-subset closure conventions enlarge the parameter class and are a separate modeling choice. Thus the set-limit result is the finite-temperature link between the shadow and the hard representation bound.

$$\underbrace{|C(tu) - \Phi_{\mathcal{R}}(tu)| = o(t)}_{\text{deformed}} \Longrightarrow \underbrace{\widetilde{C} = \Phi_{\mathcal{R}}}_{\text{shadow}} \Longrightarrow \underbrace{|C| \geq |\widetilde{C}| \geq S_{\min}(\mathcal{R})}_{\text{hard bound}}. \tag{18}$$

Uniformly bounded error is a special case.

Corollary 10. *The shadow depends on wiring, while $E(C)$ depends on deformation parameters and constants. Annealing therefore shrinks the deformation error $|C - \widetilde{C}| \leq E(C)$ with the representation error $\widetilde{C} - \Phi_{\mathcal{R}}$ fixed by the wiring.*

Theorem 11. *If $|C| < S_{\min}(\mathcal{R})$, there are $\delta > 0$ and a nonempty relatively open set U of unit directions such that*

$$|C(tu) - \Phi_{\mathcal{R}}(tu)| \geq \tfrac{\delta}{4} t \qquad (u \in U,\ t > 4E(C)/\delta). \tag{19}$$

If u_0 maximizes the shadow gap, then the bound improves to $\delta t/2$ for $t \geq 2E(C)/\delta$.

Proof. Theorem 6 gives $g := \widetilde{C} - \Phi_{\mathcal{R}} \neq 0$. Continuity and homogeneity give a relatively open U with $|g(u)| \geq \delta/2$. Lemma 4 then yields $|C(tu) - \Phi_{\mathcal{R}}(tu)| \geq t|g(u)| - E(C)$, which implies (19) and the directional bound. □

Proposition 12. *Let $g = \widetilde{C} - \Phi_{\mathcal{R}}$, and let ν have finite first moment, support K in a linear score space M, and $K = \operatorname{relint}_M K$ with nonempty relative interior. For $D_\nu(g) := \min_s \mathbb{E}_\nu|g - s|$, every $t > 0$ and $b \in \mathbb{R}$ satisfy*

$$\mathbb{E}_{u\sim\nu} \frac{|C(tu) + b - \Phi_{\mathcal{R}}(tu)|}{t} \geq D_\nu(g) - \frac{E(C)}{t}. \tag{20}$$

Moreover, $D_\nu(g) = 0$ exactly when $\widetilde{C} = \Phi_{\mathcal{R}}$ on $\operatorname{relint}_M K$.

Proof. Lemma 4 and homogeneity give the left side at least $\mathbb{E}|g(u) + b/t| - E(C)/t$. Minimizing over $b/t \in \mathbb{R}$ gives (20). If $D_\nu(g) = 0$, g is constant on the relative interior of K; a homogeneous piecewise-linear function constant there is zero. The converse is immediate. □

3.4 The class the theorems reach

Write

$$\mathcal{D}_\varepsilon := \{\psi : \sup_x |\psi(x) - \min(x)| \leq \varepsilon\},$$
$$\mathfrak{C} := \{\text{finite DAGs over } \{u_{i,a}\} \cup \mathbb{R} \cup \{+\} \cup \bigcup_{\varepsilon<\infty} \mathcal{D}_\varepsilon\}. \tag{21}$$

This includes hard min-plus inference, softmin at every temperature, the Ω-regularized operators of Mensch & Blondel (2018), arbitrary sharing, finite constants, gate-specific temperatures, and annealing schedules. Sign reversal gives the corresponding max-plus form, including smoothed Viterbi and the max-plus subclass of morphological networks (Shen et al., 2022). Positive scalar gates form the STC extension covered by Theorem 7. Signed affine preprocessing has the recession bridge of Proposition 8; its shadow and affine cost convention determine the lower bound in that larger class.

4 Instantiating the Barrier: Linear Codes

We now supply an exact hard-layer bound for q-ary linear codes and a family of ordinary constrained-decoding instances with exponential size.

4.1 Interface dimension

Let $\mathcal{C} \leq \mathbb{F}_q^X$, $n = |X|$. For $X = L \dot\cup R$, let $\mathcal{C}_L, \mathcal{C}_R$ be the subcodes supported in L, R, and define

$$\lambda_{\mathcal{C}}(L) := \dim \mathcal{C} - \dim \mathcal{C}_L - \dim \mathcal{C}_R = \dim \mathcal{C}/(\mathcal{C}_L \oplus \mathcal{C}_R). \tag{22}$$

A branch tree is an unrooted degree-three tree with leaves bijectively labelled by X. Its interface dimension is

$$\rho(\mathcal{C}) := \min_T \max_{e\in E(T)} \lambda_{\mathcal{C}}(L_e). \tag{23}$$

This is a code invariant; under the convention of Kashyap (2009), $\rho = \mathrm{bw}(\mathcal{C}) - 1$.

Theorem 13. *For every q-ary linear code and either score domain,*

$$S_{\min}(\mathcal{C}) \geq q^{\rho(\mathcal{C})}. \tag{24}$$

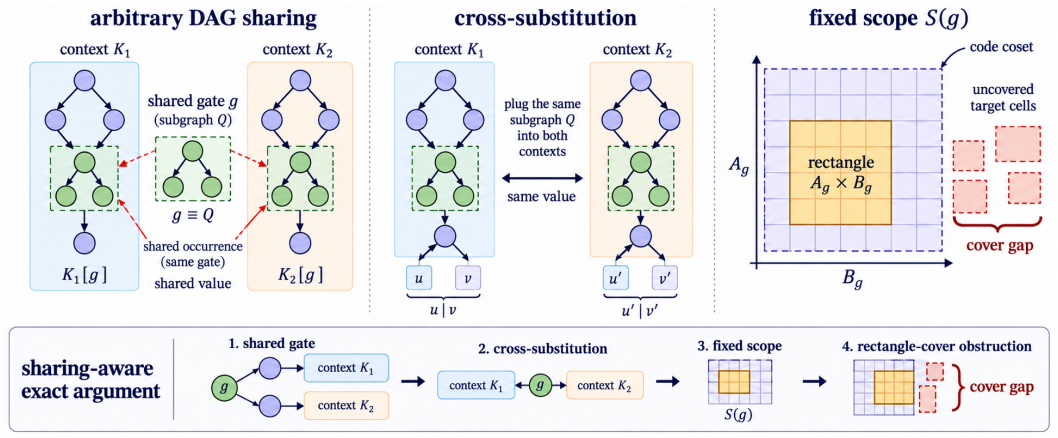

Figure 2: **Sharing-aware rectangle argument.** A gate reused in contexts K_1 and K_2 has one value. Cross-substitution therefore fixes its scope $S(g)$; after the scope is split into A_g and B_g, the gate contributes one rectangle $A_g \times B_g$ in a fixed code coset. The red target cells mark the cover obstruction used in the counting argument.

Formal lower-bound chain. For a complete parse of a size-s DAG, the appendices' safety and gate-parse lemmas produce a branch tree and, for each live gate g, a fixed scope $S(g)$ and a rectangle $A_g \times B_g$. The fixed-scope statement follows from cross-substitution $K_1[Q_1], K_2[Q_2]$: equality of the substituted degree and one-hotness of degree-n parses force $\deg Q_1 = \deg Q_2$ and $S(Q_1) = S(Q_2)$. At a branch-tree edge with $\lambda_{\mathcal{C}}(L) \geq \rho$, the $q^{\lambda_{\mathcal{C}}(L)}$ cosets of $\mathcal{C}/(\mathcal{C}_L \oplus \mathcal{C}_R)$ are maximal feasible rectangles. Each gate rectangle covers at most one such coset, while all $|\mathcal{C}|$ codewords must be covered. Hence $s \geq q^{\rho}$. Appendix B gives the degree, scope, and rectangle arguments on both score domains.

4.2 Upper bound and grid family

Proposition 14. *For $K = q^{\rho(\mathcal{C})}$, the code has a pure min-plus realization of size $O(nq + nK^2)$.*

For a cut and $a \in \Sigma^L$, define the right language $\mathsf{L}_{\mathcal{R},L}(a) := \{b : a \oplus b \in \mathcal{R}\}$. Its equivalence classes are the messages carried by a decomposition. A branch tree attaining ρ carries at most $K = q^{\rho}$ classes per oriented edge; binary joins enumerate K^2 pairs. Therefore

$$q^{\rho} \leq S_{\min}(\mathcal{C}) \leq O\big(nq + nq^{2\rho}\big)\,. \tag{25}$$

State tables are relation-dependent wiring and are free in the circuit-size measure. For fixed q, this gives polynomial size at $\rho = O(\log n)$, subexponential size at $\rho = o(n)$, and the $2^{\Theta(n)}$ regime at $\rho = \Theta(n)$.

Consider the binary cycle code $\mathcal{C}_{G_k}$ of the $k \times k$ grid: edge subsets of even degree, with $n = 2k(k-1)$.

Proposition 15. *For $k \geq 3$,*

$$\rho(\mathcal{C}_{G_k}) = k - 1 = \Theta(\sqrt{n}), \qquad 2^{k-1} \leq S_{\min}(\mathcal{C}_{G_k}) \leq 2^{O(\sqrt{n})}. \tag{26}$$

Proof. The code connectivity function equals that of its dual, and the dual of a connected graph's cycle code is its cocycle code. Hence $\lambda_{\mathcal{C}_{G_k}} = \lambda_{M(G_k)}$ and $\rho = \mathrm{bw}(M(G_k)) - 1$. A grid tangle of order k (Geelen et al., 2009) and the tangle-branchwidth theorem (Robertson & Seymour, 1991; Geelen et al., 2009) give $\mathrm{bw}(M(G_k)) \geq k$; the left-to-right edge sweep gives the matching upper bound. Substitution in (25) proves (26). □

The sweep realizes the code's optimal ρ, so its interface width is an instance invariant. The circuit size remains between the q^{ρ} and $q^{2\rho}$ terms.

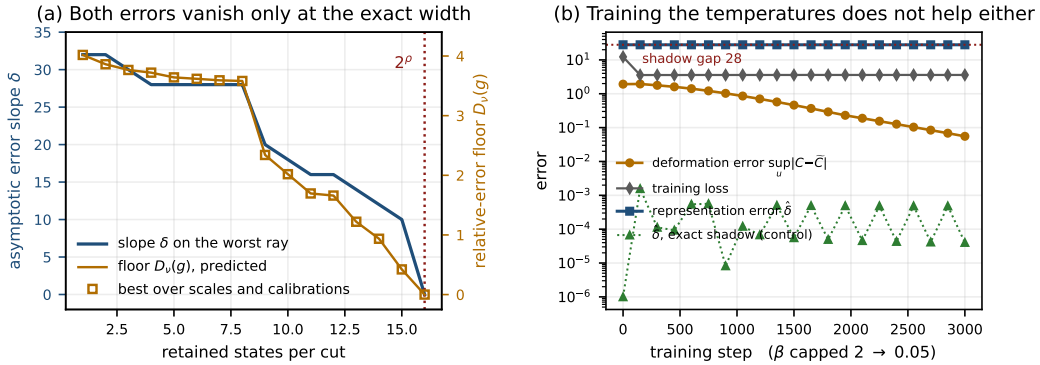

Figure 3: **Exact-width barrier and trained annealing.** (a) For the 5×5 grid, both the open-cone slope and the relative-support floor vanish only at $m = 2^\rho = 16$; markers are the exhaustive scale and calibration search. (b) On fixed deficient $m = 8$ wiring, learned temperatures reduce deformation error by $34.7\times$ while the representation slope stays at the shadow gap. The full-width control reaches zero slope.

Corollary 16. *For a q-ary linear-code inference layer built from $+$ and Ω-regularized* min *gates with bounded simplex regularizer, arbitrary sharing, finite constants, and learned temperatures,*

$$|C| \geq q^{\rho(C)}.$$

For the $k \times k$ grid cycle code this is $2^{\Theta(\sqrt{n})}$ at every temperature. A smaller wiring has the open-cone error of Theorem 11 and the relative-support floor of Proposition 12.

5 Experiments

The target is the exact inference value Φ_C, evaluated by a full trellis. We use binary grid cycle codes for the structured sweeps and standard block codes for the external-code check. The results below report the empirical consequences of the propositions; the complete protocol, precision policy, validation paths, and artifact manifest are collected in Appendix D.

5.1 Decomposition search

E1. Across grid sides $k = 3, \ldots, 12$, the best orders attain the exact $\rho = k - 1$, confirming that the order used by the trellis realizes the code invariant from Proposition 15.

5.2 Budget and width

E2. For the 5×5 grid, the width sweep recovers the exact threshold $m = 2^\rho = 16$: the open-cone slope reaches zero only at full width, while the deficient widths agree with the Hamming distance to the surviving codeword set. At width 15, the relative-support floor remains 0.42, despite retaining most of the exact trellis.

5.3 Finite-temperature operators

E3. On fixed deficient wiring with shadow gap 28, lowering the smoothing scale reduces deformation error by 96.8%, while the representation error is unchanged. The full-width trellis remains the exact-shadow control.

5.4 Standard block codes

E4. Hamming$(7, 4)$, BCH$(15, 7)$, and extended Golay G_{24} recover their expected distance slopes. For G_{24}, the searched order has interface dimension 9, matching the Muder minimal-trellis bound (Muder, 1988); the natural cyclic order has dimension 12.

5.5 Trained annealing

E5. On the same deficient $m = 8$ wiring, annealing reduces the gates-only deformation error by $34.7\times$, while the measured slope stays around the shadow gap 28. The full-width control reaches near-zero slope.

5.6 Score supports

E6. On nested Gaussian score supports, the width-8 floor decreases from 3.58 at rank 80 to 0.82 at rank 2, but remains positive on the two-dimensional support. The exact-width circuit has zero floor at every rank.

6 Related Work

Differentiable dynamic programming replaces hard $\min$ / $\max$ recurrences by regularized operators (Mensch & Blondel, 2018). Our contribution is a representation theorem for this setting: bounded value-operator deformation preserves the same-wiring tropical shadow and its exact lower bound. Tropical circuit lower bounds are surveyed by Jukna (2023); the closest approximation result, Jukna & Seiwert (2020), studies multiplicative ratios for combinatorial optimization, whereas we study additive error of the value function over score rays. The finite DAG budget is the quantitative ingredient that turns the recession limit into finite-scale slope and floor certificates.

The STC model of Hertrich & Stargalla (2026) proves exponential exact lower bounds for Birkhoff and directed-spanning-tree support functions using fractional traces. Our Theorem 7 transfers any such exact STC bound to bounded-deformation implementations. The min-plus/ABP framework of Komarath et al. (2026) and connectivity lower bounds of Kluk & Nederlof (2026) provide complementary exact circuit results. Hausdorff compression for tropical neural networks (Fotopoulos et al., 2026) also uses a data-free function-class distance; our paired-class bound identifies the finite-temperature class, its shadow class, and the explicit $\log(m)/\beta$ or DAG budget connecting them. Neural compilation (Kratsios et al., 2026), Tropical Attention benchmarks (Hashemi et al., 2025), tropical attention limits (Alpay & Senturk, 2026), and regular-matroid neural constructions (Hertrich et al., 2026) supply adjacent neural motivation and model boundaries. The present result concerns the value-function interface and provides its lower-bound certificate.

The proof combines code width with rectangle methods. The interface $\dim \mathcal{C} - \dim \mathcal{C}_L - \dim \mathcal{C}_R$ is the state quantity in minimal tree realizations of codes (Kashyap, 2009); right languages, decomposability, and rectangle covers are standard in knowledge compilation (Pipatsrisawat & Darwiche, 2008; Bova et al., 2016), and communication arguments control auxiliary variables in CNF encodings (Mengel & Wallon, 2020). Here those ingredients yield a circuit-size theorem: cross-substitution supplies a fixed scope for a gate reused by an arbitrary DAG, after which the code-coset rectangles give q^{ρ}. The stronger $2^{\Omega(k \log\log k)}$ TSP bound of Kluk & Nederlof (2026) illustrates a different, problem-specific width regime; any exact bound of that form can also be passed through Theorem 6.

7 Discussion: Scope and Next Models

Operator class. The main class $\mathfrak{C}$ contains $+$, bounded deformations of $\min$ / $\max$, finite constants, arbitrary sharing, positive scalars, and annealed gate parameters. Effective subtraction produces tropical rational maps (Zhang et al., 2018); alternating $\min$ / $\max$ architectures (Luo & Fan, 2021), division, and hard attention likewise require the larger signed or rational shadow model. Proposition 8 provides the recession bridge for signed affine softmax/maxout forms, with the affine budget explicit. A lower bound for that shadow is the next model-level extension.

Score support and exact relation. Theorem 11 gives an open-cone certificate and Proposition 12 gives a scale-free certificate on supports with relative interior in a linear score space. The nested-subspace experiment shows the floor remains nonzero on a two-dimensional support of the eighty-dimensional score space. A theorem for an arbitrary learned encoder support would extend this interface result to end-to-end distributions. The concrete $2^{\Omega(\sqrt{n})}$ statement is for the grid cycle-code

family; replacing that relation by another structured family only requires its exact hard-layer lower bound.

8 Conclusion

Bounded smoothing preserves the tropical shadow of a hard structured-inference layer. The shadow transfers every exact min-plus lower bound with no loss in size, while the recursive DAG budget gives finite-temperature, open-cone, and relative-support certificates. For linear-code inference, $q^\rho \le S_{\min} \le O(nq + nq^{2\rho})$, and grid cycle codes give the $2^{\Omega(\sqrt{n})}$ barrier.

The GPU experiments recover the exact width threshold, code-distance slopes, support-restricted floors, and the separation between deformation optimization and representation optimization. Positive scalar STCs inherit the recession transfer; signed affine, rational, and encoder-supported extensions define the next lower-bound frontiers.

Reproducibility statement

Every assumption needed for the theoretical claims is stated in Sections 2–4, with complete proofs in Appendices A–C. The supplementary artifact contains the Python scripts, the fixed seed used throughout, the generated CSV files, and the figure-generation code for Section 5, together with a test script that verifies the shadow bound, the coset rectangle partition, and the trellis against exhaustive enumeration on small instances. Three of the relations are standard block codes, each checked against its published dimension, minimum distance and weight enumerator before use; the scores are synthetic throughout, and no external dataset or pretrained model is required.

The numerical protocol and implementation checks, including the CUDA double-precision policy and the independent integer reference path, are given in Appendix D; the artifact records the GPU identity and the SHA-256 manifest beside the results.

AI use statement

In this work, we used generative AI tools to help develop theoretical models, to formulate mathematical claims and supply ingredients for proving them, to assist in the writing of proofs, to design and give feedback on the experiments, to implement methods, and to interpret results. We did not use generative AI tools to generate data: the synthetic scores come from seeded numerical code in the artifact. The remaining categories requiring disclosure are not applicable to this work. Among the recommended categories, we additionally used these tools to identify and summarize literature, to create and edit software code and scientific figures, to draft parts of the paper, and to edit it for readability. We have reviewed all AI-assisted work: every mathematical claim was re-derived by the authors, every reference was checked against its published record, and all code was executed and cross-checked against independent brute-force implementations. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

References

Faruk Alpay and Bilge Senturk. The geometry of thought: Disclosing the transformer as a tropical polynomial circuit. *arXiv preprint arXiv:2601.09775*, 2026.

Simone Bova, Florent Capelli, Stefan Mengel, and Friedrich Slivovsky. Knowledge compilation meets communication complexity. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1008–1014, 2016.

Konstantinos Fotopoulos, Petros Maragos, and Panagiotis Misiakos. TropNNC: Structured neural network compression using tropical geometry. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=u7DRq1icmY.

Jim Geelen, Bert Gerards, and Geoff Whittle. Tangles, tree-decompositions and grids in matroids. *Journal of Combinatorial Theory, Series B*, 99(4):657–667, 2009. doi: 10.1016/j.jctb.2007.10.008.

Baran Hashemi, Kurt Pasque, Chris Teska, and Ruriko Yoshida. Tropical attention: Neural algorithmic reasoning for combinatorial algorithms. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, pp. 109306–109331. Neural Information Processing Systems Foundation, Inc., 2025. doi: 10.52202/085713-3295.

Christoph Hertrich and Moritz Stargalla. Tropical circuits with scalar multiplication gates. *arXiv preprint arXiv:2607.11540*, 2026.

Christoph Hertrich, Stefan Kober, and Georg Loho. Arithmetic circuits and neural networks for regular matroids. In *Integer Programming and Combinatorial Optimization (IPCO 2026)*, volume 16588 of *Lecture Notes in Computer Science*, pp. 394–409. Springer, 2026. doi: 10.1007/978-3-032-28691-8_26.

Stasys Jukna. *Tropical Circuit Complexity: Limits of Pure Dynamic Programming*. Springer, Cham, 2023. doi: 10.1007/978-3-031-42354-3.

Stasys Jukna and Hannes Seiwert. Approximation limitations of pure dynamic programming. *SIAM Journal on Computing*, 49(1):170–205, 2020. doi: 10.1137/18M1196339.

Navin Kashyap. On minimal tree realizations of linear codes. *IEEE Transactions on Information Theory*, 55(8):3501–3519, 2009. doi: 10.1109/TIT.2009.2023718.

Kacper Kluk and Jesper Nederlof. Lower bounds on pure dynamic programming for connectivity problems on graphs of bounded path-width. In *53rd International Colloquium on Automata, Languages, and Programming (ICALP 2026)*, volume 374 of *Leibniz International Proceedings in Informatics*, pp. 130:1–130:23. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2026. doi: 10.4230/LIPIcs.ICALP.2026.130.

Balagopal Komarath, Harshil Mittal, and Jayalal Sarma. VP, VNP and algebraic branching programs over min-plus semirings. In *53rd International Colloquium on Automata, Languages, and Programming (ICALP 2026)*, volume 374 of *Leibniz International Proceedings in Informatics*, pp. 132:1–132:23. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2026. doi: 10.4230/LIPIcs.ICALP.2026.132.

Anastasis Kratsios, Simone Brugiapaglia, Bum Jun Kim, Gregory Cousins, and Haitz Sáez de Ocáriz Borde. Algorithmic foundations of deep learning: Complexity-theoretic rates and a characterization of universal approximation. *arXiv preprint arXiv:2606.26705*, 2026.

Ye Luo and Shiqing Fan. Min-max-plus neural networks. *arXiv preprint arXiv:2102.06358*, 2021.

Stefan Mengel and Romain Wallon. Graph width measures for CNF-encodings with auxiliary variables. *Journal of Artificial Intelligence Research*, 67:409–436, 2020. doi: 10.1613/jair.1.11750.

Arthur Mensch and Mathieu Blondel. Differentiable dynamic programming for structured prediction and attention. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 3462–3471. PMLR, 2018.

D. J. Muder. Minimal trellises for block codes. *IEEE Transactions on Information Theory*, 34(5): 1049–1053, 1988. doi: 10.1109/18.21228.

Knot Pipatsrisawat and Adnan Darwiche. New compilation languages based on structured decomposability. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence*, pp. 517–522, 2008.

Neil Robertson and Paul D. Seymour. Graph minors. X. obstructions to tree-decomposition. *Journal of Combinatorial Theory, Series B*, 52(2):153–190, 1991. doi: 10.1016/0095-8956(91)90061-N.

Yucong Shen, Frank Y. Shih, Xin Zhong, and I-Cheng Chang. Deep morphological neural networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 36(12):2252023, 2022. doi: 10.1142/S0218001422520231.

Liwen Zhang, Gregory Naitzat, and Lek-Heng Lim. Tropical geometry of deep neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 5824–5832. PMLR, 2018.

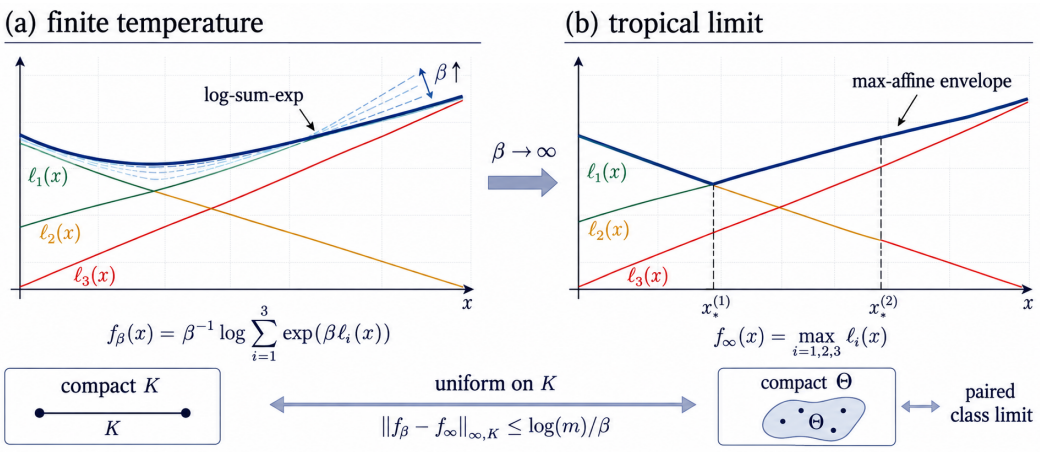

Figure 4: **Finite-temperature class limit.** For fixed affine scores ℓ_i, the log-sum-exp curve converges to the max-affine envelope as $\beta \to \infty$, with $\max_i \ell_i \le f_\beta \le \max_i \ell_i + \log(m)/\beta$. Compactness of the parameter family and uniformity on a compact score set give the paired function-class limit in Proposition 9. The lower-bound transfer is proved separately by the same-wiring recession argument below.

A Proofs for Section 3

The main text gives the induction for Lemma 4 and the limit argument for Theorem 6; this appendix records the deferred positive-scalar and signed-affine inductions, the quantitative metric transfer, the remaining proofs for Section 3, the arbitrary-fan-in form, and the max-plus form.

Proof of Proposition 2. (1) With $m := \min(a, b)$, and applying the decreasing map $z \mapsto -\beta \log z$ to the middle term,

$$e^{-m/\beta} \;\le\; e^{-a/\beta} + e^{-b/\beta} \;\le\; 2e^{-m/\beta} \qquad \Longrightarrow \qquad m - \beta \log 2 \;\le\; \mathrm{softmin}_\beta(a, b) \;\le\; m.$$

(2) Fix $i^\star \in \arg\max_i x_i$. Bounding $\max_\Omega$ below by the single feasible point $p = e_{i^\star}$ and above by $\langle p, x \rangle \le \max(x)$ for all $p \in \Delta^d$,

$$\max(x) - U_{\Omega,d} \;\le\; \langle e_{i^\star}, x \rangle - \Omega(e_{i^\star}) \;\le\; \max_{p \in \Delta^d} \{\langle p, x \rangle - \Omega(p)\} \;\le\; \max(x) - L_{\Omega,d}.$$

Substituting $x \mapsto -x$ and negating turns this into $\min(x) + L_{\Omega,d} \le \min_\Omega(x) \le \min(x) + U_{\Omega,d}$, so $\varepsilon = \max\{|L_{\Omega,d}|, |U_{\Omega,d}|\}$ is admissible. □

Proof of Lemma 4. Write $\Delta_v := \sup_{u,\, t>0} \big| C_v(tu) - t\widetilde{C}_v(u) \big|$. We show $\Delta_v \le E(v)$ by induction along a topological order of the DAG, one line per node type:

$$\begin{aligned}
v = u_{i,a}: &\quad C_v(tu) = tu_{i,a} = t\widetilde{C}_v(u) &&\Longrightarrow \quad \Delta_v = 0 = E(v),\\
v = c: &\quad C_v \equiv c,\;\; \widetilde{C}_v \equiv 0 &&\Longrightarrow \quad \Delta_v = |c| = E(v),\\
v = a + b: &\quad \Delta_v \le \Delta_a + \Delta_b &&\Longrightarrow \quad \Delta_v \le E(a) + E(b) = E(v),\\
v = \psi_v(a, b): &\quad \Delta_v \le \varepsilon_v + \max\{\Delta_a, \Delta_b\} &&\Longrightarrow \quad \Delta_v \le \varepsilon_v + \max\{E(a), E(b)\} = E(v).
\end{aligned}$$

Only the last line needs an argument. Split it through $\min$ and use Definition 1 on the first term and the 1-Lipschitz property $|\min(a, b) - \min(a', b')| \le \max(|a - a'|, |b - b'|)$ on the second:

$$\begin{aligned}
\big| C_v(tu) - t\widetilde{C}_v(u) \big| \le\; & \underbrace{\big| \psi_v\big(C_a(tu), C_b(tu)\big) - \min\big(C_a(tu), C_b(tu)\big) \big|}_{\le\, \varepsilon_v} \\
& + \underbrace{\big| \min\big(C_a(tu), C_b(tu)\big) - \min\big(t\widetilde{C}_a(u), t\widetilde{C}_b(u)\big) \big|}_{\le\, \max\{\Delta_a, \Delta_b\}}.
\end{aligned}$$

Taking v to be the output node gives (10). Dividing by t and using $\widetilde{C}(tu) = t\widetilde{C}(u)$ gives (11). □

Lemma 17. *Let C be a scalar-deformed inference network. In addition to (9), set*

$$E(v) := \lambda E(a) \quad \textit{when } v = \lambda a,\ \lambda > 0.$$

Then its scalar tropical shadow satisfies

$$\left|C_v(tu) - t\widetilde{C}_v(u)\right| \le E(v) \qquad \textit{for every node } v, u, \textit{ and } t > 0.$$

Proof. The proof is the same topological induction as for Lemma 4. The literal, constant, addition, and deformed-min cases are unchanged. At a positive scalar gate,

$$\left|C_v(tu) - t\widetilde{C}_v(u)\right| = \lambda\left|C_a(tu) - t\widetilde{C}_a(u)\right| \le \lambda E(a) = E(v).$$

Positivity of λ is also what preserves positive homogeneity of the shadow. Applying the bound at the output and dividing by t proves the corresponding recession limit. □

Proof of Theorem 7. By Lemma 17,

$$\frac{C(tu)}{t} - \widetilde{C}(u) \longrightarrow 0.$$

The assumed sublinear approximation and positive homogeneity of $\Phi_{\mathcal{R}}$ give $C(tu)/t \to \Phi_{\mathcal{R}}(u)$. Uniqueness of limits yields $\widetilde{C}(u) = \Phi_{\mathcal{R}}(u)$ for every admissible direction. The shadow has exactly the same addition and min/max gates as C, and only removes constants, so $|C|_{\mathrm{STC}} \ge |\widetilde{C}|_{\mathrm{STC}}$. Since the shadow is an exact STC for $\Phi_{\mathcal{R}}$, the second inequality follows from the definition of $S_{\mathrm{STC}}(\mathcal{R})$. □

Proof of Proposition 8. Use topological induction. At an affine gate,

$$\begin{aligned}&\left|\sum_j \lambda_j C_{a_j}(tu) + c - t\sum_j \lambda_j \widetilde{C}_{a_j}(u)\right| \\ &\qquad \le \sum_j |\lambda_j|\left|C_{a_j}(tu) - t\widetilde{C}_{a_j}(u)\right| + |c| \le \sum_j |\lambda_j| E(a_j) + |c| = E(v).\end{aligned}$$

At a deformed min/max gate, insert the exact extremum and use its 1-Lipschitz property in the supremum norm, exactly as in the proof of Lemma 4. The output bound follows. Division by t and positive homogeneity of the signed shadow give $C(tu)/t \to \widetilde{C}(u)$. Comparing this limit with the assumed $o(t)$ approximation and the positive homogeneity of the target proves exact equality. □

Proof of Proposition 9. The hypothesis $\eta_\beta \to 0$ is what makes the limit statement non-vacuous. For every $\theta \in \Theta$,

$$\|C_{\beta,\theta} - \widetilde{C}_\theta\|_\infty \le \eta_\beta.$$

Thus each element of $\mathcal{F}_\beta$ lies within η_β of $\mathcal{F}_0$, and, because the pairing uses the same parameter θ, each element of $\mathcal{F}_0$ lies within η_β of $\mathcal{F}_\beta$. Taking the two sup–inf terms in the definition of the Hausdorff distance proves $d_H(\mathcal{F}_\beta, \mathcal{F}_0) \le \eta_\beta$. For any f and any θ, the reverse triangle inequality gives

$$\|f - C_{\beta,\theta}\|_\infty \ge \|f - \widetilde{C}_\theta\|_\infty - \eta_\beta \ge \operatorname{dist}(f, \mathcal{F}_0) - \eta_\beta.$$

Taking the infimum over θ proves the distance statement. For the Kuratowski upper limit, let f_{β_j,θ_j} converge in $C(K)$. Compactness of Θ gives a subsequence $\theta_j \to \theta$, and continuity of $\theta \mapsto \widetilde{C}_\theta$, together with $\|C_{\beta_j,\theta_j} - \widetilde{C}_{\theta_j}\|_\infty \to 0$, identifies the limit as $\widetilde{C}_\theta$. The lower-limit inclusion follows by fixing any θ and using the paired sequence $C_{\beta,\theta}$. Finally, $\max z \le \beta^{-1}\log\sum_{j=1}^m e^{\beta z_j} \le \max z + \log(m)/\beta$, which gives the single-gate constant; induction through additions, scalar gates, and deformed min/max gates gives the stated DAG recursion. □

Throughout the next two proofs put $g := \widetilde{C} - \Phi_{\mathcal{R}}$, which is continuous, piecewise linear and positively homogeneous.

Proof of Theorem 11. $\widetilde{C} \neq \Phi_{\mathcal{R}}$, since otherwise $\widetilde{C}$ would be an exact pure min-plus network of size $|\widetilde{C}| \leq |C| < S_{\min}(\mathcal{R})$. Pick a nonzero direction where $g \neq 0$, normalize it to $u_0 \in \mathbb{S}^{d-1}$, and put $\delta := |g(u_0)| > 0$. On the single direction u_0, Lemma 4 gives

$$\big|C(tu_0) - \Phi_{\mathcal{R}}(tu_0)\big| \;\geq\; t\,|g(u_0)| - E(C) \;=\; t\delta - E(C) \;\geq\; \tfrac{\delta}{2}t \qquad \text{for } t \geq 2E(C)/\delta.$$

Continuity of g restricted to the unit-sphere section of the admissible score domain gives a relatively open neighborhood $U \ni u_0$ with $|g(u)| > \delta/2$ for every $u \in U$. Subtracting $E(C)$ as above and using the same homogeneous scaling,

$$|C(tu) - \Phi_{\mathcal{R}}(tu)| \geq \tfrac{\delta}{2}t - E(C) \geq \tfrac{\delta}{4}t$$

for $u \in U$ and $t > 4E(C)/\delta$. □

Proof of Proposition 12. Lemma 4 bounds the integrand pointwise, and b/t ranges over all of $\mathbb{R}$ as (t, b) does:

$$\mathbb{E}_\nu \frac{\big|C(tu) + b - \Phi_{\mathcal{R}}(tu)\big|}{t} \;\geq\; \mathbb{E}_\nu \Big|g(u) + \tfrac{b}{t}\Big| - \frac{E(C)}{t} \;\geq\; \inf_{s\in\mathbb{R}} \mathbb{E}_\nu \big|g - s\big| - \frac{E(C)}{t} \;=\; D_\nu(g) - \frac{E(C)}{t}.$$

For the characterization, let $K := \operatorname{supp}\nu$, let $s^\star$ attain $D_\nu(g)$, and suppose $D_\nu(g) = 0$. Such a minimizer exists: g has at most linear growth, so the finite first moment gives $g \in L^1(\nu)$, while $s \mapsto \mathbb{E}_\nu|g - s|$ is continuous and coercive. Then $g = s^\star$ ν-a.e. Continuity and the definition of the topological support imply $g \equiv s^\star$ on K, hence on $\operatorname{relint}_M K$. Choose a nonempty relatively open subset of $\operatorname{relint}_M K$ lying in one linearity cell of the piecewise-linear restriction $g|_M$. On that subset $g(u) = \langle \alpha, u\rangle$ for some linear functional α on M. Since it is constant on a relatively open set, $\alpha = 0$ on M, and homogeneity forces $s^\star = 0$. Thus $g = 0$ on that cell and, by repeating the argument over the cells meeting $\operatorname{relint}_M K$, on the whole relative interior. Conversely, if $g = 0$ on $\operatorname{relint}_M K$, continuity and $K = \overline{\operatorname{relint}_M K}$ for the support in the proposition give $g = 0$ ν-a.e., so $D_\nu(g) = 0$. □

Lemma 18. *Let C be a network whose non-additive gates ψ_g have fan-in $d_g \leq D$ and satisfy $\sup_x |\psi_g(x) - \min(x)| \leq \varepsilon_g$. Define $E(v) = \varepsilon_v + \max_i E(a_i)$ at such a gate and as in (9) otherwise. Then (10) holds verbatim, and the binary expansion of $\widetilde{C}$ has at most $(D-1)|C|$ gates.*

Proof. Identical to Lemma 4, using that $\min$ on $\mathbb{R}^d$ is 1-Lipschitz for the supremum norm. For the second claim, replace each d_g-ary $\min$ by a balanced tree of $d_g - 1$ binary $\min$ gates. □

Consequently every lower bound stated for the binary model applies to a bounded-fan-in deformed network after dividing by D; in the layers we consider D is a constant.

Proposition 19. *Let $\Psi_{\mathcal{R}}(u) = \max_{r\in\mathcal{R}} \sum_i u_{i,r_i}$ and let C be a network of $+$ gates and ε_g-deformations of $\max$ on the full-real score space. Then Lemma 4 and Theorems 6–11 hold with $\min$ replaced by $\max$ and Φ by Ψ.*

Proof. Negate every literal and every intermediate value. This maps $\max$ to $\min$, $+$ to $+$, a constant c to $-c$ and an ε-deformation of $\max$ to one of $\min$, preserving the number of gates, and sends $\Psi_{\mathcal{R}}(u)$ to $\Phi_{\mathcal{R}}(-u)$. □

B Proof of the Exact Lower Bound

Fix $\mathcal{C} \leq \mathbb{F}_q^X$ with $n = |X|$, and let a pure min-plus DAG compute $\Phi_{\mathcal{C}}$ on $\mathsf{M} \in \{\mathbb{R}, \mathbb{R}_{\geq 0}\}$. A *parse* chooses one child at a $\min$ gate and both children at a $+$ gate; shared subcircuits are unrolled when a parse is viewed as a tree. Every finite output parse P computes an affine form $\kappa_P + \langle u, \alpha_P\rangle$ with $\alpha_P \in \mathbb{Z}_{\geq 0}^{qn}$ counting literal occurrences; its degree is $|\alpha_P|_1$. For $f \in \Sigma^X$ let χ_f be the one-hot exponent vector.

Lemma 20. *For nonempty $F \subseteq \Sigma^X$, $\operatorname{conv}\{\chi_f : f \in F\} \cap \mathbb{Z}_{\geq 0}^{qn} = \{\chi_f : f \in F\}$.*

Table 1: Which operators the theorems of Section 3 reach. Every gate on the left is a bounded deformation of $\min$ or $\max$, so the network has a tropical shadow and inherits the exact bound; positive scalar gates use the STC shadow. Signed affine networks on the right have the separate recession bridge of Proposition 8, but not the code lower bound.

Inside $\mathfrak{C}$ or $\mathfrak{C}_+$	**Separate signed-affine boundary**
hard $\min$/$\max$; $\mathrm{softmin}_\beta$ and $\mathrm{softmax}_\beta$ at every $\beta > 0$	effective subtraction, hence every tropical rational map (Zhang et al., 2018), including bias-free ReLU networks
$\min_\Omega, \max_\Omega$ for every Ω bounded on Δ^d: negative entropy, squared ℓ_2, either instance of Mensch & Blondel (2018); smoothed Viterbi; the max-plus subclass of morphological networks (Shen et al., 2022)	alternating $\min$ / $\max$ architectures (Luo & Fan, 2021); division; hard-attention Transformers

Proof. Every convex-hull point lies in the product of simplices $\{x \geq 0 : \sum_{a\in\Sigma} x_{i,a} = 1 \ \forall i\}$, so an integer point of the hull is χ_d for some $d \in \Sigma^X$. If $\chi_d = \sum_j \theta_j \chi_{f_j}$ with $\theta_j > 0$, every positively weighted f_j agrees with d in every coordinate, so $d \in F$. □

Lemma 21. *1. On the full-real score space every finite output parse has exponent χ_f for some $f \in \mathcal{C}$.*

2. On the nonnegative cone no finite output parse has degree below n, and every degree-n parse has exponent χ_f for some $f \in \mathcal{C}$.

3. In either model every $f \in \mathcal{C}$ occurs as the exponent of some degree-n output parse.

Proof. Put $P := \mathrm{conv}\{\chi_f : f \in \mathcal{C}\}$. Correctness of the circuit means $\min_P \langle u, \cdot\rangle = \Phi_{\mathcal{C}}(u) \leq \kappa_P + \langle u, \alpha\rangle$ for every output parse exponent α.

(1) Suppose $\alpha \notin P$ on the full-real domain. Strict separation gives u with $\langle u, \alpha\rangle < \min_{x\in P}\langle u, x\rangle$, and evaluating along tu,

$$\kappa_P + t\langle u, \alpha\rangle \ < \ \kappa_P + t \min_{x\in P}\langle u, x\rangle \ = \ \kappa_P + t\,\Phi_{\mathcal{C}}(u) \ \xrightarrow[t\to\infty]{} \ \text{undercuts the target,}$$

a contradiction. Hence $\alpha \in P \cap \mathbb{Z}^{qn}_{\geq 0}$, and Lemma 20 makes it one-hot feasible.

(2) On the nonnegative cone evaluate at $t\mathbf{1}$, where the target has slope n:

$$\deg \alpha = d < n \quad \Longrightarrow \quad \kappa_P + td \ < \ tn = \Phi_{\mathcal{C}}(t\mathbf{1}) \quad \text{for large } t,$$

again a contradiction. For $\deg \alpha = n$ with $\alpha \notin P$, take a separating u and replace it by $u + c\mathbf{1}$ with c large. Since every exponent compared has degree exactly n, the shift adds cn to both sides and leaves the strict inequality intact while making the direction nonnegative, so (1) applies verbatim.

(3) Fix $f \in \mathcal{C}$ and set $u_{i,f_i} = 1$, $u_{i,a} = 2$ for $a \neq f_i$. Then

$$\langle u, \chi_f\rangle = n, \qquad \langle u, \alpha\rangle \geq n + 1 \quad \text{for every other one-hot feasible or higher-degree } \alpha.$$

Were χ_f absent from every parse, the circuit evaluated at tu would exceed $tn = \Phi_{\mathcal{C}}(tu)$ for large t. □

Call a node *live* if it occurs in at least one degree-n output parse; literals count as nodes, a live literal $u_{i,a}$ having scope $\{i\}$.

Lemma 22. *Assume no output parse has degree below n and every degree-n output parse is one-hot feasible. If a node g is live then its subtree degree d_g is independent of the occurrence, and there is a fixed scope $S(g) \subseteq X$ with $|S(g)| = d_g$. Let $A_g \subseteq \Sigma^{S(g)}$ be the partial assignments produced by degree-d_g parses rooted at g and $B_g \subseteq \Sigma^{X\setminus S(g)}$ the complementary assignments induced by output contexts containing g. Then $A_g \times B_g \subseteq \mathcal{C}$. At a live $\min$ gate the two child scopes are equal; at a live $+$ gate they are disjoint with union the parent scope.*

Proof. Degree. We make the replacement argument explicit, because it is the point at which sharing is handled. A complete parse occurrence of g is written as $K[Q]$, where Q is the parse subtree rooted at the original DAG node g, and $K[\cdot]$ is the rest of the unrolled parse tree. Since the circuit is acyclic, replacing Q by any other parse subtree rooted at the same DAG node produces another valid parse occurrence; this remains true even when the original node is reused in many different contexts.

Degree. Let $K_1[Q_1]$ and $K_2[Q_2]$ be two degree-n output parses using g, and let d_i be the degree of Q_i. Their context degrees are $n - d_i$. Cross-substitution produces valid output parses of degrees

$$n + d_1 - d_2 \quad \text{and} \quad n + d_2 - d_1, \qquad \text{both at least } n \text{ by hypothesis.}$$

Hence $d_1 = d_2 =: d_g$, independently of the occurrence or context.

Scope. In a degree-n one-hot parse $K_i[Q_i]$, let S_i be the set of coordinates used by Q_i, and let T_i be the set used by its context. Since Q_i has degree d_g and the complete parse is one-hot, $S_i \cap T_i = \varnothing$ and $S_i \cup T_i = X$. Now substitute Q_2 into K_1. The resulting parse still has degree n, so it is one-hot; its literal sets are S_2 and T_1. Thus $S_2 = X \setminus T_1 = S_1$. The same argument with the roles reversed shows that every occurrence of g has one fixed scope $S(g)$, and $|S(g)| = d_g$. This is the explicit sharing-invariance step: the node cannot acquire a different scope merely by being called from a different context.

Rectangle. Fix any context $K[\cdot]$ containing g, and any degree-d_g parse Q rooted at g. The substitution $K[Q]$ is a degree-n output parse and therefore a one-hot feasible word. Consequently, if $a \in A_g$ and $b \in B_g$, then

$$a \oplus b \in \mathcal{C}, \qquad \text{so} \qquad A_g \times B_g \subseteq \mathcal{C}.$$

Here B_g is well defined because the preceding scope argument shows that all contexts containing g use the same complementary coordinate set. At a live $\min$ gate, every selected child is substituted in the same parent context, so it has the parent scope; hence the two child scopes agree whenever both children are live. At a live $+$ gate both children occur in every parse through the gate, and one-hotness makes their scopes disjoint while their union is the parent scope. □

Lemma 23. *For a cut $X = L \dot\cup R$ put $Q_L = \mathcal{C}/(\mathcal{C}_L \oplus \mathcal{C}_R)$, so $|Q_L| = q^{\lambda_{\mathcal{C}}(L)}$. For $\theta \in Q_L$ let A_θ and B_θ be the two coordinate projections of the coset. Then*

$$\mathcal{C} = \dot{\bigcup}_{\theta \in Q_L} A_\theta \times B_\theta, \qquad |A_\theta||B_\theta| = \frac{|\mathcal{C}|}{q^{\lambda_{\mathcal{C}}(L)}},$$

and no rectangle contained in $\mathcal{C}$ meets two distinct cosets. Consequently the largest rectangle inside $\mathcal{C}$ has size $|\mathcal{C}|q^{-\lambda_{\mathcal{C}}(L)}$, and the number of distinct nonempty right languages across the cut is exactly $q^{\lambda_{\mathcal{C}}(L)}$ in either orientation.

Proof. Each coset is a rectangle. Let $(a, b_0), (a_0, b) \in \theta$. Then

$$(a, b) = (a, b_0) + (a_0, b) - (a_0, b_0) \quad \text{modulo } \mathcal{C}_L \oplus \mathcal{C}_R \quad \Longrightarrow \quad (a, b) \in \theta,$$

so $\theta = A_\theta \times B_\theta$. Distinct cosets have disjoint left and disjoint right projections, so the union is disjoint and no rectangle inside $\mathcal{C}$ can meet two of them without producing a word outside $\mathcal{C}$.

Sizes. By definition of λ,

$$|A_\theta||B_\theta| = |\mathcal{C}_L||\mathcal{C}_R| = \frac{|\mathcal{C}|}{q^{\lambda_{\mathcal{C}}(L)}} \qquad \text{for every } \theta \in Q_L, \qquad |Q_L| = q^{\lambda_{\mathcal{C}}(L)}.$$

Finally $a \equiv a'$ in the right-language sense exactly when a, a' lie in the same coset, so the count of nonempty right languages is $|Q_L|$ in either orientation. □

Proof of Theorem 13. The argument assigns to each codeword a node of the circuit, and bounds how much of $\mathcal{C}$ one node can account for. Fix $r \in \mathcal{C}$ and follow the chain

$$r \xrightarrow{\text{Lem. 21}} P_r \xrightarrow{\text{reduce}} T_r \xrightarrow{(23)} L_r \xrightarrow{\text{Lem. 22}} g(r),$$

whose four steps are as follows. P_r is a complete degree-n output parse with exponent χ_r. Deleting its constant leaves and suppressing degree-two vertices leaves exactly the n visible literals, since χ_r uses every coordinate once, so the result T_r is a subcubic branch tree on X. As ρ is a minimum

over branch trees and T_r is one of them, some edge of T_r induces a cut L_r with $\lambda_{\mathcal{C}}(L_r) \geq \rho$. Root T_r on the opposite side of this edge. The first unsuppressed parse node on the L_r-side is either a visible literal or an operation gate; call the corresponding original DAG node $g(r)$. Its parse subtree has exactly the visible leaves in L_r, so $S(g(r)) = L_r$. If the reduction suppressed a chain of unary selected $\min$-nodes, choose the first original node on that chain; the explicit scope invariance in Lemma 22 makes this choice independent of every other occurrence of the shared DAG node. Thus $g(r)$ is a counted literal or operation node, and in either case it contributes one node of C.

Lemmas 22 and 23 now bound the rectangle of $g(r)$, which contains r:

$$r \in A_{g(r)} \times B_{g(r)} \subseteq \mathcal{C}, \qquad \big|A_{g(r)}\big|\big|B_{g(r)}\big| \;\leq\; \frac{|\mathcal{C}|}{q^{\lambda_{\mathcal{C}}(L_r)}} \;\leq\; \frac{|\mathcal{C}|}{q^{\rho}}.$$

A node carries one scope and hence one rectangle, so $r \mapsto g(r)$ exhibits $\mathcal{C}$ as covered by at most $|C|$ rectangles of that size:

$$|\mathcal{C}| \;\leq\; \sum_{g \in \operatorname{im} g(\cdot)} \big|A_g\big|\big|B_g\big| \;\leq\; |C| \cdot \frac{|\mathcal{C}|}{q^{\rho}} \qquad \Longrightarrow \qquad |C| \;\geq\; q^{\rho(\mathcal{C})}.$$

Lemma 21 supplies the parse hypotheses in both score domains. □

C Upper Bound and the Grid Instance

Proof of Proposition 14. Take a branch tree attaining ρ, root it at an edge, and put $K := q^{\rho}$. On every oriented edge store one message per nonempty right-language class of the leaf set below it, of which Lemma 23 gives at most K, holding the least unary cost of a partial assignment in that class.

Right-language equivalence composes across disjoint scopes: writing $a \equiv_S a'$ for equality of right languages on S,

$$a_1 \equiv_{S_1} a_1', \quad a_2 \equiv_{S_2} a_2', \quad S_1 \cap S_2 = \varnothing \qquad \Longrightarrow \qquad a_1 \oplus a_2 \;\equiv_{S_1 \cup S_2}\; a_1' \oplus a_2',$$

so a binary join is well defined on classes: enumerate the at most K^2 child-state pairs, send each completable pair to its parent class, add the two messages and minimize within each class. Costs add up as

$$\underbrace{O(nq)}_{\text{leaves}} + \underbrace{O(n)}_{\text{joins}} \cdot \underbrace{O(K^2)}_{\text{per join}} \;=\; O\big(nq + nq^{2\rho}\big),$$

with the root edge combining compatible states from the two sides. State sets and transition tables are relation-dependent wiring and are not charged to circuit size. □

Remark 24. *Writing* bw *for the branchwidth of a code in the convention of Kashyap (2009), our ρ equals* $\mathrm{bw}(\mathcal{C}) - 1$. *We avoid the offset because it otherwise has to be carried through every exponent.*

Proof of Proposition 15. From codes to the cycle matroid. For $\mathcal{C} \leq \mathbb{F}_q^X$ with generator matrix G, the columns of G represent a matroid M whose connectivity function $\lambda_M(L) = r(L) + r(X \setminus L) - r(X)$ coincides with (22). Chaining self-duality of λ with cycle/cocycle duality of a connected graph, whose cocycle code is represented by the incidence matrix,

$$\lambda_{\mathcal{C}_G} \overset{\text{self-duality}}{=} \lambda_{(\mathcal{C}_G)^{\perp}} \;=\; \lambda_{\text{cocycle}(G)} \;=\; \lambda_{M(G)} \qquad \Longrightarrow \qquad \rho(\mathcal{C}_G) = \mathrm{bw}\big(M(G)\big) - 1,$$

the implication following by taking minima over branch trees.

Sandwiching the branchwidth. The two halves meet exactly:

$$\underbrace{k \overset{\text{tangle of order } k}{\leq} \mathrm{bw}\big(M(G_k)\big)}_{\text{Geelen et al. (2009), } k \geq 3} \qquad \text{and} \qquad \underbrace{\mathrm{bw}\big(M(G_k)\big) \;\leq\; k}_{\text{left-to-right sweep, all } \lambda \leq k-1},$$

the first using that the maximum order of a tangle in a matroid equals its branchwidth (Robertson & Seymour, 1991; Geelen et al., 2009). Hence $\mathrm{bw}(M(G_k)) = k$ and $\rho(\mathcal{C}_{G_k}) = k - 1$. With $n = 2k(k-1)$ this is $\Theta(\sqrt{n})$, and substituting into (25) gives (26). □

Remark 25. *Proposition 15 pins the* instance width *exactly and certifies the optimality of the sweep at $k - 1$. The circuit-size estimate in (25) remains within a constant factor in the exponent, and its upper-bound half is checked computationally for $k \leq 12$ in Appendix D.*

D Experimental Details

The six experiments in Section 5 use seed 20260830. Unless stated otherwise, the code is the binary cycle code of a grid, with parity-check columns given by the vertex-edge incidence matrix after one redundant row is deleted. Syndromes are carried as machine integers and all GF(2) membership tests are exact. The exact trellis is checked against exhaustive enumeration of all 2^n binary vectors for $n \le 12$, and against explicit codeword enumeration for larger grids, with agreement to 10^{-15}. All floating-point forward passes, distance/floor sweeps, subspace samples, and annealing updates use CUDA double precision; the scripts stop if CUDA is unavailable. CPU code is restricted to the independent integer construction and exhaustive reference checks.

E1. Decomposition search. This is an implementation check of Proposition 15. For $k = 3, \dots, 12$ we compute the interface dimension of thirteen edge orders: column, row and diagonal sweeps; a linearized nested dissection that recursively bisects the vertex box along its longer side and emits interior edges before separator edges; a greedy heuristic appending the edge that keeps the open-vertex set smallest; and eight uniform random permutations. Every order lands at or above $k-1$, with the two sweeps and the greedy heuristic attaining it, the linearized nested dissection well above it, and random orders reaching $\lambda = 101$ at $k = 12$. This is consistent with $\rho = k - 1$ (Figure 6). Profiles come from prefix and suffix GF(2) ranks, which requires $O(n)$ rank updates rather than $O(n)$ full recomputations. The implementation performs those updates in synchronized CUDA boolean bases; the recorded wall-clock timings are included in the CSV artifact. The explicit trellis is built for $k \le 9$ to confirm that its peak width equals 2^λ.

E2. Budget and width. The 5×5 grid has $n = 40$ edges, $\dim \mathcal{C} = 16$ and 65536 codewords, and the exact trellis has 16 states at its widest cut and 971 gates. Truncation retains the m states carrying the most codeword traffic at each cut, a purely structural rule that uses no training data. The global tropical constant is the median training residual, which is the L_1-optimal choice.

The held-out sample is 50000 Gaussian score vectors, and the combinatorial check below probes 1024 codewords per budget, where the discrepancy is 0 at every budget. δ is obtained by running the truncated circuit on the 65536 rays $u^{(c)}$ defined by $u^{(c)}_{i,c_i} = 0$ and $u^{(c)}_{i,1-c_i} = 1$, on which $\Phi_{\mathcal{C}} = 0$, and taking the largest value. Exact positive homogeneity makes regression against t redundant, so the independent check is computed directly from the codeword distances. The truncated circuit computes the minimum over the codewords whose syndrome path survives, so on $u^{(c)}$ its value must equal $\min_{c' \text{ surviving}} d_H(c, c')$, which we evaluate from the codeword set alone on 1024 codewords sampled per budget.

The floor is predicted as $\min_s \mathbb{E}|g - s|$ with g the shadow gap on the held-out sample, the minimizer being the median. Against it we minimize the circuit's mean relative error directly over the grid $t \in \{\frac{1}{4}, \frac{1}{2}, 1, 2, 4, 8, 16, 64\}$ and 121 calibration constants spanning $[-3|b| - 1,\ 1]$, evaluating the circuit at each pair.

Table 2: Truncating the exact trellis of the 5×5 grid cycle code. Both error quantities reach zero only at $m = 2^\rho$, the code's exact interface dimension. The last column brute-forces the circuit's mean relative error over eight scales and 121 calibrations, and never beats the predicted floor.

m	gates	surviving codewords	δ	$D_\nu(g)$ predicted	best over (t, b)
1	40	1	32	4.0180	4.0181
8	307	8	28	3.5815	3.5815
12	639	3948	16	1.6601	1.6602
15	888	39648	10	0.4221	0.4221
16	971	65536	0	0.0000	0.0000

E3. Finite-temperature operators. Over $\varepsilon \in [0.069, 1.386]$ the deficient softmin deformation error $\sup_u |C - \widetilde{C}|$ falls from 4.00 to 0.129, a 96.8% reduction, while its slope holds at 27.95 with spread 3.4×10^{-3}. The full-width control's final softmin slope decays to 3.8×10^{-5}. The deficient budget is $m = 8$, whose shadow gap is 28.

Deformations. $\mathrm{softmin}_\beta$ is evaluated as $\min(a,b) - \beta\log(1+e^{-|a-b|/\beta})$ for numerical stability, with $\varepsilon = \beta\log 2$. The squared-$\ell_2$ operator is $\min_\Omega(x) = -\max_\Omega(-x)$ with $\Omega(p) = \frac{\gamma}{2}\|p\|^2$, computed through the Euclidean projection onto the simplex. On the d-simplex $\min_\Delta \Omega = \gamma/(2d)$ and $\max_\Delta \Omega = \gamma/2$, so $\varepsilon = \gamma/2$ is valid. Gates at which one argument is $+\infty$, which is how truncation is implemented, degenerate to the finite branch and are not deformed.

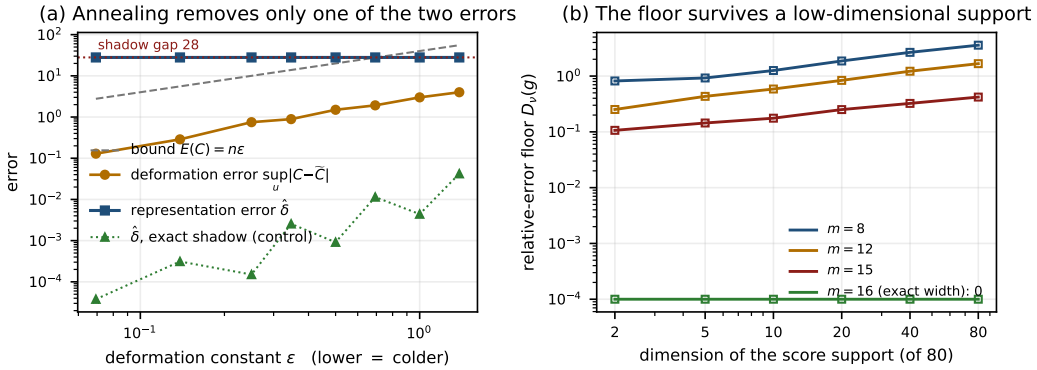

Figure 5: **The two supplementary sweeps.** (a) Freezing the deficient $m = 8$ wiring and sweeping ε over five softmin temperatures and three squared-ℓ_2 regularized operators. The deformation error obeys $E(C)$ and vanishes with ε; the representation error remains fixed. The control is the full-width trellis, whose shadow is exact. (b) The floor when the scores are confined to a nested chain of random subspaces of the 80-dimensional score space. Markers are the brute-force search over scales and calibrations. The exact-width circuit has floor exactly 0 at every rank and is drawn at the axis floor.

E4. Standard block codes. Hamming$(7,4)$, BCH$(15,7)$, and the extended Golay code G_{24} are used as independent checks because their dimensions, distances, and relevant trellis-width bounds are known. Hamming$(7,4)$ is built from its standard parity matrix; BCH$(15,7)$ is the cyclic code with generator $x^8 + x^7 + x^6 + x^4 + 1$; and G_{24} is obtained from the cyclic Golay code with generator $x^{11} + x^{10} + x^6 + x^5 + x^4 + x^2 + 1$, followed by an overall parity bit. Before the sweep, the artifact checks the divisibility by $x^{23} - 1$ and the weight enumerator $(1, 759, 2576, 759, 1)$. Coordinate orders are searched by pairwise-swap descent from 40 random restarts, and the budget grid includes both endpoints. Table 3 reports the result.

Table 3: The budget sweep on three standard block codes. Widths are the maximum cut interface dimension λ; for G_{24} the searched value matches the Muder lower bound (Muder, 1988). In each code the last nonzero slope equals the minimum distance.

code	n	$d_{\min}$	λ natural	λ searched	δ sequence	floor at full width
Hamming$(7,4)$	7	3	3	3	$7,4,3,0$	0
BCH$(15,7)$	15	5	7	5	$15,10,9,8,6,5,0$	0
Golay G_{24}	24	8	12	9 (Muder)	$24,16,12,8,0$	0

E5. Trained annealing. The wiring is the $m = 8$ truncation of Section 5.2, held fixed throughout. The trainable parameters are one softmin temperature per gate layer, written $\beta_i = \beta_{\mathrm{cap}}(\mathrm{step}) \cdot \sigma(\theta_i)$ so that the schedule is an upper bound the learner may undercut, and one additive constant leaf per layer; $2n = 80$ parameters in all. The schedule is geometric, $\beta_{\mathrm{cap}}\colon 2 \to 0.05$ over 3000 standard CUDA-Adam steps at learning rate 3×10^{-2} and batch 1024, against the exact layer on the same 50000 Gaussian training vectors. Backpropagation through the trellis uses PyTorch autograd in CUDA double precision; with $s = \mathrm{softmin}_\beta(a,b)$ and w the softmin weights, $\partial s/\partial a = \sigma((b-a)/\beta)$ and $\partial s/\partial\beta = (s - \langle w, (a,b)\rangle)/\beta$, and a gate with one infinite argument degenerates to the finite branch and is not deformed.

Two deviations from the shadow are reported. The whole gap must obey $E(C) = \sum_i(\varepsilon_i + |c_i|)$, which is the recursion of (9) on this wiring; the trained constants keep it away from zero, since

Definition 3 zeroes every constant. Setting the constants to zero isolates the deformation error, and that is the quantity the schedule drives down, bounded by $\sum_i \varepsilon_i$. The slope is read off the worst adversarial ray at $t = 256$, where Lemma 4 still leaves a residual of order $E(C)/t$, so it approaches 28 from below rather than attaining it.

E6. Score supports. A fixed random orthonormal frame of $\mathbb{R}^{80}$ is drawn once; its first r columns span the rank-r subspace, so the chain of supports is nested and every direction available at rank r survives at every larger rank. Scores are Az with z standard normal on $\mathbb{R}^r$, which at $r = 80$ is exactly the isotropic Gaussian used elsewhere, up to the draw. The floor is predicted and brute-forced by the same two routines as Section 5.2, on 50000 samples per rank, and Table 4 reports both.

Table 4: The relative-error floor $D_\nu(g)$ when the scores are confined to a nested chain of subspaces of the 80-dimensional score space. Brute-force values, searched over eight scales and 121 calibrations, are in parentheses. The exact-width circuit $m = 16$ has floor 0 at every rank.

support dimension	80	20	5	2
$m = 8$	3.577 (3.577)	1.864 (1.864)	0.924 (0.924)	0.817 (0.817)
$m = 12$	1.669 (1.669)	0.837 (0.837)	0.433 (0.433)	0.252 (0.252)
$m = 15$	0.420 (0.420)	0.251 (0.251)	0.144 (0.144)	0.106 (0.106)

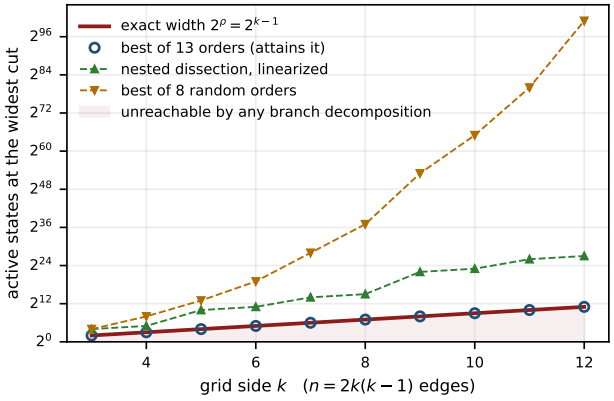

Figure 6: **Implementation check on the decomposition search.** Active states at the widest cut for grid cycle codes, over thirteen edge orders per grid. Proposition 15 gives $\rho = k - 1$, so the sweeps and the greedy heuristic are optimal and nothing can enter the region below them; the search confirms this numerically rather than establishing it.