FORMAL VERIFICATION OF KOLMOGOROV-ARNOLD NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Motivated by the Kolmogorov-Arnold representation theorem, Kolmogorov-Arnold Networks (KANs) are a novel neural network architecture that replace the weight matrices and fixed activation functions of a multi-layer perceptron with a learned univariate spline on every edge and an unweighted sum on every node. They have quickly become popular, being utilized as scientific surrogates, safe controllers and barrier certificates, but attempts to verify them are limited. Mainstream neural network verification tools have no spline operator, and conversion from a KAN with a grid of G cells to a ReLU network inflates the network by a factor of G. Therefore, this work develops formal verification utilizing set-based reachability for KANs by directly considering its unique architecture. We show that degree-1 KANs without the base branch admit exact verification, and that KANs of degree ≥ 2 or with the base branch admit sound verification. We compare head-to-head a method family instantiated on the KAN structure; using interval bound propagation, CROWN-style bounds, mixed-integer linear programs (MILPs), and star-set reachability. We demonstrate formal verification of five classes of properties for trained KANs, namely robustness, output ranges, barrier conditions, monotone ordering and one-step reachability; including *certified robustness* for an adversarially attacked 784-input MNIST KAN and all three conditions of a control barrier certificate whose barrier and controller are both KANs. Our literature-based case studies collectively can be considered a small KAN verification benchmark, *e.g.* for VNN-COMP. `https://anonymous.4open.science/r/KAN-verification`

1 INTRODUCTION

Kolmogorov–Arnold Networks (KANs) (Liu et al., 2025b) invert the layer structure of the multilayer perceptron (MLP). Where an MLP layer computes a learned affine map, then a fixed elementwise nonlinearity, a KAN layer places a *learned univariate function on every edge* and computes

$$\overbrace{z_q}^{\textbf{node } q} = \sum_{p=1}^{n} \overbrace{\phi_{q,p}(x_p)}^{\textbf{edge } (q,p)}, \qquad \phi_{q,p}(x) = w_{q,p}\,\mathrm{silu}(x) + \sum_i c_{q,p,i}\,B_{i,k}(x), \tag{1}$$

in which node q sums into its value z_q the outputs $\phi_{q,p}(x_p)$ of its n incoming edges, and every edge is a B-spline of degree k with learned coefficients $c_{q,p,i}$ on fixed basis functions $B_{i,k}$ plus a smooth *base branch* as $w_{q,p}\mathrm{silu}(x)$, where $\mathrm{silu}(x) = x/(1+e^{-x})$. A KAN is *no-base* if all $w_{q,p} = 0$. The $[2, 3, 1]$ KAN of Figure 1 stacks two such layers into the network N, with output $N(\boldsymbol{x})$ on input $\boldsymbol{x}$.

KANs are motivated by the Kolmogorov–Arnold representation theorem (Kolmogorov, 1956; Givental et al., 2009) and were introduced for scientific machine learning (Liu et al., 2025b;a), where they promise accuracy on smooth compositional targets and interpretable edges. The (growing) literature uses KANs as neural-ODE vector fields (Koenig et al., 2024), physics-informed PDE surrogates (Shukla et al., 2024), time-series forecasters (Vaca-Rubio et al., 2024), reinforcement-learning policies (Kich et al., 2024), medical image segmenters (Li et al., 2025), and, most relevant here, barrier certificates and safe controllers for cyber-physical systems (Zhang et al., 2025b;a). Comparisons with MLPs are contested, with KANs outperforming mainly on symbolic-structure tasks (Yu et al., 2024; Hou et al., 2024). However, it is precisely in this niche of compact models

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.

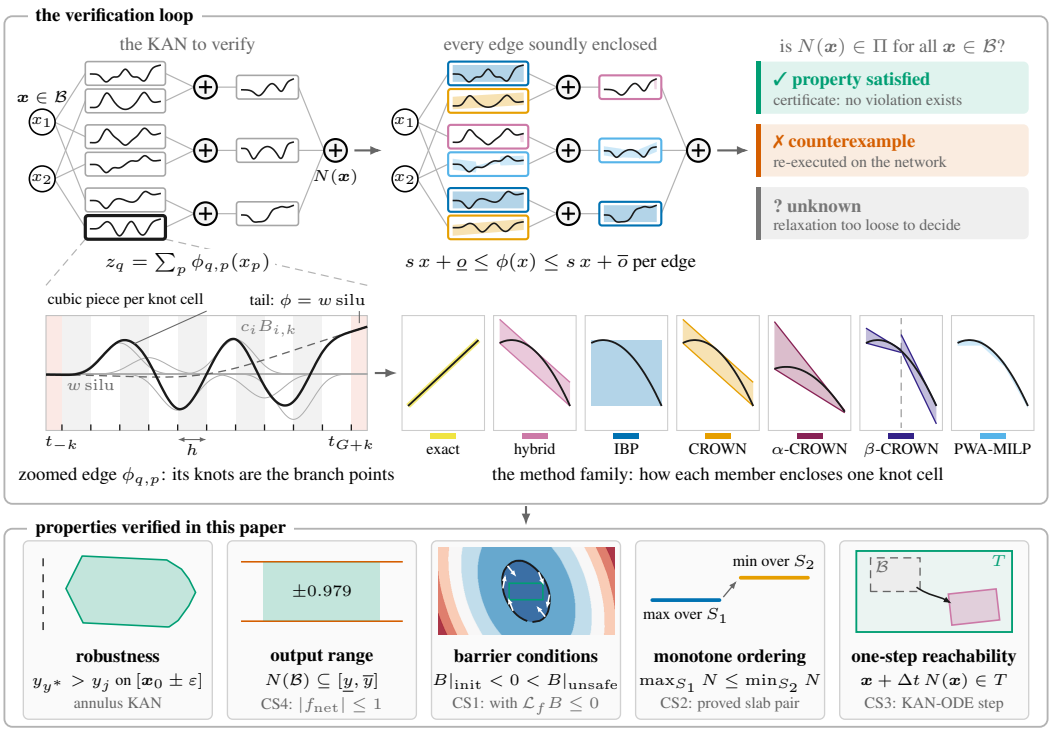


Figure 1: **The verification loop.** *Top:* a $[2, 3, 1]$ KAN N (learned edges $\phi_{q,p}$, nodes that only sum) becomes the same network with every edge soundly enclosed. Box borders take the color of the member supplying the enclosure. The question is whether $N(x) \in \Pi$ for every x in the box $\mathcal{B}$, for a linear output property Π; the answer is a certificate, a re-executed counterexample, or UNKNOWN. *Middle:* the zoomed edge (black box) with its knots and knot cells, and how each member of the method family encloses one knot cell. *Bottom:* the property classes verified, each tagged with its case study.

of physical systems, controllers, and certificates that formal guarantees are required. Usefully, the structure of these guarantees generalize to showing $N(x)$ satisfies a property Π for every input x in a box $\mathcal{B}$.

Verification tooling for this need is limited. Most neural network verification, from complete solvers (Katz et al., 2017) through bound propagation (Zhang et al., 2018) to set-based reachability with *star sets* (affine images of polytopes (Tran et al., 2019)), is built around layers that factor as an affine map followed by an elementwise nonlinearity, **which a KAN layer is not**. No mainstream verifier has a native B-spline operator, and no KAN benchmark has appeared in any VNN-COMP edition (Brix et al., 2024; Kaulen et al., 2025). Degree-1 no-base KANs convert exactly into ReLU networks (Schoots et al., 2025), but for higher degrees the conversion of (Wang et al., 2025) uses the $\max(0, x)^k$ activation function, requires no base branches, and has roughly G times as many parameters for a grid of G knot cells. Karia et al. (2025) assume KAN edges stay inside the training grid (not guaranteed in practice), and other treatments do not verify the trained KAN itself.

This paper develops KAN verification from the architecture up rather than from any single technique: we record which parts of MLP verification carry over to KANs and what the KAN layer changes. We instantiate a *family* of methods on that structure, and compare its members on a small benchmark of 26 trained KANs, which can be incorporated into VNN-COMP.

Contributions.

- **A family of verification methods for KANs.** IBP, CROWN-style bounds (with and without optimized slopes), branch and bound, MILP encodings, and star-set reachability all are adapted

and applied to verify KANs. Table 2 compares the members on the same queries and certified edge ranges, and on our benchmark the branch-and-bound member performs best.

- **The defense counterpart to the KAN adversarial literature**. We certify robustness for fully-connected KANs with the layer parameterization attacked by Alter et al. (2025), retrained by us. We verify 784-input MNIST, Fashion-MNIST and depth-2 MNIST KANs whose certified fractions match the PGD ceiling (Madry et al., 2018) at all ten radii.
- **KAN Verification Benchmark.** We analyze literature case studies as *trained KANs*, including the KAN4CBC pendulum barrier and controller networks (Zhang et al., 2025a), whose original pipeline certifies only a symbolic surrogate, and certify the trained degree-3 barrier and controller on all three barrier conditions. We also measure the accuracy and robustness costs of the exactly-verifiable regime, comparing them with the equivalent ReLU-MLP conversion, and provide, to our knowledge, the first benchmark for KAN verification with 26 networks.

2 Related Work

KANs and their variants. The original KAN (Liu et al., 2025b) and KAN 2.0 (Liu et al., 2025a) use B-spline edges with a base branch. Variants swap the basis for RBFs (Li, 2024), Chebyshev (SS et al., 2024), wavelets (Bozorgasl & Chen, 2024), ReLU-product bumps (Qiu et al., 2025) or grouped rationals (Yang & Wang, 2025), and each changes the verification problem qualitatively. Comparisons temper the original claims (Yu et al., 2024; Hou et al., 2024; Shukla et al., 2024).

Neural network verification. Complete methods (SMT, MILP, and branch-and-bound over bound propagation) (Katz et al., 2017; 2019; Tjeng et al., 2019; Xu et al., 2021; Wang et al., 2021) and scalable incomplete methods (bound propagation, abstract interpretation) (Zhang et al., 2018; Xu et al., 2020; Gehr et al., 2018; Singh et al., 2018; 2019; Wang et al., 2018) target networks of affine layers and fixed activations. Branch-and-bound over general nonlinearities (Shi et al., 2025) comes closest to KANs, but it has no dedicated B-spline operator and pre-optimizes its branch points offline for each nonlinearity. Star sets, introduced for linear-system reachability (Bak & Duggirala, 2017), support exact and over-approximate ReLU reachability with LP-tight bounds (Tran et al., 2019; 2021; Lopez et al., 2023), and have been extended to CNNs (Tran et al., 2020), probabilistic verification (Tran et al., 2023), sparse image representations (Tran et al., 2025) and general piecewise-linear activations (Antal et al., 2025), but none handle B-spline edges.

Verifying KANs. Prior efforts to verify KANs exist, but to our knowledge none adapts linear-bound propagation, branch and bound or star sets to the KAN layer. Schwartz et al. (2026) abstract each edge by an optimally allocated piecewise-affine over-approximation with certified error and discharge range properties by MILP or SMT, whereas our complete members are exact for degree-1 no-base KANs. KAN4BC (Zhang et al., 2025b) synthesizes barrier certificates represented by Fourier KANs and verifies them with Lipschitz bounds and the SMT solver dReal. KAN4CBC (Zhang et al., 2025a) synthesizes B-spline KAN barrier certificates and controllers, but its verification step checks a pruned, symbolified surrogate with dReal rather than the trained KAN. Alter et al. (2025) *attack* KANs and name certified defenses as future work; we supply the defense counterpart for KANs of the same layer parameterization. Concurrently, Meymani & Razavi-Far (2026) benchmark vision KANs with randomized-smoothing certificates and interval-bound certified training, leaving verification to future work. We *verify* trained B-spline KANs post hoc against any linear property, soundly and, in the degree-1 no-base regime, completely.

3 Preliminaries

KANs. A KAN layer $L : \mathbb{R}^n \to \mathbb{R}^m$ (n inputs, m outputs) computes Equation 1: the values x_p of its n input nodes, $p \in \{1, \dots, n\}$, feed its m output nodes z_q, $q \in \{1, \dots, m\}$, through the *edges* $\phi_{q,p}$ (the learned activation functions of Liu et al. (2025b)); node q passes z_q to the next layer. The *spline part* of an edge is a B-spline of degree k with learned coefficients $c_{q,p,i}$ and fixed basis functions $B_{i,k}$, $i = -k, \dots, G-1$ (de Boor, 1978), on a per-layer grid $[a, b]$ of G intervals of width $h = (b-a)/G$. The *knots* of this basis are the grid points $t_j = a + jh$, $j = -k, \dots, G+k$, and the *knot cells* are the $G + 2k$ intervals $[t_j, t_{j+1}]$ between consecutive knots: the G of the grid and k more on each side. We call $[t_{-k}, t_{G+k}]$ the *extended grid*, and we write $\mathcal{P}$ for the *extended partition* of $\mathbb{R}$ into the knot cells and the two *tails* $(-\infty, t_{-k}]$ and $[t_{G+k}, \infty)$; its *pieces* are closed,

so neighboring pieces share an endpoint. A KAN is a composition $N = L_D \circ \cdots \circ L_1$ of D layers and is named by its layer widths. We assume $k \geq 1$ throughout, since piecewise-constant edges are discontinuous. The two common regimes are *degree-1 no-base* ($k = 1$ and $w \equiv 0$; piecewise linear) and *degree-3+SiLU* ($k = 3$ with the base branch), *degree 1* and *degree 3* for short. Each $B_{i,k}$ is a piecewise polynomial on $k+1$ knot cells, so the spline part is one polynomial per knot cell and zero beyond the extended grid. The following lemma, which adapts standard B-spline theory, provides a foundation for cells, knots, and tails utilized later in the manuscript.

Lemma 1 (KANs are piecewise with zero tails)**.** *Let $\phi = \sum_{i=-k}^{G-1} c_i B_{i,k}$ be the spline part of an edge, $k \geq 1$. Then ϕ is continuous, on every knot cell $[t_j, t_{j+1}]$ it coincides with a polynomial of degree at most k, and on the tails $(-\infty, t_{-k}]$ and $[t_{G+k}, \infty)$ it is identically zero. Consequently, with $k = 1$ and $w \equiv 0$ every edge, and hence every layer, is continuous piecewise affine.*

Star sets. A star (Bak & Duggirala, 2017; Tran et al., 2019) is $\Theta = \langle \boldsymbol{c}, \boldsymbol{V}, P \rangle = \{\boldsymbol{c} + \boldsymbol{V}\boldsymbol{\alpha} : \boldsymbol{\alpha} \in P\}$ with center $\boldsymbol{c}$, generator matrix $\boldsymbol{V}$, *predicate variables* $\boldsymbol{\alpha}$ and *predicate polytope* $P = \{\boldsymbol{\alpha} : \boldsymbol{C}\boldsymbol{\alpha} \leq \boldsymbol{d}\}$.Linear programs (LPs) can provide bounds, emptiness and linear optimization over Θ. Stars are closed under affine maps exactly and without predicate growth. Intersecting a star with a halfspace is exact and appends one row to its predicate. A box $\mathcal{B}$ is itself the star $\langle \boldsymbol{0}, \boldsymbol{I}, \mathcal{B} \rangle$, where $\boldsymbol{I}$ is the identity matrix.

The verification problem. Given a KAN N, a box $\mathcal{B} = [\underline{\boldsymbol{x}}, \overline{\boldsymbol{x}}]$ and a linear property $\Pi = \{\boldsymbol{y} : \boldsymbol{H}\boldsymbol{y} \leq \boldsymbol{g}\}$ (rows may be strict), we decide whether $N(\mathcal{B}) \subseteq \Pi$. Local robustness at $(\boldsymbol{x}_0, \varepsilon)$ with label y^* uses $\mathcal{B} = [\boldsymbol{x}_0 \pm \varepsilon]$ and $\Pi = \bigcap_{j \neq y^*} \{\boldsymbol{y} : y_{y^*} > y_j\}$, i.e. positive *margins* $y_{y^*} - y_j$. A verifier returns one of three verdicts (Figure 1). ROBUST certifies that the property holds on $\mathcal{B}$. FALSIFIED returns a *counterexample* (a witness), an input that violates the property, confirmed by re-executing the network, or an explicit no-counterexample marker. For the star-set member, UNKNOWN requires a relaxed layer or an LP that ends neither optimal nor proved infeasible. The five property classes of Figure 1 all reduce to this problem and differ in the input box, the rows of Π and the map that is checked. An output range uses a box $\Pi = [\underline{\boldsymbol{y}}, \overline{\boldsymbol{y}}]$, either given or computed as a certified enclosure, and the two sign conditions of a barrier B, namely $B < 0$ on the initial set and $B > 0$ on the unsafe region, are one-row range queries, the latter over a box cover. One-step reachability checks the Euler map $\boldsymbol{x} \mapsto \boldsymbol{x} + \Delta t\, N(\boldsymbol{x})$ with time step Δt, an affine map of $(\boldsymbol{x}, N(\boldsymbol{x}))$ that stars carry exactly, against a target box T. Monotone ordering compares the maximum of N over an input slab S_1 with its minimum over a second slab S_2. The third barrier condition (on the Lie derivative), is a little more complex; requiring the derivative network $N'(\boldsymbol{x})$.

4 HOW KAN VERIFICATION RELATES TO MLP VERIFICATION

What carries over. The verification problem is the same as for an MLP, and every *set representation* (the object propagated through the network) used for MLPs, whether intervals, linear bounds, MILP encodings or star sets, applies once each edge is enclosed. The degree-1 no-base regime is closest to the ReLU case: such a KAN is continuous piecewise affine (Lemma 1), and degree-1 KANs and ReLU networks convert into each other with polynomial size (Schoots et al., 2025; Wang et al., 2025). Therefore the restriction loses no worst-case expressiveness, and deciding whether such a KAN violates a linear property on a box is NP-complete, as it is for ReLU networks (Katz et al., 2017).

What changes. A KAN layer applies nm nonlinearities first and a fixed unweighted sum after (with 0/1 weights), which is the reverse of an MLP layer. Four differences matter. *First*, a relaxation needs one variable per edge, nm per layer instead of m, but all m edges leaving node p read the same x_p, so a layer's case splits form a product of n knot partitions, independent of the output width; whereas exact ReLU analysis splits per output neuron at learned hyperplanes. *Second*, the natural branch points are the knots, known in advance, so no offline branch-point pre-optimization (Shi et al., 2025) is needed. *Third*, edges are piecewise polynomial: for $k{=}1$ every knot cell is affine and exact analysis is possible, whereas a $k{\geq}3$ knot cell can inflect (Figure 6, Appendix C), like sigmoid and tanh, so no ReLU-style closed-form optimal relaxation exists, and the base branch forces relaxation even at $k{=}1$. *Fourth*, the spline part of every edge vanishes outside the extended grid, a soundness trap with no MLP analogue: extrapolating the boundary knot cell's polynomial is unsound, clamping inputs to the grid verifies a different network, and to our knowledge no prior work addresses this.

5 A METHOD FAMILY FOR KAN VERIFICATION

We present the family through its star-set member (Algorithm 1, Appendix A), whose stars carry the knot-cell splits and edge relaxations that the other members also use; the details of each method are in their own paper (Table 1), and we state only what changes for KAN layers. The star-set member has three modes, `exact`, `hybrid` and `approx`, each presented below with their underlying theory.

5.1 EXACT MODE FOR DEGREE-1 NO-BASE KANS

Exact mode adapts star-set reachability (Tran et al., 2019) to KANs: it splits each layer input at the knots and never relaxes, so it returns the exact reachable set, verification is complete (never UNKNOWN), and every negative margin yields a counterexample. For a *split cap* $S_{\max}$, it stops with an error if one star splits into at least $S_{\max}$ stars before the last node of the layer is split. Each edge reads one input node and changes formula only at the knots, so once each input node of a star lies in one knot cell, every edge is one polynomial there.

Lemma 2 (Knot-cell case splits)**.** *Let L be a KAN layer with extended partition $\mathcal{P}$, let $\Theta \subseteq \mathbb{R}^n$ be a nonempty bounded star, let $[l_p, u_p]$ be the range of x_p over Θ, let $R_p \subseteq \mathcal{P}$ be the set of pieces that meet $[l_p, u_p]$, and for a* piece assignment $J \in R_1 \times \cdots \times R_n$ *let $\Theta_J = \Theta \cap \{\boldsymbol{x} : x_p \in J_p \text{ for all } p\}$ be the* region star *of J. Then (i) $\bigcup_J \Theta_J = \Theta$; (ii) each Θ_J is a star whose predicate has at most $2n$ rows more than that of Θ; (iii) for all p and q, $\phi_{q,p} = w_{q,p}\mathrm{silu} + \rho_{q,p}$ on J_p for a polynomial $\rho_{q,p}$ of degree at most k; and (iv) the number of piece assignments is $\prod_p r_p$ with $r_p = |R_p| \le G + 2k + 2$, which does not depend on the output width m.*

Lemma 2 gives the region stars of Theorem 1 and the splits of exact and hybrid modes, and by part (iv) the number of cases stays flat as the output width m grows, as the experiments confirm. Without the base branch, a degree-1 edge is affine on each knot cell, so on each region star the whole layer is one affine map, which a star represents without loss.

Theorem 1 (Exact degree-1 reachability)**.** *Let $N = L_D \circ \cdots \circ L_1$ be a degree-1 no-base KAN and $\mathcal{B}$ a box. Let $\mathcal{F}_0$ be the one-entry list $(\langle \mathbf{0}, \boldsymbol{I}, \mathcal{B}\rangle)$, and, for $d = 1, \ldots, D$, let $\mathcal{F}_d$ be the list, with one entry per pair (Θ, J), of the stars $L_d(\Theta_J)$ over all entries Θ of $\mathcal{F}_{d-1}$ and all nonempty region stars Θ_J of Θ for L_d. Then every entry of $\mathcal{F}_D$ is a star $\Omega_i = \langle \boldsymbol{c}_i, \boldsymbol{V}_i, P_i\rangle$ with $P_i \subseteq \mathcal{B}$ and $N(\boldsymbol{\alpha}) = \boldsymbol{c}_i + \boldsymbol{V}_i\boldsymbol{\alpha}$ for all $\boldsymbol{\alpha} \in P_i$, the polytopes P_i cover $\mathcal{B}$, and $\bigcup_i \Omega_i = N(\mathcal{B})$.*

Exact mode computes $\mathcal{F}_D$, so Theorem 1 is its guarantee; its output stars keep the input coordinates as their predicate variables, so the minimizer of an LP over an output star is a point $\boldsymbol{\alpha}^* \in \mathcal{B}$, and a negative margin there is a counterexample.

Corollary 1 (Decision procedure with counterexamples)**.** *Let N, $\mathcal{B}$ and $\Omega_1, \ldots, \Omega_r$ be as in Theorem 1, let $\Pi = \bigcap_{j \ne y^*}\{\boldsymbol{y} : y_{y^*} > y_j\}$ be a robustness property with C classes, and for $j \ne y^*$ let $\mu_{ij} = \min_{\boldsymbol{\alpha} \in P_i}\left[(\boldsymbol{c}_i + \boldsymbol{V}_i\boldsymbol{\alpha})_{y^*} - (\boldsymbol{c}_i + \boldsymbol{V}_i\boldsymbol{\alpha})_j\right]$, an LP. Then (i) $N(\mathcal{B}) \subseteq \Pi$ if and only if all $r(C-1)$ values μ_{ij} are positive; (ii) if $\mu_{ij} < 0$, every minimizer $\boldsymbol{\alpha}^*$ lies in $\mathcal{B}$ and satisfies $N(\boldsymbol{\alpha}^*)_j > N(\boldsymbol{\alpha}^*)_{y^*}$; (iii) if $\min_{i,j} \mu_{ij} = 0$, then $N(\mathcal{B}) \not\subseteq \Pi$ while $N(\boldsymbol{x})_{y^*} \ge N(\boldsymbol{x})_j$ for all $\boldsymbol{x} \in \mathcal{B}$ and $j \ne y^*$.*

Exact verdicts and re-executed counterexamples instantiate Corollary 1. The problem stays hard: deciding violation is NP-complete, as it is for ReLU networks (Katz et al., 2017).

5.2 HYBRID AND APPROX MODES FOR DEGREE ≥ 2 OR THE BASE BRANCH

Hybrid mode splits at the knots and then relaxes each edge within its knot cell; approx mode relaxes each edge over the whole range of its input node and never splits. Both soundly verify standard KANs, of degree ≥ 2 or with the base branch, and hybrid mode trades splits for tightness. Inside one knot cell $[l, u]$ an edge is $\phi(x) = w\,\mathrm{silu}(x) + \rho(x)$ with ρ a polynomial of degree at most k (Lemma 2), and the family adapts the two-line enclosure of MLP verifiers to it as the *parallel relaxation* (Figure 1, middle), which bounds ϕ on $[l, u]$ between two lines of a common slope s, by default the *chord slope* $s_c = (\phi(u) - \phi(l))/(u - l)$. For continuous ϕ and $s \in \mathbb{R}$, the *offsets* are $\underline{o}(s) = \min_{x \in [l,u]}(\phi(x) - sx)$ and $\overline{o}(s) = \max_{x \in [l,u]}(\phi(x) - sx)$, and the *gap* is $\mathrm{gap}(s) = \overline{o}(s) - \underline{o}(s)$. Two parallel lines enclose the edge with a single error number, the gap, and for a fixed slope the tightest lines touch the highest and lowest points of the residual $\phi(x) - sx$.

Lemma 3 (Tightest parallel bounds). *Let ϕ be continuous on $[l, u]$ with $l < u$, with chord slope s_c and with the offsets $\underline{o}$, $\overline{o}$ and the gap defined above. Then (i) for every $s \in \mathbb{R}$, $sx + \underline{o}(s) \le \phi(x) \le sx + \overline{o}(s)$ for all $x \in [l, u]$, and any reals $o_1 \le o_2$ with $sx + o_1 \le \phi(x) \le sx + o_2$ on $[l, u]$ satisfy $o_1 \le \underline{o}(s)$ and $o_2 \ge \overline{o}(s)$; (ii) if ϕ is affine on $[l, u]$ with slope s, then $\mathrm{gap}(s) = 0$; (iii) if ϕ is convex or concave on $[l, u]$, then $\mathrm{gap}(s_c) = \min_{s\in\mathbb{R}} \mathrm{gap}(s)$.*

Part (i) certifies the offsets used by Theorem 2, part (ii) makes the relaxation exact on degree-1 no-base pieces, and part (iii) justifies the chord-slope default, optimal on degree-2 no-base knot cells, which are convex or concave; elsewhere convexity affects optimality but never soundness (Figure 6, Appendix C). The gap measures how far the edge bends over the interval, which the second derivative controls, so it shrinks with the square of the interval's width.

Lemma 4 (Quadratic relaxation gap). *Let ϕ be continuously differentiable on $[l, u]$, with an absolutely continuous derivative and $|\phi''| \le M_2$ almost everywhere on $[l, u]$, and let $u - l \le h$. Then the gap defined above satisfies $\mathrm{gap}(s_c) \le M_2 h^2/8$, with equality for $\phi(x) = M_2 x^2/2$ on $[l, u] = [-h/2, h/2]$.*

Lemma 4 is why hybrid mode splits before it relaxes: within a knot cell the gap is $O(h^2)$, whereas over a reachable range of width W approx mode can incur $M_2 W^2/8$, a worst case larger by the factor $(W/h)^2$. The offsets are certified by *Lipschitz sampling*, the extremum over uniform samples with a Lipschitz margin, with Lipschitz constants from interval arithmetic on the polynomial coefficients plus $1.1\,|w|$ for the base branch, since $\sup_{x\in\mathbb{R}} |\mathrm{silu}'(x)| < 1.1$. The relaxed layer then emits a *lifted star* with one fresh predicate variable $\delta_{q,p}$ per edge: for a star Θ with per-edge slopes $s_{q,p}$, offsets $[\underline{o}_{q,p}, \overline{o}_{q,p}]$ and certified ranges $[\underline{\phi}_{q,p}, \overline{\phi}_{q,p}]$, it is $\Omega(\Theta) = \{\boldsymbol{S}\boldsymbol{x} + \boldsymbol{E}\boldsymbol{\delta} : \boldsymbol{x} \in \Theta,\ \underline{o}_{q,p} \le \delta_{q,p} \le \overline{o}_{q,p},\ \underline{\phi}_{q,p} \le s_{q,p}x_p + \delta_{q,p} \le \overline{\phi}_{q,p}\ \forall q, p\}$ with $\boldsymbol{S} = (s_{q,p})$ and $(\boldsymbol{E}\boldsymbol{\delta})_q = \sum_p \delta_{q,p}$, a star with center $\boldsymbol{S}\boldsymbol{c}$, generator matrix $[\boldsymbol{S}\boldsymbol{V}\ \boldsymbol{E}]$ and predicate variables $(\boldsymbol{\alpha}, \boldsymbol{\delta})$, whose range constraints are the *tighten rows*. In hybrid and approx modes a piece assignment may also leave an input node unsplit, $J_p = [l_p, u_p]$ (every node in approx mode, and in hybrid mode the nodes left when the split cap $S_{\max}$ is reached, which is recorded); Lemma 2 applies unchanged, so these region stars still cover Θ with at most $2n$ rows each.

Theorem 2 (Soundness of the lifted star). *Let L be a KAN layer, $\Theta \subseteq [l_1, u_1] \times \cdots \times [l_n, u_n]$ a star, and $s_{q,p} \in \mathbb{R}$, $\underline{o}_{q,p} \le \overline{o}_{q,p}$, $\underline{\phi}_{q,p} \le \overline{\phi}_{q,p}$ reals such that for every edge (q, p) of L and every $x \in [l_p, u_p]$, $\phi_{q,p}(x) - s_{q,p}x \in [\underline{o}_{q,p}, \overline{o}_{q,p}]$ and $\phi_{q,p}(x) \in [\underline{\phi}_{q,p}, \overline{\phi}_{q,p}]$. Then $L(\Theta) \subseteq \Omega(\Theta)$. Consequently, if $\mathcal{F}_0 = \{\langle \mathbf{0}, \boldsymbol{I}, \mathcal{B}\rangle\}$ and $\mathcal{F}_d$ replaces each $\Theta \in \mathcal{F}_{d-1}$ by the stars $\Omega(\Theta_J)$ under L_d, with offsets and ranges as above, or $L_d(\Theta_J)$ when every edge of L_d is affine on the ranges of Θ_J, for finitely many piece assignments J whose region stars cover Θ, then $N(\mathcal{B}) \subseteq \bigcup_{\Omega\in\mathcal{F}_D} \Omega$, and, for every property Π, if $\Omega \subseteq \Pi$ for every $\Omega \in \mathcal{F}_D$, then $N(\mathcal{B}) \subseteq \Pi$.*

Every hybrid and approx verdict rests on Theorem 2. Theorems 1 and 2 cover boxes beyond the extended grid because $\mathcal{P}$ includes the zero tails; extrapolating a boundary knot cell's polynomial is unsound.

Proposition 1 (Unsound extrapolation). *There exist a degree-1 no-base KAN N, a box $\mathcal{B}$ and a property Π with $N(\mathcal{B}) \cap \Pi = \emptyset$ such that the network $\tilde{N}$ obtained from N by replacing, on every edge, the zero tails by the affine continuations of the adjacent knot cells satisfies $\tilde{N}(\mathcal{B}) \subseteq \Pi$.*

Proposition 1 is why $\mathcal{P}$ keeps the tails; the experiments reproduce its false verdict.

5.3 The method family

Table 1 lists the members, which all use the same certified edge ranges. Star-set reachability (Tran et al., 2019), in the modes `exact`, `hybrid` and `approx`, keeps the dependence between the nodes in the predicate of each star and is the member that returns reachable sets. Interval bound propagation (IBP) (Gowal et al., 2019) bounds every node by the sum of its edge ranges, and its final-layer elision bounds each margin of the last layer directly, which is tighter because both outputs read the same inputs. CROWN (Zhang et al., 2018) composes the parallel relaxation of Lemma 3 backwards through the network without solving an LP. α-CROWN (Xu et al., 2021) optimizes the slopes of the two lines of every edge and re-certifies their offsets, and β-CROWN (Wang et al., 2021) adds branch

Table 1: The method family, with the labels of Tables 2: exact, hybrid and approx are the star-set modes, IBP interval bound propagation (with final-layer elision on robustness margins), CROWN the CROWN-style bounds, α-CROWN and β-CROWN our KAN implementations of these methods, MILP the exact MILP, and the ReLU rows the released tools on the ReLU conversion. All members are sound. *Exact if unlimited branching/splitting. †Only as $K \to \infty$. §On the ReLU conversion, exact at degree 1 and a sound enclosure at degree 3 (Appendix C.9). In the counterex. column, ✓ means the member returns a candidate counterexample, which counts once re-executed on the network.

member	builds on	complete	counterex.	returns	degree 3	cost driver
exact	Tran et al., 2019	✓	✓	image	—	LPs × splits
hybrid		✓*	✓	set	✓	LPs × splits
approx		—	✓	set	✓	LPs
IBP	Gowal et al., 2019	—	—	bound	✓	matrix ops
CROWN	Zhang et al., 2018	—	✓	bound	✓	backsubstitution
α-CROWN	Xu et al., 2021	—	✓	bound	✓	optimized slopes
β-CROWN	Wang et al., 2021	✓*	✓	bound	✓	branch and bound
PWA-MILP	Schwartz et al., 2026	†	✓	bound	✓	binaries $\times K$
MILP	Tjeng et al., 2019	✓	✓	verdict	—	binaries
ReLU α-CROWN§	Xu et al., 2021	—	—	bound	✓	hinges × slopes
ReLU α, β-CROWN§	Wang et al., 2021	✓	✓	verdict	✓	hinges × splits

and bound over the node ranges, split at knots where possible, with a multiplier $\beta \geq 0$ per hidden-node split; both are our KAN implementations for depth at most 2 (see Appendix C.8). Four points of their application are specific to KANs. Lemma 3(i) holds for every slope, so any α-CROWN slopes are sound. As in the final-layer elision, a linear function $\boldsymbol{h}^\top N(\boldsymbol{x})$ of the output is relaxed through the univariate *combined edges* $\sum_q h_q \phi_{q,p}$ of the last layer. Hidden-node intervals need no backward tightening, as the first layer acts on each input separately. On a degree-1 no-base network every edge is affine once its input range lies in one knot cell, so β-CROWN is complete given enough time, up to the Lipschitz margin. MILP, the exact encoding for degree-1 KANs after Tjeng et al. (2019), uses one binary variable per knot cell, and PWA-MILP (Schwartz et al., 2026) one per piece of a piecewise-affine over-approximation. The released α-CROWN and α, β-CROWN tools have no B-spline operator, so we also run them on the ReLU conversion of the KAN.

5.4 Derivative networks and barrier conditions

For a two-layer KAN barrier $B(\boldsymbol{x}) = \sum_q \phi_q(z_q)$ with hidden nodes $z_q = \sum_p \psi_{q,p}(x_p)$ and closed-loop vector field f, barrier verification needs the Lie derivative $\mathcal{L}_f B = \nabla B \cdot f$. The derivative of an edge's spline part is a degree-$(k-1)$ spline on the same knots, so every factor of $\partial B / \partial x_p = \sum_q \phi_q'(z_q)\, \psi_{q,p}'(x_p)$ is within the certified-range machinery, and the boxes whose certified B-range straddles 0 cover the zero level set. Proposition 2 (Appendix B) makes this a sound discharge of the barrier derivative condition on trained networks, whereas KAN4CBC checks only a symbolic surrogate. On the trained degree-3 pendulum barrier the check passes, so the barrier meets all three barrier conditions.

6 Experiments

The benchmark and what each member proves. We evaluate the family of Table 1 on the 26 networks and eleven tasks of Table 2, among them four literature case studies retrained by us: the KAN4CBC pendulum networks (CS1; Zhang et al., 2025a), the pykan example (CS2; Liu et al., 2025b), a KAN-ODE Lotka–Volterra field (CS3; Koenig et al., 2024) and Feynman I.16.6 (CS4; Udrescu & Tegmark, 2020). FALSIFIED always carries a re-executed counterexample, and every certified bound is checked against a value attained by sampling and a PGD-style attack (the *width ratio* is the certified over the attained width). PWA-MILP is our reimplementation of Schwartz et al. (2026) with uniform pieces; the protocol and every failed check are in Appendices C.1 and C.12.

Table 2: **The benchmark and what each member proves.** Entries count the queries or boxes proved out of #, and F counts those with a re-executed counterexample. **Bold**: all decided ($\#-F$ proved), the most any sound method can prove. *Italic*: mean width ratio of the certified enclosure (1: exact). Labels as in Table 1; on robustness rows IBP uses final-layer elision. The CROWN column shows β-CROWN, the best of the CROWN methods on every row. PWA-MILP uses $K{=}4$ pieces. $\times$: not run; –: not applicable; ‡: capped (PWA-MILP: dual bound; β-CROWN: 9 of 36 boxes at its time limit, sound enclosures). Benchmark details available in Appendix C.1 and Table 4.

		degree 1, no base						degree 3 + SiLU					
Task	#	F	exact	approx	IBP	CROWN	MILP	F	hybrid	approx	IBP	CROWN	PWA-MILP
Local robustness, 2–30 inputs													
Annulus	660	220	**440**	349	251	**440**	**440**	209	450	435	408	**451**	441
Two-moons	500	5	**495**	471	424	**495**	**495**	9	**491**	489	483	**491**	**491**
Gaussian-4	120	4	**116**	115	112	**116**	**116**	5	**115**	**115**	114	**115**	**115**
UCI×3	480	58	**422**	398	369	**422**	**422**	18	456	454	431	**462**	442‡
Local robustness, 784 inputs													
MNIST	200	143	×	**57**	**57**	**57**	×	70	×	129	**130**	**130**	×
Fashion-MNIST	200		*(degree-3 network only)*					102	×	**98**	**98**	**98**	×
MNIST, depth 2	50		*(degree-3 network only)*					3	×	**47**	44	**47**	×
Case studies													
CS1 barrier	129	0	**129**	128	123	**129**	**129**	0	**129**	**129**	122	**129**	**129**
CS2 enclosure	36	–	*1.00*	*1.21*	*1.27*	*1.00*‡	*1.00*	–	*1.07*	*1.13*	*1.55*	*1.00*	*1.04*
CS3 Euler step	21	0	**21**	**21**	**21**	**21**	**21**	0	**21**	**21**	**21**	**21**	**21**
CS4 range	64	0	**64**	**64**	**64**	**64**	**64**	0	**64**	61	57	**64**	**64**

Appendix C visualizes every verified property on its benchmark: local robustness per radius (Figure 3), monotone ordering (Figure 9), reachability (Figure 10), output range (Figure 11) and the barrier conditions (Figure 12).

What the family certifies. A row's *ceiling*, $\#-F$ in Table 2, is the most any sound method can prove, and the CROWN column is β-CROWN, the best of CROWN, α-CROWN and β-CROWN on every row (Table 8, Appendix C.8). In the degree-1 regime the complete members agree and reach the ceiling (Figure 2(a,b)): star-set exact and the exact MILP decide all $2{,}010$ queries and boxes, and β-CROWN decides every query, including the 200 MNIST queries at 784 inputs (Appendix C.6). Star-set exact and hybrid are not run on the 784-input rows, where the number of cases of Lemma 2 reaches 2^{30} (MNIST) and 2^{289} (Fashion-MNIST). In the degree-3+SiLU regime β-CROWN still reaches the ceiling on every row, and star-set hybrid, PWA-MILP, α-CROWN and β-CROWN prove every control and scientific property (Appendix C.8). β-CROWN alone reaches the ceiling on the degree-3 annulus and UCI rows. α-CROWN also reaches it on all four 784-input networks, and star-set approx and CROWN miss it by one degree-3 MNIST image at $8/255$ (Appendix C.11).

What is certified on the literature networks. On the retrained pendulum barrier B_{net} and controller u_{net}, both sign conditions hold on all 65 boxes and $|u_{\mathrm{net}}| \le 14$ on all 64 in both regimes, and the interval test of Proposition 2 discharges the Lie-derivative condition on all 34 boxes covering the degree-3 barrier's zero level set, so this barrier and controller meet all three barrier conditions (Figure 2(d), Appendix C.5). Exact mode certifies the exact range of the degree-1 pykan network and proves its monotone ordering on all 7 separated slab pairs. Every member proves the Lotka–Volterra Euler step, a 30-step reach tube certifies that both target sets are reached and two unsafe boxes are avoided, and $|N| \le 1$ is certified on all 64 Feynman boxes (Appendix C.4).

Where the guarantees and their cost come from. In Figure 2(c), star-set hybrid, PWA-MILP at $K{=}8$ and β-CROWN reach the ceiling of 90 on 120 degree-3 annulus queries at medians of 495, 907 and 55 ms per query. As Lemmas 2 and 4 and Proposition 1 predict, exact mode's star count grows with the grid, depth and input width but not with the output width (Figure 2(f)), the relaxation gap is quadratic in the interval width, and extrapolating past the extended grid is unsound (Appendix C.3).

What each choice buys. Further ablations (Appendix C.7) show exact mode removes all 91 UNKNOWN verdicts of approx mode on the degree-1 annulus network and hybrid mode 15 of 16 on the degree-3 network. The exactly verifiable regime costs robustness, with 34% of the MNIST images certified at $2/255$ at degree 1 against 90% at degree 3, both at the PGD ceiling (Figure 2(e)); it also

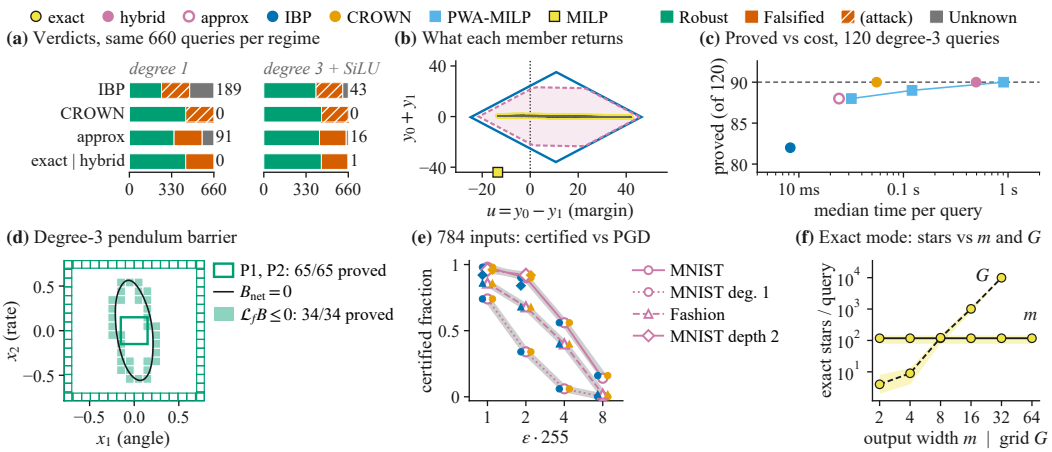


Figure 2: **The family on the benchmark.** Labels are as in Table 2. CROWN shows β-CROWN, the best of the CROWN methods. (a) Verdicts on 660 annulus queries per regime. Hatching marks counterexamples supplied by the shared attack, and numbers count UNKNOWN verdicts. β-CROWN tries the attack point before it searches, so all of its counterexamples are hatched. (b) Outputs on one falsified degree-1 query, with the box of IBP. The exact image has 25 stars, approx has one, and the dotted line is the decision boundary. CROWN is not drawn, since β-CROWN returns a verdict but no set. (c) Times are medians per query. PWA-MILP uses $K = 2, 4, 8$ pieces from left to right, and the dashed line is the attack ceiling. The query set and timing of (c) are in Appendices C.7 and C.8. (d) The initial box and the 64 unsafe boxes are proved, and $\mathcal{L}_f B \le 0$ is proved on all 34 boxes that cover $\{B_{\text{net}} = 0\}$. (e) The grey band is the PGD ceiling. (f) Solid lines vary the output width m at $G = 8$, dashed lines vary the grid G at $m = 2$, and bands show the range over points.

multiplies RMSE by a median factor of 2.83 on six regression tasks and changes accuracy by -0.8 to $+0.2$ points on three classification tasks (Appendix C.10). The ReLU conversion (Appendix C.9) keeps the complete verdicts and often speeds up exact methods but discards the univariate structure: native interval bounds certify 317 of 381 degree-1 queries and converted ones none, and the native members together prove 402 of 429 degree-3 queries against 394 for all converted methods.

Limitations. Appendix C.11 collects the negative results and the limits of the runs, notably the Lie-derivative violations of the degree-1 pendulum barrier. The family and the benchmark target the original KAN architecture of Liu et al. (2025b), with B-spline edges and SiLU base branches summed at the nodes (Equation 1), although Lemma 3 holds for any continuous edge and Appendix C.4 applies it to one Gaussian-RBF KAN. Future work could apply these verification concepts to richer architectures such as Kolmogorov–Arnold transformers (Yang & Wang, 2025), KANs with multiplication nodes (Liu et al., 2025a) and convolutional KANs.

7 CONCLUSION

This work builds a family of verification methods on the structure of the KAN layer, proves the facts the family rests on, measures it on a 26-network benchmark, and demonstrates it up to 784-input KANs. We verify uniform per-layer grids as fixed in the checkpoint (adaptive grids are covered once frozen, and non-uniform knots need only an implementation extension). β-CROWN decides every degree-1 query of the benchmark, star-set exact and the exact MILP those with up to 30 inputs, star-set exact decides synthetic networks of up to six layers, the relaxed members verify 784-input networks of depth at most 2, and deeper networks at that width remain open. The trained degree-3 pendulum barrier and controller meet all three barrier conditions. A natural next step is to try and verify more complex KAN architectures like convolutional KANs, KAN transformers, *etc*.

AI USE STATEMENT

Anthropic's Claude (the Opus 5 and Fable 5 families) models were used to assist in the development of this manuscript. Claude supported proof generation in Lean 4 with Mathlib (Moura & Ullrich, 2021; mathlib Community, 2020), then translating the proofs to a human-readable form, these were then verified and their readability improved manually. Claude also assisted with writing code, highlighting relevant references, generating figures, polishing writing, and drafting some sections of the paper. However, much effort and time was spent by the authors to iterate, improve, verify and enhance the generated outputs. The authors remain ultimately responsible for the text, code, and any other aspect relating to the creation of this work.

ETHICS STATEMENT

The authors have read and adhere to the ICLR Code of Ethics. This work develops methods that certify properties of trained Kolmogorov–Arnold networks, with the aim of making their use in safety-critical settings more trustworthy. All experiments use publicly available datasets (MNIST, Fashion-MNIST and three UCI datasets) or synthetic and simulated tasks, and involve no human subjects or personally identifiable information; the released benchmark contains only trained networks and property specifications derived from these datasets. A certificate holds only for the stated network, property and input set, up to the numerical tolerances we report, and should not be read as a guarantee about a deployed system beyond that scope. The adversarial attacks in our experiments serve only to search for counterexamples to the properties we verify, and every failed check of our experiments is reported in Appendix C.12.

REPRODUCIBILITY STATEMENT

The code, data and proofs are available in an anonymized repository, `https://anonymous.4open.science/r/KAN-verification`. The repository contains the verifier with its three star-set modes and the other members of the method family, every experiment script with its cached measurement rows, fixed seeds and rendered report, the scripts that regenerate every table and figure of the paper from those rows, and the Lean development; its README gives the installation steps, the recorded software versions and the recorded run times. Since no KAN verification benchmark exists in any VNN-COMP edition, it also contains a standalone **benchmark release** of the 26 trained KANs of Table 2: every case-study network in both regimes, the two $[784, 10]$ MNIST KANs, the $[784, 10]$ Fashion-MNIST KAN with the attacked parameterization, the depth-2 $[784, 16, 10]$ MNIST KAN whose two layers carry different grids, and tabular KANs on three UCI datasets in both regimes, each serialized with its grid metadata (a KAN checkpoint is not fully specified by its weights; Section 4) and with machine-readable property specifications that carry the concrete query sets, the input normalization and the reference verdicts of the members of Table 4 (Appendix C.1), so that future KAN verifiers have a common target. Every number quoted in the paper is regenerable from a script, and every certified bound is accompanied by the attainability check described in Section 6.

REFERENCES

Tal Alter, Raz Lapid, and Moshe Sipper. On the robustness of kolmogorov-arnold networks: An adversarial perspective. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL `https://openreview.net/forum?id=uafxqhImPM`.

László Antal, Erika Ábrahám, and Hana Masara. Generalizing neural network verification to the family of piece-wise linear activation functions. *Sci. Comput. Program.*, 243(C), July 2025. ISSN 0167-6423. doi: 10.1016/j.scico.2025.103269. URL `https://doi.org/10.1016/j.scico.2025.103269`.

Stanley Bak and Parasara Sridhar Duggirala. Simulation-equivalent reachability of large linear systems with inputs. In Rupak Majumdar and Viktor Kunčak (eds.), *Computer Aided Verification*, pp. 401–420, Cham, 2017. Springer International Publishing. ISBN 978-3-319-63387-9.

Zavareh Bozorgasl and Hao Chen. Wav-kan: Wavelet kolmogorov-arnold networks. *arXiv preprint arXiv:2405.12832*, 2024.

Christopher Brix, Stanley Bak, Taylor T. Johnson, and Haoze Wu. The fifth international verification of neural networks competition (vnn-comp 2024): Summary and results, 2024. URL https://arxiv.org/abs/2412.19985.

Edward Cheung and Yuying Li. Solving separable nonsmooth problems using frank-wolfe with uniform affine approximations. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 2035–2041. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/281. URL https://doi.org/10.24963/ijcai.2018/281.

Carl de Boor. *A Practical Guide to Splines*. Springer Verlag, New York, 1978.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy (SP)*, pp. 3–18, 2018. doi: 10.1109/SP.2018.00058.

Alexander B. Givental, Boris A. Khesin, Jerrold E. Marsden, Alexander N. Varchenko, Victor A. Vassiliev, Oleg Ya. Viro, and Vladimir M. Zakalyukin (eds.). *On functions of three variables*, pp. 5–8. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009. ISBN 978-3-642-01742-1. doi: 10.1007/978-3-642-01742-1_2. URL https://doi.org/10.1007/978-3-642-01742-1_2.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Arthur Mann, and Pushmeet Kohli. Scalable verified training for provably robust image classification. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4841–4850, 2019. doi: 10.1109/ICCV.2019.00494.

Björn Heiderich, Max-Lion Schumacher, and Marco Huber. Training and verifying robust kolmogorov-arnold networks. In *ICLR 2025 Workshop: VerifAI: AI Verification in the Wild*, 2025. URL https://openreview.net/forum?id=N9hl7CShRA.

Yuntian Hou, Tianrui Ji, Di Zhang, and Angelos Stefanidis. Kolmogorov-arnold networks: A critical assessment of claims, performance, and practical viability. *arXiv preprint arXiv:2407.11075*, 2024.

Tanuj Karia, Giacomo Lastrucci, and Artur M. Schweidtmann. Deterministic global optimization over trained kolmogorov arnold networks, 2025. URL https://arxiv.org/abs/2503.02807.

Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International conference on computer aided verification*, pp. 97–117. Springer, 2017.

Guy Katz, Derek A Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zeljić, et al. The marabou framework for verification and analysis of deep neural networks. In *International conference on computer aided verification*, pp. 443–452. Springer, 2019.

Konstantin Kaulen, Tobias Ladner, Stanley Bak, Christopher Brix, Hai Duong, Thomas Flinkow, Taylor T. Johnson, Lukas Koller, Edoardo Manino, ThanhVu H Nguyen, and Haoze Wu. The 6th international verification of neural networks competition (vnn-comp 2025): Summary and results, 2025. URL https://arxiv.org/abs/2512.19007.

Victor A Kich, Jair A Bottega, Raul Steinmetz, Ricardo B Grando, Ayano Yorozu, and Akihisa Ohya. Kolmogorov-arnold networks for online reinforcement learning. In *2024 24th international conference on control, automation and systems (ICCAS)*, pp. 958–963. IEEE, 2024.

Benjamin C Koenig, Suyong Kim, and Sili Deng. Kan-odes: Kolmogorov–arnold network ordinary differential equations for learning dynamical systems and hidden physics. *Computer Methods in Applied Mechanics and Engineering*, 432:117397, 2024.

Andrei Nikolaevich Kolmogorov. On the representation of continuous functions of several variables as superpositions of continuous functions of a smaller number of variables. *Dokl. Akad. Nauk*, 108(2), 1956.

Chenxin Li, Xinyu Liu, Wuyang Li, Cheng Wang, Hengyu Liu, Yifan Liu, Zhen Chen, and Yixuan Yuan. U-kan makes strong backbone for medical image segmentation and generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 4652–4660, 2025.

Ziyao Li. Kolmogorov-arnold networks are radial basis function networks. *arXiv preprint arXiv:2405.06721*, 2024.

Ziming Liu, Max Tegmark, Pingchuan Ma, Wojciech Matusik, and Yixuan Wang. Kolmogorov-arnold networks meet science. *Physical Review X*, 15(4):041051, 2025a. doi: 10.1103/4t7t-v19l.

Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljacic, Thomas Hou, and Max Tegmark. Kan: Kolmogorov–arnold networks. In *International conference on learning representations*, volume 2025, pp. 70367–70413, 2025b.

Diego Manzanas Lopez, Sung Woo Choi, Hoang-Dung Tran, and Taylor T. Johnson. NNV 2.0: The neural network verification tool. In *35th International Conference on Computer-Aided Verification (CAV)*, July 2023.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

The mathlib Community. The lean mathematical library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, POPL '20, pp. 367–381. ACM, January 2020. doi: 10.1145/3372885.3373824. URL http://dx.doi.org/10.1145/3372885.3373824.

Mohammad Meymani and Roozbeh Razavi-Far. Kan-robust-bench: A benchmark for evaluating the robustness of kolmogorov-arnold networks, 2026. URL https://arxiv.org/abs/2608.21488.

Leonardo de Moura and Sebastian Ullrich. The lean 4 theorem prover and programming language. In *Automated Deduction – CADE 28: 28th International Conference on Automated Deduction, Virtual Event, July 12–15, 2021, Proceedings*, pp. 625–635, Berlin, Heidelberg, 2021. Springer-Verlag. ISBN 978-3-030-79875-8. doi: 10.1007/978-3-030-79876-5_37. URL https://doi.org/10.1007/978-3-030-79876-5_37.

Mark Niklas Mueller, Mislav Balunovic, and Martin Vechev. Certify or predict: Boosting certified robustness with compositional architectures. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=USCNapootw.

Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition, 2014. ISBN 1461346916.

Qi Qiu, Tao Zhu, Helin Gong, Liming Chen, and Huansheng Ning. Relu-kan: New kolmogorov-arnold networks that only need matrix addition, dot multiplication, and relu. In *2025 IEEE Smart World Congress (SWC)*, pp. 1686–1694. IEEE, 2025.

Nandi Schoots, Mattia Jacopo Villani, and Niels uit de Bos. Relating piecewise linear kolmogorov arnold networks to relu networks. In Yingzhen Li, Stephan Mandt, Shipra Agrawal, and Emtiyaz Khan (eds.), *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 199–207. PMLR, 03–05 May 2025. URL https://proceedings.mlr.press/v258/schoots25a.html.

Noah Schwartz, Chandra Kanth Nagesh, Sriram Sankaranarayanan, Ramneet Kaur, Tuhin Sahai, and Susmit Jha. Optimized piecewise affine abstractions of neural networks with learnable activation functions. In *Proceedings of the 26th Conference on Formal Methods in Computer-Aided Design–FMCAD 2026*, pp. 678–693. TU Wien Academic Press, 2026.

Zhouxing Shi, Qirui Jin, Zico Kolter, Suman Jana, Cho-Jui Hsieh, and Huan Zhang. Neural network verification with branch-and-bound for general nonlinearities. In Arie Gurfinkel and Marijn Heule (eds.), *Tools and Algorithms for the Construction and Analysis of Systems*, pp. 315–335, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-90643-5.

Khemraj Shukla, Juan Diego Toscano, Zhicheng Wang, Zongren Zou, and George Em Karniadakis. A comprehensive and fair comparison between mlp and kan representations for differential equations and operator networks. *Computer Methods in Applied Mechanics and Engineering*, 431: 117290, 2024.

Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/f2f446980d8e971ef3da97af089481c3-Paper.pdf.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. 3(POPL), January 2019. doi: 10.1145/3290354. URL https://doi.org/10.1145/3290354.

Sidharth SS, Keerthana AR, Anas KP, et al. Chebyshev polynomial-based kolmogorov-arnold networks: An efficient architecture for nonlinear function approximation. *arXiv preprint arXiv:2405.07200*, 2024.

Vincent Tjeng, Kai Y. Xiao, and Russ Tedrake. Evaluating robustness of neural networks with mixed integer programming. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=HyGIdiRqtm.

Hoang-Dung Tran, Diago Manzanas Lopez, Patrick Musau, Xiaodong Yang, Luan Viet Nguyen, Weiming Xiang, and Taylor T. Johnson. Star-based reachability analysis of deep neural networks. In *Formal Methods – The Next 30 Years: Third World Congress, FM 2019, Porto, Portugal, October 7–11, 2019, Proceedings*, pp. 670–686, Berlin, Heidelberg, 2019. Springer-Verlag. ISBN 978-3-030-30941-1. doi: 10.1007/978-3-030-30942-8_39. URL https://doi.org/10.1007/978-3-030-30942-8_39.

Hoang-Dung Tran, Stanley Bak, Weiming Xiang, and Taylor T. Johnson. Verification of deep convolutional neural networks using imagestars. In *32nd International Conference on Computer-Aided Verification (CAV)*. Springer, July 2020.

Hoang-Dung Tran, Neelanjana Pal, Diego Manzanas Lopez, Patrick Musau, Xiaodong Yang, Luan Viet Nguyen, Weiming Xiang, Stanley Bak, and Taylor T. Johnson. Verification of piecewise deep neural networks: a star set approach with zonotope pre-filter. *Form. Asp. Comput.*, 33(4–5):519–545, August 2021. ISSN 0934-5043. doi: 10.1007/s00165-021-00553-4. URL https://doi.org/10.1007/s00165-021-00553-4.

Hoang-Dung Tran, Sungwoo Choi, Hideki Okamoto, Bardh Hoxha, Georgios Fainekos, and Danil Prokhorov. Quantitative verification for neural networks using probstars. In *Proceedings of the 26th ACM International Conference on Hybrid Systems: Computation and Control*, HSCC '23, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400700330. doi: 10.1145/3575870.3587112. URL https://doi.org/10.1145/3575870.3587112.

Hoang-Dung Tran, Sung Woo Choi, Yuntao Li, Qing Liu, Hideki Okamoto, Bardh Hoxha, and Georgios Fainekos. Starv: A qualitative and quantitative verification tool for learning-enabled systems. In *Computer Aided Verification: 37th International Conference, CAV 2025, Zagreb, Croatia, July 23–25, 2025, Proceedings, Part II*, pp. 376–394, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-3-031-98678-9. doi: 10.1007/978-3-031-98679-6_17. URL https://doi.org/10.1007/978-3-031-98679-6_17.

Silviu-Marian Udrescu and Max Tegmark. Ai feynman: A physics-inspired method for symbolic regression. *Science Advances*, 6(16):eaay2631, 2020. doi: 10.1126/sciadv.aay2631. URL https://www.science.org/doi/abs/10.1126/sciadv.aay2631.

Cristian J Vaca-Rubio, Luis Blanco, Roberto Pereira, and Màrius Caus. Kolmogorov-arnold networks (kans) for time series analysis. In *2024 IEEE Globecom Workshops (GC Wkshps)*, pp. 1–6. IEEE, 2024.

Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Formal security analysis of neural networks using symbolic intervals. In *Proceedings of the 27th USENIX Conference on Security Symposium*, SEC'18, pp. 1599–1614, USA, 2018. USENIX Association. ISBN 9781931971461.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J. Zico Kolter. Beta-crown: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 29909–29921. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/fac7fead96dafceaf80c1daffeae82a4-Paper.pdf.

Yixuan Wang, Jonathan Siegel, Ziming Liu, and Thomas Hou. On the expressiveness and spectral bias of kans. In *International Conference on Learning Representations*, volume 2025, pp. 27492–27511, 2025.

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. *Advances in Neural Information Processing Systems*, 33, 2020.

Kaidi Xu, Huan Zhang, Shiqi Wang, Yihan Wang, Suman Jana, Xue Lin, and Cho-Jui Hsieh. Fast and Complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=nVZtXBI6LNn.

Xingyi Yang and Xinchao Wang. Kolmogorov-arnold transformer. In *International Conference on Learning Representations*, volume 2025, pp. 76063–76086, 2025.

Runpeng Yu, Weihao Yu, and Xinchao Wang. Kan or mlp: A fairer comparison. *arXiv preprint arXiv:2407.16674*, 2024.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. *Advances in Neural Information Processing Systems*, 31:4939–4948, 2018. URL https://arxiv.org/pdf/1811.00866.pdf.

Xiongqi Zhang, Ning Lv, Wang Lin, and Zuohua Ding. Formal synthesis of safe kolmogorov-arnold network controllers with barrier certificates. In *IJCAI*, pp. 302–310, 2025a.

Xiongqi Zhang, Junwei Xu, Yang Wang, Dongming Xiang, Wang Lin, and Zuohua Ding. Formal synthesis of barrier certificates using fourier kolmogorov-arnold network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1138–1146, 2025b.

A Algorithms

Algorithm description. Algorithm 1 realizes Theorems 1 and 2 one layer at a time, and Algorithm 2 turns its output stars into a verification decision. Within a layer, lines 4 to 16 of Algorithm 1 split each input star Θ into region stars. Lines 17 to 34 then propagate every region star through the layer.

The split proceeds node by node in index order and produces the region stars of Lemma 2. Two LPs compute the range $[l_p, u_p]$ of x_p over the current region star exactly. These LPs are solved for all coordinates of a star at once, and their results are cached. Each piece of $\mathcal{P}$ selected at line 9 yields one region star through at most two predicate rows. A region star that one LP proves empty is discarded. Line 10 moves an end of $J_p \cap [l_p, u_p]$ that lies within 10^{-12} of the range bound on its

Algorithm 1 Layer propagation $\mathcal{F}_{d-1} \mapsto \mathcal{F}_d$ of the star-set member in modes `exact`, `hybrid` and `approx`. Started from $\mathcal{F}_0 = (\langle \mathbf{0}, \boldsymbol{I}, \mathcal{B} \rangle)$ and applied for $d = 1, \ldots, D$, it computes a list $\mathcal{F}_D$ as in Theorem 2; when the flag *relaxed* stays unset, $\mathcal{F}_D$ has the properties stated in Theorem 1.

Require: list $\mathcal{F}_{d-1}$ of stars; layer L with n input nodes, m output nodes and extended partition $\mathcal{P}$; mode $\in \{\texttt{exact}, \texttt{hybrid}, \texttt{approx}\}$, where `exact` requires $k = 1$ and $w \equiv 0$; split cap $S_{\max}$; the flag *relaxed* returned for layer $d-1$, unset when $d = 1$
Ensure: the list $\mathcal{F}_d$ and the flag *relaxed*, set if a lifted star has been emitted in this or an earlier layer

1: $\mathcal{F}_d \leftarrow ()$; $n' \leftarrow n$, or $n' \leftarrow 0$ in mode `approx` ▷ `approx` splits no node
2: **for** each $\Theta \in \mathcal{F}_{d-1}$ **do**
3: frontier $\leftarrow$ $((\Theta, ()))$, the list of pairs $(\Theta_J, \mathcal{I})$ of a region star of Θ and the list $\mathcal{I}$ of the intervals $I_1, \ldots, I_{p-1}$ of the nodes split so far
4: **for** $p = 1, \ldots, n'$ **do**
5: **if** |frontier| $\geq S_{\max}$ **then** in mode `exact` stop with an error, in mode `hybrid` record that nodes $p, \ldots, n$ stay unsplit and **break**
6: frontier$'$ $\leftarrow ()$
7: **for** each $(\Theta_J, \mathcal{I})$ in frontier **do**
8: $[l_p, u_p] \leftarrow$ the range of x_p over Θ_J ▷ two LPs
9: **for** each piece $J_p \in \mathcal{P}$ that meets $[l_p, u_p]$, in more than one point if $l_p < u_p$ **do**
10: $I_p \leftarrow J_p \cap [l_p, u_p]$, its lower end set to l_p if within 10^{-12} of it, its upper end to u_p likewise
11: $\Theta' \leftarrow \Theta_J \cap \{\boldsymbol{x} : x_p \in I_p\}$, one row per end of I_p not in $\{l_p, u_p\}$ ▷ Lemma 2(ii)
12: **if** $\Theta' \neq \emptyset$ (one LP if a row was added) **then** append $(\Theta', (\mathcal{I}, I_p))$ to frontier$'$
13: **end for**
14: **end for**
15: frontier $\leftarrow$ frontier$'$
16: **end for**
17: **for** each $(\Theta_J, \mathcal{I})$ in frontier **do**
18: $J_p = I_p \leftarrow [l_p, u_p]$, the range of x_p over Θ_J, for each unsplit node p ▷ two LPs; Section 5.2
19: **for** $p = 1, \ldots, n$, for the m edges $\phi_{q,p}$ on I_p **do**
20: **if** L has no base branch and I_p meets, as at line 9, one piece of $\mathcal{P}$: a tail or a knot cell on which every $\rho_{q,p}$ has degree at most 1 **then**
21: $s_{q,p} \leftarrow$ the slope of $\rho_{q,p}$; $\underline{o}_{q,p} = \overline{o}_{q,p} \leftarrow$ its intercept; $[\underline{\phi}_{q,p}, \overline{\phi}_{q,p}] \leftarrow$ the smaller and the larger of its values at the ends of I_p ▷ Lemma 3(ii)
22: **else**
23: $\Lambda_{q,p} \leftarrow 1.1|w_{q,p}|$ plus the largest, over the knot cells $[t_j, t_{j+1}]$ that I_p meets as at line 9 (0 if none), of the bound stated after Lemma 6 with $\tau_{\max}$ taken over $I_p \cap [t_j, t_{j+1}]$
24: $[\underline{\phi}_{q,p}, \overline{\phi}_{q,p}] \leftarrow$ the range of $\phi_{q,p}$ on I_p by Lemma 6 with $\Lambda_{q,p}$ and $\nu(\max_q \Lambda_{q,p})$ intervals, where $\nu(\Lambda) = \min\{8192, \max\{8, \lceil \Lambda |I_p| / (2 \cdot 10^{-10}) \rceil\}\}$
25: $s_{q,p} \leftarrow s_c$ of $\phi_{q,p}$ on I_p, or 0 if I_p is a point
26: $[\underline{o}_{q,p}, \overline{o}_{q,p}] \leftarrow$ the range of $\phi_{q,p}(x) - s_{q,p}x$ on I_p by Lemma 6 with $\Lambda_{q,p} + |s_{q,p}|$ and $\nu(\max_q(\Lambda_{q,p} + |s_{q,p}|))$ intervals ▷ Lemma 3(i)
27: **end if**
28: **end for**
29: **if** $\underline{o}_{q,p} = \overline{o}_{q,p}$ for every edge (q, p) **then**
30: append the affine image $\boldsymbol{S}\Theta_J + \boldsymbol{b} = L(\Theta_J)$ to $\mathcal{F}_d$, where $\boldsymbol{S} = (s_{q,p})$ and the vector $\boldsymbol{b}$ holds the row sums of $(\underline{o}_{q,p})$ ▷ Lemma 5
31: **else**
32: append the lifted star $\Omega(\Theta_J)$ of Section 5.2, with these slopes, offsets and ranges and its tighten rows, to $\mathcal{F}_d$; set *relaxed* ▷ Theorem 2
33: **end if**
34: **end for**
35: **end for**
36: **return** $\mathcal{F}_d$, *relaxed*

side to that bound. This rounding avoids adding a row that $[l_p, u_p]$ implies up to LP tolerance. The star Θ' of line 11 then contains $\Theta_J \cap \{\boldsymbol{x} : x_p \in J_p\}$, and I_p contains the range of x_p over Θ'. Exact and hybrid modes split the nodes in this way, whereas approx mode splits none (line 1).

The split cap acts at line 5, before a node is split. It takes effect when the frontier of one input star holds $S_{\max}$ or more region stars while a node remains to be split. Exact mode then stops with an error, because truncating the split would forfeit the exactness of Theorem 1. Hybrid mode instead

Algorithm 2 Robustness verdict from the output stars.

Require: KAN N with n inputs; box $\mathcal{B} = [\boldsymbol{x}_0 \pm \varepsilon]$; label y^*; the list $\mathcal{F}_D = (\Omega_1, \ldots, \Omega_r)$ of Algorithm 1, $\Omega_i = \langle \boldsymbol{c}_i, \boldsymbol{V}_i, P_i \rangle$, with its flag *relaxed*
Ensure: ROBUST; FALSIFIED, with a counterexample $\boldsymbol{x}^* \in \mathcal{B}$ or with the no-counterexample marker; or UNKNOWN; and the margin μ^*
1: $\mu^* \leftarrow +\infty$
2: **for** each $i = 1, \ldots, r$ and each class $j \neq y^*$ **do**
3: $\quad \mu_{ij} \leftarrow \min_{\boldsymbol{\alpha} \in P_i} \left[(\boldsymbol{c}_i + \boldsymbol{V}_i \boldsymbol{\alpha})_{y^*} - (\boldsymbol{c}_i + \boldsymbol{V}_i \boldsymbol{\alpha})_j \right]$ by one LP, $+\infty$ if $P_i = \emptyset$ ▷ Corollary 1
4: $\quad$ **if** $\mu_{ij} < \mu^*$ **then** $\mu^* \leftarrow \mu_{ij}$ and $\boldsymbol{\alpha}^* \leftarrow$ a minimizer
5: **end for**
6: **if** $\mu^* > 0$ **then**
7: $\quad$ **return** ROBUST ▷ Theorem 2; Corollary 1(i) if unrelaxed
8: **end if**
9: $\boldsymbol{x}^* \leftarrow$ the first n entries of $\boldsymbol{\alpha}^*$, clipped to $\mathcal{B}$ ▷ the inputs; each $\boldsymbol{\delta}$ follows them
10: **if** $\min\{j : N(\boldsymbol{x}^*)_j = \max_{j'} N(\boldsymbol{x}^*)_{j'}\} \neq y^*$ **then**
11: $\quad$ **return** FALSIFIED with $\boldsymbol{x}^*$ ▷ re-executed; Corollary 1(ii) if unrelaxed and $\mu^* < 0$
12: **end if**
13: search $\mathcal{B}$ for a point $\boldsymbol{x}^*$ with $N(\boldsymbol{x}^*)_{y^*} < \max_{j \neq y^*} N(\boldsymbol{x}^*)_j$: its center, its 2^n corners if $n \leq 12$ and 2,000 uniform samples, then a coordinate-wise descent of the margin from the best of them in steps of $\varepsilon/4$, for at most 50 rounds
14: **if** such a point was found **then**
15: $\quad$ **return** FALSIFIED with $\boldsymbol{x}^*$
16: **else if** not *relaxed* **then**
17: $\quad$ **return** FALSIFIED with the no-counterexample marker ▷ Corollary 1(iii), up to LP tolerance
18: **else**
19: $\quad$ **return** UNKNOWN ▷ the minimizer may lie in the relaxation gap
20: **end if**

records the truncation, leaves the remaining nodes unsplit and continues. A frontier that reaches or exceeds the cap with the split of the last node is completed as it stands.

The propagation treats every region star by the same rule, in every mode. Each node p carries the interval I_p on which its m edges are enclosed. For a split node, I_p is the interval of line 10. For an unsplit node, in approx mode or after the cap, I_p is the range of x_p over the region star itself (line 18). This range is the unsplit assignment $J_p = [l_p, u_p]$ of Section 5.2. Both intervals contain the range of x_p over Θ_J, as Theorem 2 requires. Lines 20 and 21 handle the exact case. The case requires a layer without the base branch. It also requires I_p to lie on a tail or on a knot cell where none of the m polynomials $\rho_{q,p}$ has a nonzero coefficient of order at least 2. Every edge leaving p is then affine, and its slope and intercept are read off the coefficients. The two offsets of each such edge coincide by Lemma 3(ii).

In exact mode, every interval I_p lies in its piece J_p up to the rounding of line 10. The exact case, together with the point intervals discussed below, therefore applies to every node up to this rounding. A completed exact run thus computes the entries of $\mathcal{F}_D$ of Theorem 1, with two exceptions. First, line 9 drops a piece that a range $[l_p, u_p]$ of positive width meets only in one endpoint. The region star of the dropped piece lies in that of the adjacent piece, where continuity makes both affine maps agree. The union and the cover of Theorem 1 are therefore unchanged. Slivers of positive width, in contrast, are kept, because the adjacent piece's polynomial would be the wrong affine map on them. Second, suppose that the range crosses a knot t by at most 10^{-12}, say $0 < u_p - t \leq 10^{-12}$. Line 10 then moves the end t of the interval of the piece below t to u_p. The case at l_p is symmetric. The resulting I_p meets two pieces, so even in exact mode its node is relaxed at lines 23 to 26. Unless every edge leaving the node is constant on I_p, the layer then emits a lifted star and marks the run relaxed. Algorithm 2 accordingly treats the run as relaxed and may return UNKNOWN, while Theorem 2 keeps the enclosure sound. All 660 exact-mode annulus queries of E1 ran unrelaxed. The case also cost no verdict in the matrix run (E15), where none of the 2,010 star-set exact verdicts is UNKNOWN.

Lines 23 to 26 are the parallel relaxation of Section 5.2 at the chord slope. The Lipschitz constant $\Lambda_{q,p}$ is the interval-arithmetic bound stated after Lemma 6, plus $1.1|w_{q,p}|$ for the base branch. In that bound, $\tau_{\max}$ is taken over $I_p \cap [t_j, t_{j+1}]$, and the bound is maximized over the knot cells that I_p meets. Lemma 6 certifies both the range of $\phi_{q,p}$ and the range of the residual $\phi_{q,p}(x) - s_{q,p}x$, whose

ends are the offsets. Lipschitz sampling is used instead of locating critical points by root finding. A root lost to rounding would miss an extremum, whereas Lemma 6 holds for every ν. The number $\nu(\Lambda)$ of sampling intervals, defined at line 24, is shared by the m edges of the node. Its argument Λ is the largest of their constants, $\max_q \Lambda_{q,p}$ for the range and $\max_q(\Lambda_{q,p} + |s_{q,p}|)$ for the offsets. The margin $\Lambda|I_p|/(2\nu)$ of Lemma 6 is therefore at most 10^{-10} exactly when $\Lambda|I_p| \le 1.6384 \cdot 10^{-6}$. Otherwise the cap $\nu = 8192$ binds, and each edge receives the margin $c|I_p|/16384$ for its own constant c. This margin is about $6 \cdot 10^{-5}$ when $c|I_p| = 1$. At lines 23 to 26, a point interval receives the slope 0 and two offsets equal to the edge value, without sampling.

The affine fast path of lines 29 and 30 emits the image $\boldsymbol{S}\Theta_J + \boldsymbol{b} = L(\Theta_J)$ instead of a lifted star. It is taken only when the two offsets of every edge of the region star are equal in floating point. They are equal at line 21 and on point intervals. Every edge is then affine on its interval I_p, and the image is exact. A test with a tolerance would lose soundness by up to the gap of each edge it collapses. It would also mark the run unrelaxed, so Algorithm 2 could return FALSIFIED with the no-counterexample marker where UNKNOWN is due. When the offsets of some edge differ, line 32 instead emits the lifted star $\Omega(\Theta_J)$ of Theorem 2 and marks the run as relaxed. The lifted star has one fresh variable $\delta_{q,p}$ per edge, bounded by its offsets. Since the tighten rows are on by default, it also has two rows per edge that confine $s_{q,p}x_p + \delta_{q,p}$ to the certified range. The ablation in Table 6 switches these rows off.

Algorithm 2 solves the $r(C-1)$ margin LPs μ_{ij} of Corollary 1 over the output stars (line 3). It returns ROBUST when all of them are positive. This verdict is sound in every mode by Theorem 2, and in an unrelaxed run it is the criterion of Corollary 1(i). Output-range queries replace the margin LPs by the interval hull of $\mathcal{F}_D$, which takes two LPs per output coordinate and star.

When the smallest margin μ^* is not positive, line 9 maps its minimizer $\boldsymbol{\alpha}^*$ back to an input. The predicate variables of $\langle \mathbf{0}, \boldsymbol{I}, \mathcal{B} \rangle$ are the input coordinates, and every lifted star appends its variables $\boldsymbol{\delta}$ after them. The first n entries of $\boldsymbol{\alpha}^*$ therefore form a point of $\mathcal{B}$. This point is a counterexample by Corollary 1(ii) when the run is unrelaxed and $\mu^* < 0$, and a candidate otherwise. Line 10 re-executes N on the point and accepts it when the predicted class differs from y^*. The predicted class is the first index of a maximal entry of $N(\boldsymbol{x}^*)$. A point of margin 0 whose tying class has a smaller index than y^* therefore also counts. If the point is rejected, the search of line 13 looks for a point of $\mathcal{B}$ with a strictly negative margin. When neither yields a counterexample, an unrelaxed run returns FALSIFIED with the no-counterexample marker. This is case (iii) of Corollary 1, diagnosed up to LP tolerance. A relaxed run returns UNKNOWN instead, because the minimizer may lie in the relaxation gap.

The verifier also fails closed. An LP in either algorithm that ends neither optimal nor proved infeasible makes the verdict UNKNOWN, instead of dropping a star or returning an infinite margin. The implementation uses float64 and an LP solver at default tolerances. This leaves the standard unmodeled residual shared with the auto_LiRPA and α, β-CROWN line of work (Xu et al., 2020; Wang et al., 2021).

B PROOFS

How these proofs were produced. We used agentic AI tools to convert the statements we wished to prove into Lean 4 with Mathlib (Moura & Ullrich, 2021; mathlib Community, 2020) and to prove them there, and then to convert the Lean proofs back into natural language. We verified by hand that the resulting proofs are correct and improved their readability by hand.

B.1 STAR TRANSFORMATIONS

Lemma 5 (Star transformations)**.** *Let $\Theta = \langle \boldsymbol{c}, \boldsymbol{V}, P \rangle$ be a star, $\boldsymbol{W}$ a matrix, $\boldsymbol{b}$ a vector and $\{\boldsymbol{x} : \boldsymbol{H}\boldsymbol{x} \le \boldsymbol{g}\}$ a polyhedron. Then $\boldsymbol{W}\Theta + \boldsymbol{b} = \langle \boldsymbol{W}\boldsymbol{c} + \boldsymbol{b}, \boldsymbol{W}\boldsymbol{V}, P \rangle$ and $\Theta \cap \{\boldsymbol{x} : \boldsymbol{H}\boldsymbol{x} \le \boldsymbol{g}\} = \langle \boldsymbol{c}, \boldsymbol{V}, P \cap \{\boldsymbol{\alpha} : \boldsymbol{H}\boldsymbol{V}\boldsymbol{\alpha} \le \boldsymbol{g} - \boldsymbol{H}\boldsymbol{c}\} \rangle$.*

Proof. This is proven by Tran et al. (2019; 2021). The idea is that both operations act on the point $\boldsymbol{c} + \boldsymbol{V}\boldsymbol{\alpha}$ and leave the predicate variables $\boldsymbol{\alpha}$ untouched. The first identity holds because $\boldsymbol{W}(\boldsymbol{c} + \boldsymbol{V}\boldsymbol{\alpha}) + \boldsymbol{b} = (\boldsymbol{W}\boldsymbol{c} + \boldsymbol{b}) + (\boldsymbol{W}\boldsymbol{V})\boldsymbol{\alpha}$ for every $\boldsymbol{\alpha} \in P$. For the second, the point $\boldsymbol{x} = \boldsymbol{c} + \boldsymbol{V}\boldsymbol{\alpha}$ satisfies $\boldsymbol{H}\boldsymbol{x} \le \boldsymbol{g}$ if and only if $\boldsymbol{H}\boldsymbol{V}\boldsymbol{\alpha} \le \boldsymbol{g} - \boldsymbol{H}\boldsymbol{c}$, so both sides consist of the same points. □

B.2 Proof of Lemma 1 (piecewise with zero tails)

Proof of Lemma 1. The lemma restates the support and continuity properties of B-splines (de Boor, 1978) for the basis $i = -k, \dots, G-1$ in the KAN setting. We include a proof sketch for completeness. The idea is that ϕ is a finite sum of basis functions, each a continuous piecewise polynomial supported on $k+1$ consecutive knot cells. Each $B_{i,k}$ coincides on every knot cell with a polynomial of degree at most k, vanishes outside its support $[t_i, t_{i+k+1}]$, and is continuous because $k \geq 1$ and the knots are simple; a finite sum of such functions is continuous and coincides on each knot cell with a polynomial of degree at most k. The supports of $B_{-k,k}, \dots, B_{G-1,k}$ lie in $[t_{-k}, t_{G+k}]$, so ϕ vanishes outside the extended grid and, by continuity, at its two endpoints, that is, on both closed tails. For $k = 1$ and $w \equiv 0$ every piece of every edge is affine, so every edge is continuous piecewise affine, and a layer sums such functions of single inputs, which is again continuous piecewise affine. □

B.3 Proof of Lemma 2 (knot-cell case splits)

Proof of Lemma 2. The idea is that the pieces of $\mathcal{P}$ cover every input coordinate, and fixing one piece per input node fixes the polynomial of every edge. (i) $\Theta_J \subseteq \Theta$ by definition. Conversely, for $\boldsymbol{x} \in \Theta$ each x_p lies in some piece J_p of $\mathcal{P}$, which meets $[l_p, u_p]$ because x_p does, so $J \in R_1 \times \cdots \times R_n$ and $\boldsymbol{x} \in \Theta_J$. (ii) Each piece is a closed interval or half-line, so $x_p \in J_p$ is the conjunction of at most two halfspace constraints, and Lemma 5 appends one predicate row per halfspace. (iii) On a knot cell J_p the spline part of $\phi_{q,p}$ is a polynomial of degree at most k and on a tail it is the zero polynomial (Lemma 1); adding the base branch $w_{q,p}\mathrm{silu}$ gives the stated form. (iv) $\mathcal{P}$ has $G + 2k + 2$ pieces, so $r_p \leq G + 2k + 2$, and the piece assignments are the elements of $R_1 \times \cdots \times R_n$; each R_p depends only on $\mathcal{P}$ and on $[l_p, u_p]$, so their product does not depend on m. □

B.4 Proof of Theorem 1 (exact degree-1 reachability)

Proof of Theorem 1. The idea is that on a region star of a degree-1 no-base layer every edge is affine, so the layer is one affine map there, which Lemma 5 applies to the star without loss; an induction over the layers then keeps the input coordinates as the predicate variables. We show by induction on d that every entry $\langle \boldsymbol{c}, \boldsymbol{V}, P \rangle$ of $\mathcal{F}_d$ satisfies $P \subseteq \mathcal{B}$ and $L_d \circ \cdots \circ L_1(\boldsymbol{\alpha}) = \boldsymbol{c} + \boldsymbol{V}\boldsymbol{\alpha}$ for all $\boldsymbol{\alpha} \in P$, and that these polytopes P cover $\mathcal{B}$; for $d = D$ this is the claim, since then $N(\mathcal{B}) = \bigcup_i N(P_i) = \bigcup_i \Omega_i$. For $d = 0$ the entry $\langle \boldsymbol{0}, \boldsymbol{I}, \mathcal{B} \rangle$ has these properties. Let $\Theta = \langle \boldsymbol{c}, \boldsymbol{V}, P \rangle$ be an entry of $\mathcal{F}_{d-1}$ and Θ_J a nonempty region star of Θ for L_d. By Lemma 5, $\Theta_J = \langle \boldsymbol{c}, \boldsymbol{V}, P_J \rangle$ with $P_J = P \cap \{\boldsymbol{\alpha} : (\boldsymbol{c} + \boldsymbol{V}\boldsymbol{\alpha})_p \in J_p \ \forall p\} \subseteq P$. By Lemma 2(iii) with $k = 1$ and $w \equiv 0$, every edge is affine on Θ_J, $\phi_{q,p}(x_p) = s_{q,p}x_p + \beta_{q,p}$ with $s_{q,p} = \beta_{q,p} = 0$ on a tail, so $L_d(\boldsymbol{x}) = \boldsymbol{S}_J\boldsymbol{x} + \boldsymbol{b}_J$ on Θ_J with $(\boldsymbol{S}_J)_{q,p} = s_{q,p}$ and $(\boldsymbol{b}_J)_q = \sum_p \beta_{q,p}$, and Lemma 5 gives $L_d(\Theta_J) = \langle \boldsymbol{S}_J\boldsymbol{c} + \boldsymbol{b}_J, \boldsymbol{S}_J\boldsymbol{V}, P_J \rangle$. For $\boldsymbol{\alpha} \in P_J$ the induction hypothesis yields $L_d \circ \cdots \circ L_1(\boldsymbol{\alpha}) = \boldsymbol{S}_J(\boldsymbol{c} + \boldsymbol{V}\boldsymbol{\alpha}) + \boldsymbol{b}_J$, the affine map of this star. Finally, the region stars cover Θ by Lemma 2(i), so every $\boldsymbol{\alpha} \in P$ lies in some P_J, and the polytopes of the entries of $\mathcal{F}_d$ cover $\mathcal{B}$. □

B.5 Proof of Corollary 1

Proof of Corollary 1. The idea is that Theorem 1 turns the minimum of each margin over $\mathcal{B}$ into finitely many LPs whose minimizers are inputs. By Theorem 1 the polytopes P_i cover $\mathcal{B}$ and $N(\boldsymbol{x}) = \boldsymbol{c}_i + \boldsymbol{V}_i\boldsymbol{x}$ for $\boldsymbol{x} \in P_i$, so for each $j \neq y^*$, $\min_{\boldsymbol{x} \in \mathcal{B}} \big(N(\boldsymbol{x})_{y^*} - N(\boldsymbol{x})_j\big) = \min_i \mu_{ij}$, and each μ_{ij} is attained because P_i is a nonempty bounded polytope. (i) $N(\mathcal{B}) \subseteq \Pi$ holds if and only if every margin is positive on $\mathcal{B}$, that is, if and only if $\min_i \mu_{ij} > 0$ for all $j \neq y^*$. (ii) A minimizer $\boldsymbol{\alpha}^*$ of $\mu_{ij} < 0$ lies in $P_i \subseteq \mathcal{B}$, and its margin $N(\boldsymbol{\alpha}^*)_{y^*} - N(\boldsymbol{\alpha}^*)_j$ equals $\mu_{ij} < 0$. (iii) If $\min_{i,j} \mu_{ij} = 0$, some $\boldsymbol{x} \in \mathcal{B}$ has a margin equal to 0, so $N(\boldsymbol{x}) \notin \Pi$, while no margin is negative on $\mathcal{B}$. □

Case (iii) is not confined to a null set: if every coordinate range of $\mathcal{B}$ lies in a tail of $\mathcal{P}$, then every edge of a one-layer no-base KAN, hence every logit, is identically zero on $\mathcal{B}$ by Lemma 1, so every margin is 0 on all of $\mathcal{B}$.

B.6 Proof of Lemma 3 (tightest parallel bounds)

Proof of Lemma 3. Part (iii) is the characterization of the best uniform affine approximation of a convex function, proved for closed proper convex ϕ by Cheung & Li (2018) and derived for twice continuously differentiable convex ϕ by Mueller et al. (2021); we include a proof sketch for completeness. The idea is that the gap at slope s is the range of the residual $\phi(x)-sx$, and that for convex ϕ the best affine approximation has its largest error at both endpoints, which forces the chord slope. (i) The residual is continuous, so it attains its extrema on $[l,u]$, and $sx+o_1 \le \phi(x) \le sx+o_2$ on $[l,u]$ holds if and only if $[o_1,o_2]$ contains the range $[\underline{o}(s),\overline{o}(s)]$ of the residual. (ii) If $\phi(x)=sx+\beta$ on $[l,u]$, the residual is the constant β, so $\operatorname{gap}(s)=0$. (iii) Let ϕ be convex; the concave case follows by applying the result to $-\phi$. Since $\operatorname{gap}(s)=2\inf_{\beta\in\mathbb{R}}\max_{x\in[l,u]}|\phi(x)-sx-\beta|$, minimizing the gap over s is the problem of best uniform approximation of ϕ on $[l,u]$ by affine functions, which has a solution $x\mapsto s^*x+\beta^*$, and it suffices to show that $s^*=s_{\mathrm{c}}$. If ϕ is affine, the minimum 0 is attained only at its own slope, which is s_{c}. Otherwise the error $e(x)=\phi(x)-s^*x-\beta^*$ is convex with $E=\max_{[l,u]}|e|>0$, and by the Chebyshev equioscillation theorem e alternates between E and $-E$ at some points $\theta_1<\theta_2<\theta_3$ of $[l,u]$. Convexity rules out $e(\theta_2)=E$, which would give $E\le\max\{e(\theta_1),e(\theta_3)\}=-E$, so $e(\theta_1)=e(\theta_3)=E$ and $e(\theta_2)=-E$. Convexity also forces $\theta_1=l$: if $\theta_1>l$, then $\theta_1=\lambda l+(1-\lambda)\theta_2$ with $0<\lambda<1$ and $E=e(\theta_1)\le\lambda e(l)+(1-\lambda)e(\theta_2)\le\lambda E-(1-\lambda)E<E$, a contradiction. Symmetrically $\theta_3=u$, so $e(l)=e(u)$, that is, $\phi(u)-\phi(l)=s^*(u-l)$ and $s^*=s_{\mathrm{c}}$. □

B.7 Proof of Lemma 4 (quadratic relaxation gap)

Proof of Lemma 4. For twice continuously differentiable ϕ the pointwise bound below is classical (de Boor, 1978); we give the argument under the weaker hypothesis. The idea is that the residual at the chord slope is, up to a constant, ϕ minus its chord, which vanishes at both endpoints and is therefore at most $M_2(u-l)^2/8$ in absolute value; splitting $[l,u]$ at a zero of this function then bounds its range by the same constant. Let chord be the affine interpolant of ϕ at l and u and $e=\phi-\mathrm{chord}$; then $e(l)=e(u)=0$, e' is absolutely continuous with $e''=\phi''$ a.e., and $\operatorname{gap}(s_{\mathrm{c}})=\max e-\min e$, because e and the residual $\phi(x)-s_{\mathrm{c}}x$ differ by a constant.

Pointwise bound. Fix $x\in(l,u)$, let $\kappa=e(x)/((x-l)(x-u))$ and $\zeta(y)=e(y)-\kappa(y-l)(y-u)$. Then $\zeta(l)=\zeta(x)=\zeta(u)=0$, so by Rolle's theorem ζ' vanishes at some $\eta_1\in(l,x)$ and some $\eta_2\in(x,u)$, and since ζ' is absolutely continuous with $\zeta''=\phi''-2\kappa$ a.e., $0=\zeta'(\eta_2)-\zeta'(\eta_1)=\int_{\eta_1}^{\eta_2}(\phi''(y)-2\kappa)\,dy$, whence $2|\kappa|(\eta_2-\eta_1)\le M_2(\eta_2-\eta_1)$ and $|\kappa|\le M_2/2$. Therefore $|e(x)|\le\frac{M_2}{2}(x-l)(u-x)\le\frac{M_2(u-l)^2}{8}$ for all $x\in[l,u]$.

Gap bound. Since $e(l)=0$, $\min e\le 0\le\max e$. If one of them is 0, then $\operatorname{gap}(s_{\mathrm{c}})=\max|e|\le M_2h^2/8$. Otherwise $e(x^+)=\max e>0>\min e=e(x^-)$, and by continuity e has a zero x^0 strictly between x^+ and x^-, so x^+ and x^- lie in different members of $\{[l,x^0],[x^0,u]\}$, whose widths satisfy $d_1+d_2=u-l\le h$. On each of these subintervals e vanishes at both endpoints, so there it equals ϕ minus the affine interpolant of ϕ on that subinterval, and the pointwise bound gives $e(x^+)\le M_2d_i^2/8$ and $-e(x^-)\le M_2d_j^2/8$ with $\{i,j\}=\{1,2\}$. Hence $\operatorname{gap}(s_{\mathrm{c}})\le M_2(d_1^2+d_2^2)/8\le M_2(d_1+d_2)^2/8\le M_2h^2/8$.

Sharpness. For $\phi(x)=M_2x^2/2$ on $[-h/2,h/2]$ the chord slope is 0, and the residual ϕ has minimum 0 and maximum $M_2h^2/8$. □

The hypothesis of Lemma 4 holds on every knot cell and, for $k\ge 2$, across knots, since there ϕ' is continuous and piecewise smooth with bounded derivative, hence Lipschitz; it fails in general for degree-1 edges on intervals containing a knot, and hybrid mode relaxes only within knot cells.

B.8 Lipschitz sampling and derivative bounds

Lemma 6 (Uniform-grid Lipschitz bound)**.** *Let ϕ be Λ-Lipschitz on $[l,u]$, let $\nu\ge 1$ be an integer and $\xi_i=l+i(u-l)/\nu$ for $i=0,\dots,\nu$. Then* $\min_{[l,u]}\phi\ge\min_i\phi(\xi_i)-\frac{\Lambda(u-l)}{2\nu}$ *and* $\max_{[l,u]}\phi\le\max_i\phi(\xi_i)+\frac{\Lambda(u-l)}{2\nu}$.

Proof. This is Nesterov (2014) in dimension one with grid parameter ν, applied after rescaling $[l, u]$ to $[0, 1]$, which multiplies the Lipschitz constant by $u - l$. The idea is that no point of $[l, u]$ is farther than half a sample spacing from a sample. Every $x \in [l, u]$ is within $(u - l)/(2\nu)$ of some ξ_i, so $\phi(x) \geq \phi(\xi_i) - \Lambda|x - \xi_i| \geq \min_i \phi(\xi_i) - \Lambda(u - l)/(2\nu)$; the maximum case follows by applying this to $-\phi$. □

For the Lipschitz constants, interval arithmetic on the derivative polynomial gives $|\rho'(x)| \leq \sum_{i=1}^{k} i|\gamma_i|\tau_{\max}^{i-1}$ on a knot cell $[t_j, t_{j+1}]$, where γ_i are the coefficients of the polynomial ρ of that knot cell in the local variable $\tau = x - t_j$ and $\tau_{\max}$ is the largest $|\tau|$ on the knot cell. For the base branch, $\sup|\mathrm{silu}'| \approx 1.09984 < 1.1$. If ϕ is Λ-Lipschitz, the residual $\phi(x) - sx$ is $(\Lambda + |s|)$-Lipschitz, and Lemma 6 applied to it certifies offsets that satisfy the hypotheses of Theorem 2.

B.9 Proof of Theorem 2 (soundness of the lifted star)

Proof of Theorem 2. The idea is that the true edge values provide the lifted variables, $\delta_{q,p} = \phi_{q,p}(x_p) - s_{q,p}x_p$, and the hypotheses say exactly that this choice satisfies every constraint of $\Omega(\Theta)$. Let $\boldsymbol{x} \in \Theta$ and set $\delta_{q,p} = \phi_{q,p}(x_p) - s_{q,p}x_p$. Since $x_p \in [l_p, u_p]$, the hypotheses give $\delta_{q,p} \in [\underline{o}_{q,p}, \overline{o}_{q,p}]$ and $s_{q,p}x_p + \delta_{q,p} = \phi_{q,p}(x_p) \in [\underline{\phi}_{q,p}, \overline{\phi}_{q,p}]$, and $L(\boldsymbol{x})_q = \sum_p \phi_{q,p}(x_p) = \sum_p s_{q,p}x_p + \sum_p \delta_{q,p} = (\boldsymbol{S}\boldsymbol{x} + \boldsymbol{E}\boldsymbol{\delta})_q$, so $L(\boldsymbol{x}) \in \Omega(\Theta)$. For the composed statement we show by induction on d that $L_d \circ \cdots \circ L_1(\mathcal{B}) \subseteq \bigcup_{\Theta \in \mathcal{F}_d} \Theta$. For $d = 0$ the star $\langle \mathbf{0}, \boldsymbol{I}, \mathcal{B} \rangle$ is the set $\mathcal{B}$. If $\boldsymbol{x} \in \Theta \in \mathcal{F}_{d-1}$, then $\boldsymbol{x}$ lies in some region star Θ_J of the covering set of assignments, and $L_d(\boldsymbol{x})$ lies in $\Omega(\Theta_J)$ by the first part, applied to Θ_J with its own offsets and ranges, or in $L_d(\Theta_J)$, and both belong to $\mathcal{F}_d$. Hence $N(\mathcal{B}) \subseteq \bigcup_{\Omega \in \mathcal{F}_D} \Omega$, and if every $\Omega \in \mathcal{F}_D$ lies in Π, then so does $N(\mathcal{B})$. □

B.10 Proof of Proposition 1 (unsound extrapolation)

Proof of Proposition 1. The idea is a single edge whose last knot cell has nonzero slope: its affine continuation beyond the extended grid is positive there, whereas the edge itself is zero. Take the $[1, 1]$ KAN $N = \phi$ with $k = 1$, $w = 0$, a grid of G knot cells of width h, coefficient $c_{G-1} = -\sigma h$ for some $\sigma > 0$, and all other coefficients zero. On the last knot cell $[t_G, t_{G+1}]$ the only nonzero basis function is $B_{G-1,1}(x) = (t_{G+1} - x)/h$, so $\phi(x) = \sigma(x - t_{G+1})$ there, while $\phi \equiv 0$ on $[t_{G+1}, \infty)$ by Lemma 1; the affine continuation of this knot cell is $\tilde{\phi}(x) = \sigma(x - t_{G+1})$ for $x \geq t_{G+1}$. For $\mathcal{B} = [t_{G+1} + h,\, t_{G+1} + 2h]$ and $\Pi = \{y : y > \sigma h/2\}$ we obtain $N(\mathcal{B}) = \{0\}$, which is disjoint from Π, and $\tilde{N}(\mathcal{B}) = [\sigma h, 2\sigma h] \subseteq \Pi$. □

B.11 Lie-derivative discharge

Proposition 2 (Sound Lie-derivative discharge)**.** *Let $B(\boldsymbol{x}) = \sum_q \phi_q(z_q)$ with $z_q = \sum_p \psi_{q,p}(x_p)$ be a two-layer KAN of degree k, let $f : \mathbb{R}^n \to \mathbb{R}^n$ be a vector field, and let $X \subseteq \mathbb{R}^n$ be covered by finitely many boxes $\mathcal{B}_i$. Suppose that for every i there are intervals $[\underline{B}_i, \overline{B}_i] \supseteq B(\mathcal{B}_i)$, $[\underline{z}_{q,i}, \overline{z}_{q,i}] \supseteq z_q(\mathcal{B}_i)$ and $F_{p,i} \supseteq f_p(\mathcal{B}_i)$, and intervals $\Psi_{q,p,i}$ and $\Phi_{q,i}$ that contain $\psi'_{q,p}(x)$ for every $x \in \mathrm{pr}_p\mathcal{B}_i$ and $\phi'_q(z)$ for every $z \in [\underline{z}_{q,i}, \overline{z}_{q,i}]$ at which these derivatives exist, where $\mathrm{pr}_p\mathcal{B}_i$ is the p-th side of the box $\mathcal{B}_i$, and let $\Gamma_i = \sum_p \big(\sum_q \Phi_{q,i}\, \Psi_{q,p,i}\big) F_{p,i}$ be evaluated in interval arithmetic. If $\sup \Gamma_i \leq 0$ for every i with $0 \in [\underline{B}_i, \overline{B}_i]$, then $\mathcal{L}_f B(\boldsymbol{x}) = \nabla B(\boldsymbol{x}) \cdot f(\boldsymbol{x}) \leq 0$ for every $\boldsymbol{x} \in X$ with $B(\boldsymbol{x}) = 0$ at which all $\psi_{q,p}$ are differentiable at x_p and all ϕ_q at $z_q(\boldsymbol{x})$; for $k \geq 2$ these are all $\boldsymbol{x} \in X$ with $B(\boldsymbol{x}) = 0$.*

Proof of Proposition 2. The idea is that at a zero of B every factor in the chain-rule expansion of $\mathcal{L}_f B$ lies in its enclosure, and interval addition and multiplication are inclusion monotone. Let $\boldsymbol{x} \in X$ with $B(\boldsymbol{x}) = 0$ be a point at which the stated derivatives exist. Then $\boldsymbol{x} \in \mathcal{B}_i$ for some i with $0 = B(\boldsymbol{x}) \in [\underline{B}_i, \overline{B}_i]$, so $\sup \Gamma_i \leq 0$ by hypothesis. By the chain rule, $\partial B/\partial x_p(\boldsymbol{x}) = \sum_q \phi'_q(z_q(\boldsymbol{x}))\, \psi'_{q,p}(x_p)$, and since $x_p \in \mathrm{pr}_p\mathcal{B}_i$, $z_q(\boldsymbol{x}) \in [\underline{z}_{q,i}, \overline{z}_{q,i}]$ and $f_p(\boldsymbol{x}) \in F_{p,i}$ for every p and q, each factor lies in its enclosure, so $\mathcal{L}_f B(\boldsymbol{x}) = \sum_p \partial B/\partial x_p(\boldsymbol{x})\, f_p(\boldsymbol{x}) \in \Gamma_i$ and $\mathcal{L}_f B(\boldsymbol{x}) \leq \sup \Gamma_i \leq 0$. For $k \geq 2$ the spline part of every edge is $k - 1 \geq 1$ times continuously differentiable and the base branch is smooth, so the differentiability condition holds at every point. □

The enclosures are obtained as follows. The ranges of B and of the hidden nodes z_q over a box are coordinate bounds of the output stars of Theorem 1 or 2. The derivative of an edge's spline part is a spline of degree $k-1$ on the same knots, $\frac{d}{dx}\sum_{i=-k}^{G-1} c_i B_{i,k} = \sum_{i=-k}^{G} k\,\frac{c_i - c_{i-1}}{t_{i+k}-t_i} B_{i,k-1}$ with $c_{-k-1} = c_G = 0$ (de Boor, 1978), so its range over an interval is certified by Lemma 6 with the derivative bound of a polynomial of degree $k-1$; the base branch contributes $w\,\mathrm{silu}'(x)$ with $\mathrm{silu}' \in [-0.1, 1.1]$, sharpened on each input range. For no-base edges with $k = 1$ the derivative is piecewise constant with exact ranges per knot cell, and the enclosure covers the one-sided derivatives at the knots. The range of f is an interval extension of the plant dynamics composed with the star-certified controller range. The bound Γ_i is over-approximate, because the interval product decouples z_q from $\boldsymbol{x}$, and it tightens under box splitting.

B.12 Hinge decomposition of a degree-1 edge

By Lemma 1, a degree-1 no-base edge is continuous and piecewise affine with slope changes only at its knots and zero tails, so on any interval $[l, u]$ it has the form $\phi(x) = a_0 + \sum_{g=1}^{Z} a_g \max\{x - t_g, 0\}$ with Z hinges, which is the form in which a ReLU network represents it (Appendix C.9). The next proposition compares the native relaxations of ϕ with relaxations applied hinge by hinge. For an interval $I \subseteq \mathbb{R}$ and $a \in \mathbb{R}$ we write $aI = \{ay : y \in I\}$, sums of intervals are Minkowski sums, $\mathrm{gr}(f) = \{(x, f(x)) : x \in [l, u]\}$ is the graph of a function f on $[l, u]$, and $\mathrm{gap}_f(s) = \max_{x\in[l,u]}(f(x) - sx) - \min_{x\in[l,u]}(f(x) - sx)$ is the gap of Section 5.2 for a continuous f on $[l, u]$.

Proposition 3 (Native vs. hinge-wise relaxation)**.** *Let $l < u$, let $Z \geq 1$, let $t_1, \ldots, t_Z$ and $a_0, a_1, \ldots, a_Z$ be real numbers, let $r_g(x) = \max\{x - t_g, 0\}$ for $g = 1, \ldots, Z$, and let $\phi(x) = a_0 + \sum_{g=1}^{Z} a_g r_g(x)$ for $x \in [l, u]$. Then: (i) $\phi([l, u]) \subseteq a_0 + \sum_{g=1}^{Z} a_g\, r_g([l, u])$; (ii) if $C_1, \ldots, C_Z \subseteq \mathbb{R}^2$ are convex sets with $\mathrm{gr}(r_g) \subseteq C_g$ for every g, then $\mathrm{conv}\,\mathrm{gr}(\phi) \subseteq \{(x,\, a_0 + \sum_{g=1}^{Z} a_g y_g) : (x, y_g) \in C_g$ for $g = 1, \ldots, Z$, for some $y_1, \ldots, y_Z\}$; (iii) for all $s_1, \ldots, s_Z \in \mathbb{R}$, $\mathrm{gap}_\phi\big(\sum_{g=1}^{Z} a_g s_g\big) \leq \sum_{g=1}^{Z} |a_g|\,\mathrm{gap}_{r_g}(s_g)$.*

Proof of Proposition 3. The idea is that ϕ is an affine combination of the hinges, so an enclosure of each hinge, whether an interval, a convex set or a parallel strip, combines into an enclosure of ϕ. (i) For $x \in [l, u]$ and every g, $a_g r_g(x) \in a_g\, r_g([l, u])$, and a sum of elements of sets lies in the Minkowski sum of the sets. (ii) The set $K = \{(x, y_1, \ldots, y_Z) \in \mathbb{R}^{Z+1} : (x, y_g) \in C_g \text{ for every } g\}$ is convex, as an intersection of preimages of convex sets under linear maps, so its image D under the affine map $(x, y_1, \ldots, y_Z) \mapsto (x,\, a_0 + \sum_g a_g y_g)$, which is the right-hand side, is convex. For $x \in [l, u]$, $(x, r_g(x)) \in \mathrm{gr}(r_g) \subseteq C_g$ for every g, so $(x, \phi(x)) \in D$; thus D is a convex set containing $\mathrm{gr}(\phi)$, hence also $\mathrm{conv}\,\mathrm{gr}(\phi)$. (iii) Let $s = \sum_g a_g s_g$ and $e_g(x) = r_g(x) - s_g x$, so that $\phi(x) - sx = a_0 + \sum_g a_g e_g(x)$ on $[l, u]$. Bounding the maximum and the minimum over $[l, u]$ term by term and subtracting, $\mathrm{gap}_\phi(s) \leq \sum_g \big(\max_{[l,u]}(a_g e_g) - \min_{[l,u]}(a_g e_g)\big) = \sum_g |a_g|\,\mathrm{gap}_{r_g}(s_g)$, because $\max(ae) - \min(ae) = |a|(\max e - \min e)$ for every real a. □

Part (i) is the interval bound that a ReLU verifier computes on the converted network, part (ii) states that every hinge-wise convex relaxation, such as the triangle relaxation of each hinge, contains the convex hull of the graph, and part (iii) bounds the native parallel gap at the combined slope by the hinge-wise parallel gaps. The parallel pair of Lemma 3 is not the convex hull of the graph, so part (ii) does not order it against the hinge-wise triangle. When exactly one t_g lies in (l, u), every other hinge is affine on $[l, u]$ and its triangle degenerates to its graph, so the hinge-wise triangle is the hull of $\mathrm{gr}(\phi)$, which lies inside the parallel strip and, at the chord slope, has half its area. The CROWN relaxation measured in Appendix C.9 bounds each unstable hinge from below by one line instead and is not this hull. With the per-hinge chord slopes $s_g = (r_g(u) - r_g(l))/(u - l)$, which are the slopes of the DeepZ relaxation (Singh et al., 2018), $\sum_{g=1}^{Z} a_g s_g$ is the chord slope of ϕ, so part (iii) states that the native gap at the chord slope is at most $\sum_{g=1}^{Z} |a_g|\,\mathrm{gap}_{r_g}(s_g)$, the gap of the hinge-wise DeepZ relaxation.

Table 3: **The benchmarks.** 26 trained KANs with their properties and concrete query sets. Instance: the suffix `pwl` marks the degree-1 no-base network of a task and `cubic` its degree-3+SiLU network. Source: Ex. 1 is Example 1 (the pendulum) of KAN4CBC; UCI is the UCI Machine Learning Repository, and WDBC is its Breast Cancer Wisconsin (Diagnostic) dataset. Widths: the layer widths. Params: the number of trained parameters (spline coefficients and base weights). k/G: spline degree k and number G of knot cells of the grid in each layer. +b: every edge includes the base branch $w_{q,p}\,\mathrm{silu}(x)$ of Equation 1; without +b the network is no-base ($w_{q,p} = 0$). Held-out: accuracy for classifiers, RMSE of the standardized output for regressors. Property: rob., r radii is local robustness of the network's prediction at r radii around each test point; P1 and P2: the pendulum barrier is negative on the initial set and positive on the unsafe set; P3: the pendulum controller satisfies $|u_{\mathrm{net}}| \leq 14$; R1: one Euler step of the Lotka–Volterra field maps the initial box into the target box T; R2: the Euler step keeps the prey population positive; V1: $|N| \leq 1$ for Feynman I.16.6; Q1: a certified enclosure of the output of the pykan example over $[-1, 1]^2$. #: queries or boxes. Input: how the network's input is formed from the data; affine: a fixed map $(x - x_{\mathrm{c}})/x_{\mathrm{h}}$ per coordinate, with center x_{c} and half-span x_{h}; std.: features standardized to zero mean and unit variance; pixels/255: pixel intensities divided by 255.

Instance	Source	Widths	Params	k/G	Held-out	Property	#	Input
`e1-annulus-cubic`	synthetic annulus	[2, 6, 2]	192	3/4+b	99.3%	rob., 6 radii	660	affine
`e1-annulus-pwl`	synthetic annulus	[2, 6, 2]	120	1/4	99.2%	rob., 6 radii	660	affine
`cs5-moons-cubic`	two-moons	[2, 8, 2]	288	3/5+b	100.0%	rob., 5 radii	500	affine
`cs5-moons-pwl`	two-moons	[2, 8, 2]	288	1/8	100.0%	rob., 5 radii	500	affine
`cs5b-tab4-cubic`	Gaussian, 4 features	[4, 6, 3]	378	3/5+b	94.3%	rob., 3 radii	120	affine
`cs5b-tab4-pwl`	Gaussian, 4 features	[4, 6, 3]	378	1/8	94.0%	rob., 3 radii	120	affine
`drybean-cubic`	Dry Bean (UCI)	[16, 6, 7]	1,242	3/5+b	92.4%	rob., 2 radii	160	std.
`drybean-pwl`	Dry Bean (UCI)	[16, 6, 7]	1,242	1/8	90.7%	rob., 2 radii	160	std.
`magic-cubic`	MAGIC (UCI)	[10, 10, 2]	1,080	3/5+b	85.7%	rob., 2 radii	160	std.
`magic-pwl`	MAGIC (UCI)	[10, 10, 2]	1,080	1/8	81.6%	rob., 2 radii	160	std.
`wdbc-cubic`	WDBC (UCI)	[30, 6, 2]	1,728	3/5+b	99.1%	rob., 2 radii	160	std.
`wdbc-pwl`	WDBC (UCI)	[30, 6, 2]	1,728	1/8	97.4%	rob., 2 radii	160	std.
`mnist-d1-cubic`	MNIST	[784, 10]	70,560	3/5+b	93.0%	rob., 4 radii	200	pixels/255
`mnist-d1-pwl`	MNIST	[784, 10]	47,040	1/5	92.3%	rob., 4 radii	200	pixels/255
`fashion-fckan-cubic`	Fashion-MNIST	[784, 10]	70,560	3/5+b	85.6%	rob., 4 radii	200	pixels/255
`mnist-depth2-cubic`	MNIST, depth 2	[784, 16, 10]	114,336	3/5+b	93.7%	rob., 2 radii	50	pixels/255
`cs1-barrier-cubic`	KAN4CBC Ex. 1	[2, 5, 1]	135	3/5+b	0.0045	P1, P2	65	affine
`cs1-barrier-pwl`	KAN4CBC Ex. 1	[2, 5, 1]	255	1/16	0.0062	P1, P2	65	affine
`cs1-ctrl-cubic`	KAN4CBC Ex. 1	[2, 5, 1]	135	3/5+b	0.0024	P3	64	affine
`cs1-ctrl-pwl`	KAN4CBC Ex. 1	[2, 5, 1]	255	1/16	0.0016	P3	64	affine
`cs2-toy-cubic`	pykan example	[2, 5, 1]	135	3/5+b	0.0051	Q1	36	affine
`cs2-toy-pwl`	pykan example	[2, 5, 1]	255	1/16	0.0136	Q1	36	affine
`cs3-lv-cubic`	KAN-ODE	[2, 10, 2]	360	3/5+b	0.0035	R1, R2	21	affine
`cs3-lv-pwl`	KAN-ODE	[2, 10, 2]	680	1/16	0.0061	R1, R2	21	affine
`cs4-feynman-cubic`	Feynman I.16.6	[2, 5, 1]	135	3/5+b	0.0040	V1	64	affine
`cs4-feynman-pwl`	Feynman I.16.6	[2, 5, 1]	255	1/16	0.0062	V1	64	affine

C Additional Results

This appendix covers the benchmark and protocol (C.1), the full results matrix (C.2), the structural and relaxation experiments (C.3), the case studies (C.4 and C.5), the 784-input experiments (C.6), the family comparison (C.7), the native α-CROWN and β-CROWN members (C.8), the ReLU conversion (C.9), the accuracy price of the verifiable regime (C.10), the negative results (C.11) and the failed checks (C.12); experiments carry the identifiers of their accompanying scripts (E1, E2, ...). All experiments ran on one workstation using an Intel Core Ultra 9 285K CPU (24 cores), 64 GB of RAM and Windows 11. They used Python 3.10.11, NumPy 2.2.6, SciPy 1.15.3 (HiGHS for the LPs and MILPs) and, for α-CROWN and β-CROWN, PyTorch 2.12.0 on the CPU with one thread per process; the comparison with the ReLU conversion (E18, Appendix C.9) used α, β-CROWN 0.7.0 and auto_LiRPA 0.7.2 with PyTorch 2.11.0 under Python 3.11.9.

C.1 Benchmark and protocol

Instances. Table 3 lists the 26 released networks. Every task except Fashion-MNIST and depth-2 MNIST is trained in both regimes from the same data and template, the three 784-input tasks use the efficient-kan parameterization attacked by Alter et al. (2025), and the depth-2 network is the repaired network of Appendix C.6. Each instance has a checkpoint with its grid metadata (a KAN

is not specified by its weights alone, Section 4), a property file with the concrete query set that Table 2 counts and the input and output normalization (so that every threshold is stated in physical units), and the reference verdict and bound of every member of Table 4 on every query. Re-running the reference member from the released files alone reproduces the recorded bounds exactly on 77 queries, and the released weights are those of the verified networks, as checked by hash or, for the nine networks without a recorded hash, by retraining at the recorded seed.

Queries and boxes. A robustness query takes as its label y^* the network's prediction at a test point and uses a fixed ℓ_∞ radius in the network's input coordinates. The points are correctly classified except on the Gaussian and UCI tasks, where they are drawn from held-out data regardless of their labels. The query sets are 120 annulus points at $\varepsilon \in \{0.02, 0.05, 0.10, 0.15, 0.20\}$ and 60 at 0.30; 100 two-moons points at five radii from 0.02 to 0.20; 40 Gaussian-task points at $\varepsilon \in \{0.01, 0.02, 0.05\}$; 80 points per UCI dataset at the two rungs of the certified-radius ladder of Appendix C.6 closest to 0.01 and 0.02 standard deviations; and 50 images per 784-input network at $\varepsilon \in \{1, 2, 4, 8\}/255$, clipped to $[0, 1]$ (25 images at $\{1, 2\}/255$ at depth 2). The box properties of Table 3 are checked on box covers of their input regions, and the pykan enclosure, on a 36-box cover of $[-1, 1]^2$, is scored by its mean width ratio. For the Euler step (R1), IBP, plain CROWN and PWA-MILP use the non-relational interval map $\boldsymbol{x} + \Delta t\,[\underline{N}, \overline{N}]$, α-CROWN and β-CROWN add the input term before concretizing, and star sets carry the affine map exactly.

Protocol. All members run in float64 and share the same certified per-edge ranges and Lipschitz constants. The number of sampling intervals of Lemma 6 is $\nu = \lceil \Lambda(u - l)/(2 \cdot 10^{-10}) \rceil$, clipped to $[8, 8192]$, so that the sampling margin is 10^{-10} unless the clip binds. Each star-set run uses the split cap $S_{\max}$ of its experiment (4,096 to 200,000 stars), which none of the 4,020 star-set exact and hybrid queries and boxes behind Table 2 reached. PWA-MILP is our reimplementation of Schwartz et al. (2026): it replaces their optimally allocated pieces by a uniform K-piece interpolant per edge, with independently re-verified errors, and may therefore be looser than their abstraction. Unless stated otherwise it uses $K{=}4$, a node limit of 1,000 per class-pair solve on the 10- to 30-input networks and a 60 s limit elsewhere, and a capped solve reports its dual bound, which is sound. The CROWN column of Table 4 is plain CROWN, and its CR. best column, like the CROWN column of Table 8 on robustness rows, keeps per query the better of plain CROWN and CROWN with final-layer elision (Appendix C.7). The attainable value of a query is the hull of the network's outputs on 4,000 to 20,000 sampled inputs together with, on robustness queries, the best iterate of a PGD-style attack. The width ratio of a certified interval is its width divided by that of the attainable value, and no certified bound in the matrix exceeds its attainable value by more than 7.5×10^{-13}. The count F of Table 2 pools the queries with a counterexample re-executed on the network from any source, whether a star-set LP minimizer, the minimizer of a CROWN-style linear bound, an exact-MILP or PWA-MILP argmin, or the attack, so a member's own FALSIFIED count can be smaller; Table 4 gives both.

C.2 THE FULL MEMBER-BY-INSTANCE MATRIX

PWA-MILP at $K{=}4$ falls short of the ceiling on `magic-cubic` and `drybean-cubic` because of its node limit: 1,063 of its solves were capped, in 405 of its 2,010 queries and boxes, and 5 capped solves returned no finite dual bound (Appendix C.12). The exact MILP and star-set exact agree on the verdict of all 2,010 degree-1 queries and boxes, with margin differences of at most 7.77×10^{-5} (relative 6.1×10^{-5}) and box-enclosure differences of at most 3.43×10^{-4}, both due to HiGHS's default relative MIP gap of 10^{-4}, and the exact MILP is faster, by $3\times$ (Dry Bean) to $49\times$ (MAGIC) in median time per query on the annulus, two-moons and UCI networks (see also Appendix C.7). Figure 3 shows the robustness rows of Table 2 with 2 to 30 inputs per network and radius: β-CROWN reaches the ceiling at every radius of all twelve networks, as do star-set exact and the exact MILP on the six degree-1 networks, whereas star-set hybrid falls short only on the degree-3 annulus, Dry Bean and MAGIC networks, by at most three queries per radius.

C.3 STRUCTURAL CLAIMS AND RELAXATION

Exactness (E1). On the degree-1 annulus network the margin width ratio of star-set exact is 1.019–1.027 per radius, yet 57% of its queries lie on the certified-equals-attained diagonal to within 10^{-6} and none lies above it by more than 6.8×10^{-14}, so the residual ratio is the sampling floor of

Table 4: **The full member × instance matrix** behind Table 2, except its CROWN column, which is β-CROWN (Table 8). Each cell is the member's own Robust/Falsified/Unknown count (queries) or Proved/–/Unproved (boxes); pykan cells are mean width ratios. F is the pooled count of re-executed counterexamples. approx, exact/hyb.: star-set modes. The last column is the median seconds per query of the star-set member (exact on degree-1 rows, hybrid on degree-3 rows). ‡: queries with a PWA-MILP solve stopped at the node limit (sound dual bound).

Instance	#	F	IBP	IBP+el.	CROWN	CR. best	approx	exact/hyb.	MILP/PWA$_4$	s/q
degree 1, no base										
el-annulus-pwl	660	220	250/220/190	251/220/189	322/220/118	323/220/117	349/220/91	440/220/0	440/220/0	0.29
cs5-moons-pwl	500	5	422/5/73	424/5/71	438/5/57	439/5/56	471/4/25	495/5/0	495/5/0	0.52
cs5b-tab4-pwl	120	4	111/4/5	112/4/4	115/4/1	115/4/1	115/4/1	116/4/0	116/4/0	0.069
drybean-pwl	160	13	103/12/45	129/12/19	120/12/28	134/12/14	138/13/9	147/13/0	147/13/0	0.57
magic-pwl	160	39	83/39/38	87/39/34	104/39/17	105/39/16	106/39/15	121/39/0	121/39/0	1.57
wdbc-pwl	160	6	152/6/2	153/6/1	154/6/0	154/6/0	154/6/0	154/6/0	154/6/0	0.35
mnist-d1-pwl	200	143	35/143/22	57/143/0	57/143/0	57/143/0	57/143/0	·	·	·
cs1-barrier-pwl	65	0	60/0/5	·	63/0/2	·	64/0/1	65/0/0	65/0/0	0.25
cs1-ctrl-pwl	64	0	63/0/1	·	64/0/0	·	64/0/0	64/0/0	64/0/0	2.02
cs2-toy-pwl	36	–	1.270	·	1.473	·	1.206	1.001	1.001	3.34
cs3-lv-pwl	21	0	21/0/0	·	21/0/0	·	21/0/0	21/0/0	21/0/0	2.36
cs4-feynman-pwl	64	0	64/0/0	·	63/0/1	·	64/0/0	64/0/0	64/0/0	1.84
degree 3 + SiLU										
el-annulus-cubic	660	209	407/209/44	408/209/43	423/209/28	423/209/28	435/209/16	450/209/1	441/209/10	0.70
cs5-moons-cubic	500	9	483/9/8	483/9/8	487/9/4	487/9/4	489/9/2	491/9/0	491/9/0	0.40
cs5b-tab4-cubic	120	5	113/5/2	114/5/1	115/5/0	115/5/0	115/5/0	115/5/0	115/5/0	0.11
drybean-cubic	160	3	146/3/11	151/3/6	155/3/2	157/3/0	155/3/2	156/3/1	152/3/5‡	0.51
magic-cubic	160	13	121/13/26	122/13/25	139/13/8	140/13/7	141/13/6	142/13/5	132/13/15‡	1.09
wdbc-cubic	160	2	158/2/0	158/2/0	158/2/0	158/2/0	158/2/0	158/2/0	158/2/0‡	0.63
mnist-d1-cubic	200	70	99/70/31	130/70/0	128/70/2	129/70/1	129/70/1	·	·	·
fashion-fckan-cubic	200	102	84/102/14	98/102/0	98/102/0	98/102/0	98/102/0	·	·	·
mnist-depth2-cubic	50	3	38/3/9	44/3/3	47/3/0	47/3/0	47/3/0	·	·	·
cs1-barrier-cubic	65	0	59/0/6	·	65/0/0	·	65/0/0	65/0/0	65/0/0	0.12
cs1-ctrl-cubic	64	0	63/0/1	·	64/0/0	·	64/0/0	64/0/0	64/0/0	0.16
cs2-toy-cubic	36	–	1.553	·	1.313	·	1.125	1.069	1.038	0.32
cs3-lv-cubic	21	0	21/0/0	·	21/0/0	·	21/0/0	21/0/0	21/0/0	0.28
cs4-feynman-cubic	64	0	57/0/7	·	60/0/4	·	61/0/3	64/0/0	64/0/0	0.20

Table 5: **Certified fractions at 784 inputs**, star-set approx: certified/images, Wilson 95% interval, and in parentheses the PGD ceiling (images the attack does not break), per radius $\varepsilon \cdot 255$.

Network	1	2	4	8
MNIST [784, 10]	49/50 [0.90, 1.00] (49)	45/50 [0.79, 0.96] (45)	28/50 [0.42, 0.69] (28)	7/50 [0.07, 0.26] (8)
MNIST [784, 10], degree 1	37/50 [0.60, 0.84] (37)	17/50 [0.22, 0.48] (17)	3/50 [0.02, 0.16] (3)	0/50 [0.00, 0.07] (0)
Fashion [784, 10]	43/50 [0.74, 0.93] (43)	34/50 [0.54, 0.79] (34)	20/50 [0.28, 0.54] (20)	1/50 [0.00, 0.10] (1)
MNIST [784, 16, 10]	24/25 [0.80, 0.99] (24)	23/25 [0.75, 0.98] (23)	–	–

the attainable range rather than slack in the bound. The mean margin width ratio is 1.63 for star-set approx on the same network, and 1.10 for hybrid and 1.27 for approx mode on the degree-3 network. Figure 4 shows the reachable sets of one query.

The quadratic law (E2). Figure 5 measures Lemma 4. Within one knot cell the log-log slope of the gap against the width of the relaxed interval is 1.92 on a random degree-3 edge and 1.96 on an edge of the trained annulus network; at the finest widths the gap reaches the Lipschitz-sampling margin, and the margin check of the sweep fails there. End to end, the hybrid gap has slope 1.84 over $G \geq 8$, where the hybrid:approx ratio tracks $(h/W)^2$ with a log-correlation of 0.9992; the fit over $G \geq 4$ gives 1.46, outside its registered window, because the B-spline fits at G=4 and 6 have not converged (maximum fit error 0.061 and 0.065 against 5×10^{-5} at G=32). The chord slope is suboptimal by at most 3.5×10^{-15} on degree-2 knot cells (Lemma 3), or 2.7×10^{-2} with the sampling margin of the implementation, and by up to 53% on degree-3 knot cells; no relaxation among 1,622 audited against dense evaluation is violated by more than 4.4×10^{-16}.

Slope optimization (E11b). Optimizing the slope of the parallel pair per edge and knot cell, with the chosen slope re-certified, beats the chord by up to 52.4%, but only on 1,038 of the 22,351 (edge, knot cell) pairs of 120 degree-3 annulus queries and the 65 degree-3 barrier boxes; end to end it removes 0.35% of the relaxation slack and changes no verdict (Table 7). The chord slope is therefore the right default for hybrid mode, whose knot-cell splits already leave a gap of order h^2 (Lemma 4).



Figure 3: **Local robustness per radius on the classifier benchmarks with 2 to 30 inputs.** Each pair of panels is one robustness task of Table 2, with the UCI row split into its three datasets; the header gives the task and its layer widths, the left panel shows the degree-1 no-base network and the right panel the degree-3+SiLU network. At each ℓ_∞ radius ε, in the network's input coordinates (training standard deviations on the UCI tasks), the markers give the fraction of queries that each member of Table 2 proves robust, from left to right in the table's column order; IBP uses final-layer elision, CROWN is β-CROWN and PWA-MILP uses $K{=}4$ pieces. The grey bar is the ceiling, the fraction of queries without a re-executed counterexample, and the tint above it is the fraction with one. Per radius there are 120 queries on the annulus (60 at $\varepsilon{=}0.3$), 100 on two-moons, 40 on Gaussian-4 and 80 on each UCI dataset; summed over the radii and datasets, the counts are those of Table 2. Queries are labeled with the network's prediction, and the Gaussian and UCI points are drawn regardless of their true labels, so on these tasks the fractions are not certified accuracies. Vertical axes are truncated and shared within a task. The 784-input networks are shown in Figure 2(e).

This does not carry over to α-CROWN, whose separate lower and upper slopes, optimized over the whole range of each input node against the objective, prove up to 55 more queries on a row than CROWN without slope optimization (Table 8).

Structural scaling (E3). For $[2, m]$ and $[2, 6, m]$ KANs with $m \in \{2, \ldots, 64\}$ (Figure 7), the output star count is exactly constant in m at every query point, while the number of nonlinear edges grows from 24 to 396 and the wall time varies by at most $1.10\times$, as part (iv) of Lemma 2 predicts. The count grows instead with the grid (a median of 4 stars at $G{=}2$, and 9,939 stars in 115 s on the single query run at $G{=}32$), with the depth (a median of 1.5 stars at one layer and 318.5 at six layers on $[2, 4, \ldots, 4, 2]$ networks) and with the input width (first-layer regions $2, 4, 8, 16$ at $n = 1, \ldots, 4$, exactly $\prod_p r_p$). On 23 audited queries the certified hull contains the sampled outputs, with width ratio at most 1.0204. Per star, the split costs $2n$ bounding LPs and one emptiness LP per candidate region, and the layer map one $m \times n$ matrix product, the only term that grows with m.

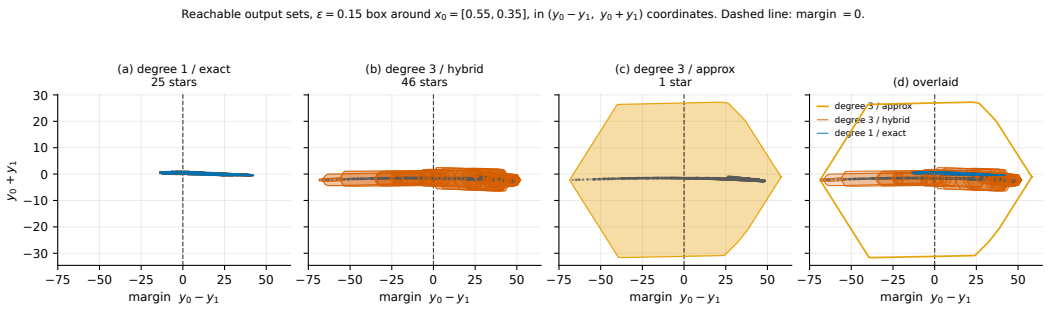


Figure 4: **Reachable output sets of the trained annulus KANs** over one input box ($x_0 = [0.55, 0.35]$, $\varepsilon = 0.15$), in (logit difference, logit sum) coordinates. Dots are sampled outputs and the dashed line is the decision boundary. (a) For the degree-1 no-base network in `exact` mode, the union of 25 stars is the image itself, so the polygons follow the samples closely. (b) For the degree-3+SiLU network, `hybrid` mode splits at the knots and relaxes within each knot cell. (c) `approx` mode uses one relaxation over the whole range of the same network. (d) In the overlay the hybrid set nests inside the approx set, which is the geometric statement of Theorem 2 and Lemma 4; the exact set of (a) belongs to a different network and is shown for scale. The same query, with the box of IBP and the exact MILP's number, is Figure 2(b).

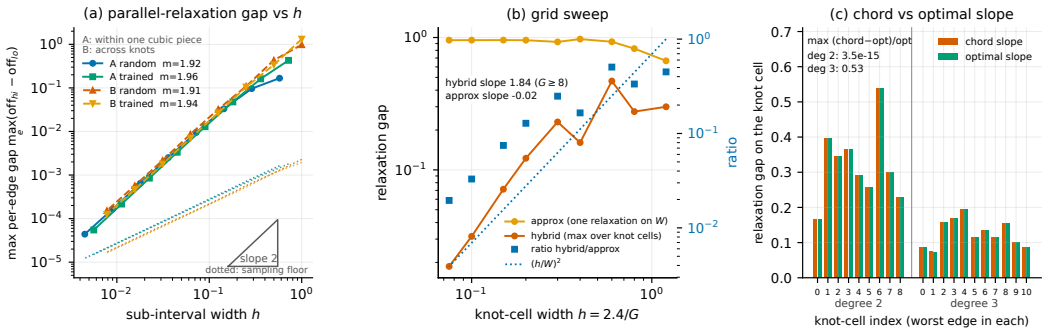


Figure 5: **The quadratic relaxation law, measured.** (a) The per-edge parallel-relaxation gap is plotted against the width $u - l$ of the relaxed interval (in panel (b), against the knot-cell width $h = (b - a)/G$), within one cubic piece (A) and across knots (B), for random and trained edges. The log-log slopes are 1.91–1.96 against the slope-2 reference line, and the dotted lines mark the Lipschitz-sampling measurement floor. (b) In the end-to-end grid sweep, the hybrid gap follows slope 1.84 for $G \geq 8$ while the approx gap stays flat, and the hybrid:approx ratio is shown against the worst-case $(h/W)^2$ line of Lemma 4. (c) The chord slope is compared with the numerically optimal slope per knot cell. The two are indistinguishable at degree 2 (Lemma 3), and the chord is up to 53% suboptimal on degree-3 knot cells.

The zero-tail trap (E4). On a hand-built $[2, 1]$ KAN and a box beyond the extended grid in its first coordinate, a propagator that extrapolates the boundary knot cells certifies $[0.21, 1.31]$ and declares a threshold property ROBUST, whereas star-set exact certifies the true range $[-0.15, 0.05]$ and a re-executed counterexample attains -0.15; inside the grid the two agree to 10^{-9}, so the tail rule is the sole cause. On the trained degree-1 annulus network and a box that sends five of its six hidden nodes beyond the extended grid of layer 2, the verifier stays sound and returns FALSIFIED with a re-executed counterexample, at width ratio 1.008.

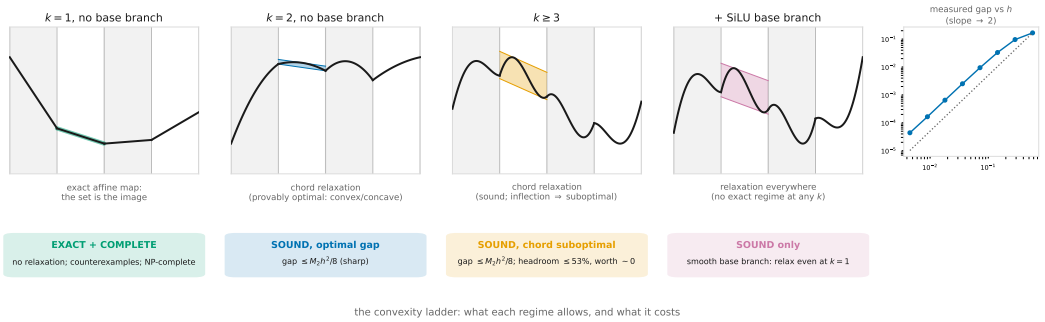


Figure 6: **The convexity ladder.** The knot cells are schematic, and the inset is drawn from measured rows of the relaxation-gap sweep. At $k{=}1$ the knot cells are affine, and the analysis is exact and complete. At $k{=}2$ every knot cell is convex or concave, so the chord-slope relaxation is provably optimal (Lemma 3). At $k{\geq}3$ a knot cell can inflect internally, so the chord is sound but suboptimal (for the shared slope of the parallel pair the measured headroom is $\leq 53\%$ per knot cell and worth ≈ 0 end to end, Appendix C.3; the separate slopes of α-CROWN are measured in Appendix C.8). With the base branch (+SiLU), the smooth silu removes the exact regime at every degree. In all relaxed regimes the gap obeys the quadratic law $\leq M_2 h^2/8$ (Lemma 4), and the inset shows the measured slope-2 line. End to end, no global convexity is ever needed. The global shape of an edge is carried by the *disjunction* over knot cells (unions of stars are non-convex), per-interval soundness rests only on continuity plus a derivative bound (Lemma 6), and the composition of layers into the network is a containment argument (Theorem 2). Therefore, non-convexity costs stars or width, but never soundness.

C.4 Case studies

CS1: the KAN4CBC pendulum barrier and controller (Figure 8). The inverted pendulum of Zhang et al. (2025a) (their Ex. 1), $\dot{x}_1 = x_2$, $\dot{x}_2 = 9.8 \sin x_1 + u_{\mathrm{net}}(\boldsymbol{x})$ on $[-\pi/4, \pi/4]^2$, uses $[2, 5, 1]$ KANs as barrier B_{net} and controller u_{net}, which we retrain in both regimes. Both sign conditions hold, $B_{\mathrm{net}} < 0$ on the initial set and $B_{\mathrm{net}} > 0$ on all 64 boxes that cover the unsafe region; on the initial set the certified maximum of B_{net} is -0.0064 in hybrid mode and -0.0107 in exact mode (degree 1), against -0.0105 for the design barrier. The actuator bound $|u_{\mathrm{net}}| \leq 14$ is proved on all 64 boxes, whereas the deliberately tight bound 13.0 fails on 2 boxes, exactly where sampling attains 13.2, so the certificate is not vacuous. Exactness costs time rather than width: on these boxes the width ratios of exact, hybrid and approx mode are 1.000, 1.031 and 1.057, at 2.27, 0.16 and 0.03 s per box.

CS2: the KAN "hello world." For the target $\exp(\sin(\pi x_1) + x_2^2)$ of Liu et al. (2025b), the certified enclosure $[0.3620, 7.3049]$ of the degree-1 $[2, 5, 1]$ KAN over $[-1, 1]^2$ is its exact range rounded outward: exact mode returns inputs at which the network attains both ends, whereas sampling attains only $[0.3621, 7.3048]$, and the distance to the analytic range $[e^{-1}, e^2]$ is fit error. Monotone ordering (Section 3) is checked on 7 pairs of consecutive x_2-slabs at $x_1 \in [-0.2, 0.2]$: the certified maximum of N over the first slab must lie below its certified minimum over the second. With touching slabs the ordering is false for the target itself (best true gap -1.36), and all 21 checks in three configurations correctly return not proved. On separated slabs of x_1 half-width 0.005, where the ordering holds, degree-1 exact mode and degree-3 hybrid and approx modes each prove all 7 pairs (Figure 1). Figure 9 shows both slab families, the certified range of every slab and the certified gap of every pair in each configuration, against the sampled values and the target.

CS3: KAN-ODE Lotka–Volterra. For a KAN vector field N in the setting of Koenig et al. (2024), star sets carry the Euler step $\boldsymbol{x} \mapsto \boldsymbol{x} + \Delta t\, N(\boldsymbol{x})$ exactly. Its image of the initial box is proved inside the target box with width ratio 1.0000 in exact mode (degree 1) and 1.002 in hybrid mode (degree 3), and one-step prey positivity is proved on all 20 sub-boxes in every mode.

Multi-step reach tubes (E14). A reach tube of 30 steps with $\Delta t{=}0.01$ from the initial box $[0.9, 1.1]^2$ is computed in exact mode on the degree-1 network and in hybrid and approx mode

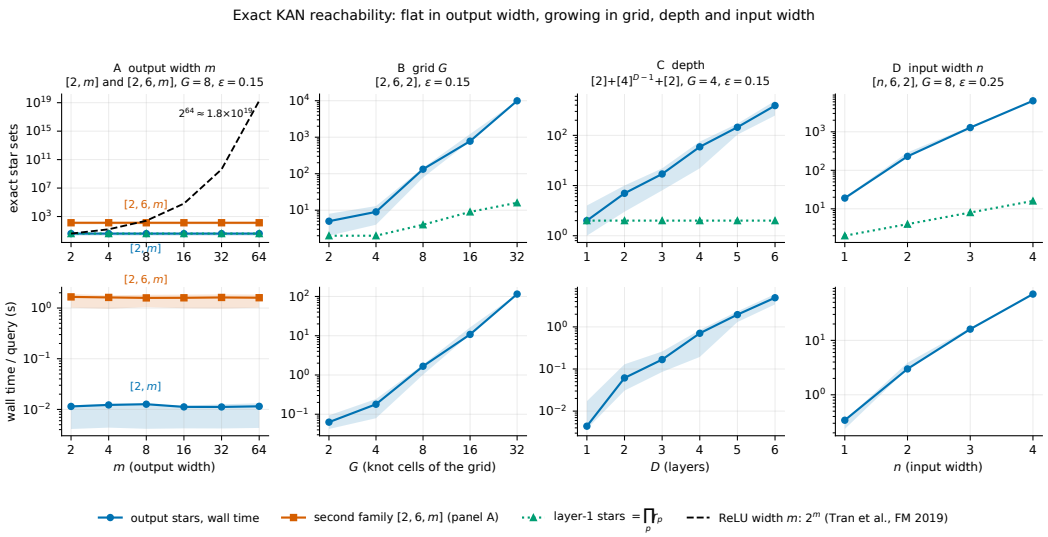


Figure 7: **Where the cost of star-set exact lives.** Both axes are logarithmic. The top row shows the number of output stars and the bottom row the seconds per query. (A) As the output width m grows from $m{=}2$ to 64, the star count and the time stay flat while the number of nonlinear edges grows $16\times$. The dashed line is the 2^m worst case of exact ReLU star analysis on a generic MLP; a ReLU network converted from a KAN has axis-aligned hinges and does not follow it (Appendix C.9). (B–D) The quantities that do drive the cost are the grid G, the depth, and the input width. The first-layer star count (green) is exactly $\prod_p r_p$, as part (iv) of Lemma 2 predicts.

on the degree-3 network, with the stars merged into their box hull after every step, which keeps the tube sound but no longer exact. The tube of the exact configuration reaches the two target sets at steps 10 and 26, avoids two unsafe boxes at every step, refutes a third avoid claim with a counterexample trajectory re-executed through the network, and at step 30 is 6.0% and 8.3% wider than the sampled hull. The tube certifies the network, not the physical system: none of 2,000 sampled network-Euler trajectories leaves a tube by more than 2 ULP, whereas the physical-Euler trajectories record 24 escapes, summed over steps and runs, which measure the model error of the network. Without merging (Table 7) only the approx configuration completes: exact mode stops at step 2 with more than 400 stars, and hybrid mode aborts at step 15 on a float64 NaN from a failed coordinate LP. A continuous-time certificate, which inflates each step by at most 6.0×10^{-4} using the derivative-spline Jacobian, covers times up to 0.21 on the exact configuration and no step on the hybrid configuration, where its inflate-and-check test fails. Figure 10 shows the one-step properties R1 and R2 in both regimes, the tube of the exact configuration with its target and unsafe boxes, and the width ratio of each tube per step, with and without merging.

CS4: Feynman I.16.6 (Udrescu & Tegmark, 2020) (relativistic velocity addition). KANs fit to $(x_1+x_2)/(1+x_1x_2)$ as in Liu et al. (2025b) must satisfy $|N| \leq 1$. For the $[2, 5, 1]$ KANs, exact and hybrid modes prove this on all 64 boxes, with certified maximum 0.9789 in exact mode, whereas approx mode on the degree-3 network proves 61: on the remaining boxes the relaxation, not the network, crosses 1 (certified maximum 1.0267, against 0.9883 in hybrid mode). Figure 11 shows these bounds box by box for every member, together with the width ratio of every box of the CS2 enclosure. Given only a derivative bound for Gaussian radial basis functions, the parallel relaxation of Lemma 3 also applies to a FastKAN-style $[2, 5, 1]$ KAN (Li, 2024) (CS4b): in approx mode its enclosures are sound on all 64 boxes and prove $|N| \leq 1$ on 61 (certified maximum 1.0798), so the relaxation is not specific to B-splines.

CS5: certified classification accuracy. Certified accuracy, the fraction of test points that are correctly classified *and* provably robust (Heiderich et al., 2025), is 95–100% on two-moons in every configuration except degree-1 approx (75% at $\varepsilon{=}0.2$), and 80–85% on the Gaussian task, against clean accuracies of 100% and 85–87.5% on the same points. Exact mode decides all 620 of its

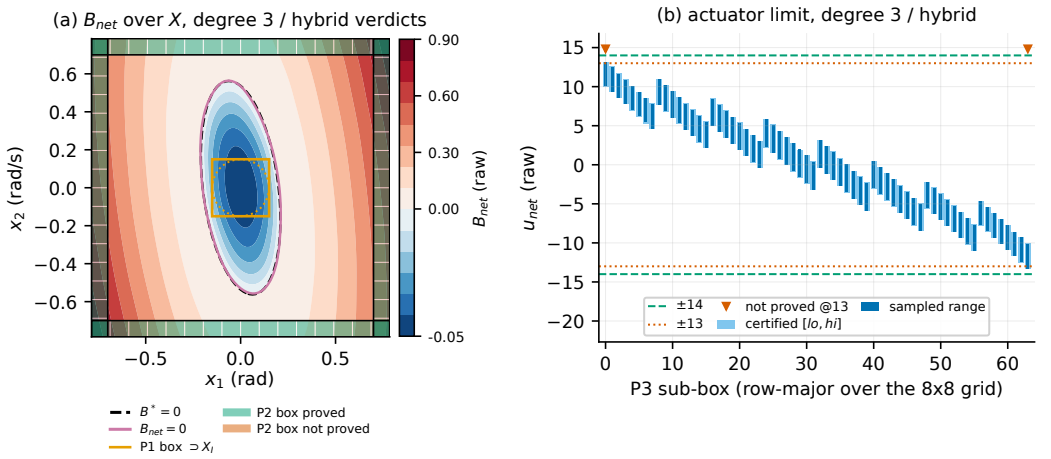


Figure 8: **Verification of the KAN4CBC pendulum networks as trained KANs** (degree-3 regime). (a) The learned barrier B_{net} is shown over the operating region, with its zero level set against that of the design barrier ($B^* = 0$), the initial-set box (P1), and the 64-box cover of the unsafe region tinted by verdict. All of these boxes are proved. (b) The certified output intervals of the controller network on the 64 P3 boxes are shown against the actuator limits. The bound $|u_{\text{net}}| \leq 14$ is proved everywhere, and the deliberately tight 13.0 variant correctly fails exactly where sampling attains 13.2 (marked). This shows that the certificate fails where the bound is violated, and therefore that it is not vacuous. The Lie-derivative condition on the same degree-3 barrier is discharged in Appendix C.5 and shown in Figure 2(d).

queries, so on the degree-1 networks the gap between clean and certified accuracy consists entirely of re-executed counterexamples (counts in Table 4).

C.5 Lie-derivative discharge on the pendulum

We instantiate Proposition 2 on the CS1 barriers with the controller in the loop. The operating region is cut into 16×16 boxes, the *band* of boxes whose certified B_{net} range contains 0 covers the zero level set, and a band box that fails is split 2×2 and retried. For the degree-3 barrier in hybrid mode, depth 2 discharges 27 of the 34 band boxes (worst certified $\sup \mathcal{L}_f B = +0.0097$ against a sampled maximum of -0.0101) and depth 3 all 34 (292 leaves, 16.6 s, worst certified $\sup \mathcal{L}_f B \leq -8.7 \times 10^{-5}$); intersecting with a mean-value-form enclosure, whose excess width is second rather than first order in the box width, closes the cover already at depth 2. All 394 points located on $\{B_{\text{net}}{=}0\}$ by bisection lie in proved leaves, with $\mathcal{L}_f B$ below the certified bound of their leaf, and the level set lies strictly inside the operating region, so with the sign conditions of CS1 the trained degree-3 barrier and controller satisfy all three barrier conditions. The degree-1 barrier in exact mode fails: only 14 of its 32 band boxes are discharged, and 8 leaves contain counterexamples on the zero level set (50 points) with $\mathcal{L}_f B$ up to $+0.312$. This regression-fit barrier thus passes both sign conditions yet violates the derivative condition, a defect that no refinement can remove and that a surrogate-based pipeline cannot see. Figure 12 shows the certified bound of every leaf of both refinements, the progress of the discharge with depth, and the leaves in which the degree-1 barrier violates the condition.

C.6 Scale: 784-input KANs

MNIST $[784, 10]$ **(E9).** This single-layer KAN is a spline additive model. Exact and hybrid modes were not attempted: the case count of Lemma 2 reaches 2^{30}, and the dense 784-dimensional predicate is costly by itself, since an uncapped exact probe on the degree-1 network, for a query that yields a single star, took 18.6 s against 0.15 s for star-set approx with the lifted star built sparsely after SparseImageStar (Tran et al., 2025). Star-set approx matches the PGD ceiling at 1, 2 and $4/255$

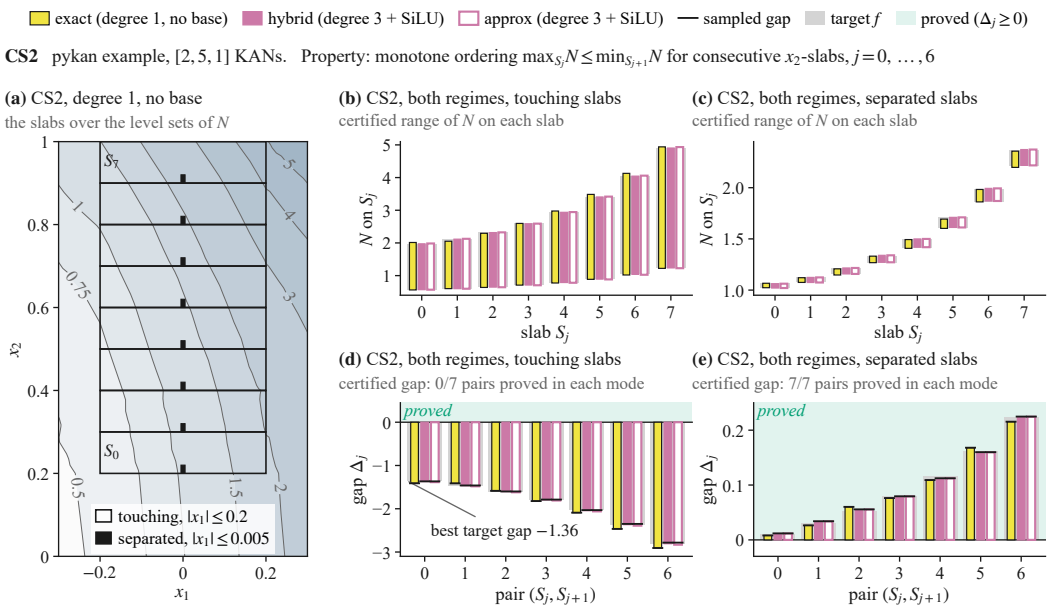


Figure 9: **Monotone ordering on CS2**, the pykan example with target $f = \exp(\sin(\pi x_1) + x_2^2)$, for the $[2, 5, 1]$ networks `cs2-toy-pwl` (degree 1, no base; exact mode, yellow) and `cs2-toy-cubic` (degree 3+SiLU; hybrid mode filled pink, approx mode hollow pink) of Table 3. The ordering $\max_{S_j} N \le \min_{S_{j+1}} N$ of consecutive x_2-slabs, $j = 0, \ldots, 6$, is proved when the certified gap Δ_j, the certified minimum of N over S_{j+1} minus its certified maximum over S_j, is nonnegative (green band). (a) Touching slabs $S_j = [-0.2, 0.2] \times [0.2 + 0.1j,\ 0.3 + 0.1j]$ (outlined) and separated slabs $S_j = [-0.005, 0.005] \times [0.2 + 0.1j,\ 0.22 + 0.1j]$ (filled), over labeled level sets of the degree-1 network, shaded darker where N is larger. (b, c) The certified range of N on each slab. (d, e) The certified gap of each pair. Black ticks mark the sampled gap, the same difference taken over sampled values of N, so the network's own gap lies between the end of a bar and its tick; hybrid and approx mode share one network and hence one tick. Grey bars give the range (b, c) and the gap (d, e) of f. On touching slabs every sampled gap is negative, so both networks violate the ordering, as does f (best gap -1.36), and all 21 checks correctly return not proved. On separated slabs every mode proves all 7 pairs.

and is one image short at $8/255$ (Table 5), with 5 UNKNOWN verdicts in the 880 queries of this experiment, at about 6.3 s per query end to end. We do not attempt to certify the radius $32/255$ at which Alter et al. (2025) attack. The degree-1 no-base network loses 0.74 accuracy points and most of its certified fraction (Table 5); because β-CROWN decides all of its queries (Appendix C.8), the loss is the network's rather than the verifier's, and since degree and base branch change together we attribute it to both.

Fashion-MNIST $[784, 10]$ **(E12a).** Star-set approx equals the PGD ceiling at all four radii without an UNKNOWN verdict (Table 5), at 6.7–6.9 s per query, and 101 of its 102 counterexamples are its own LP minimizers rather than attack points. At $32/255$ the attack breaks all 50 images. Hybrid mode is out of reach, since the case count of Lemma 2 has median 1.9×10^{11} at $4/255$ and reaches 2^{289}.

Depth-2 MNIST $[784, 16, 10]$ **(E12b).** This network first trained to only 3.7% accuracy because every hidden node left the spline support of layer 2, which is the zero-tail trap at training time. Placing the grid of layer 2 on the measured hidden range $[-157.1, 75.1]$, and re-gridding it without changing the function after each training stage, repairs the same architecture from the same layer-1 initialization to 93.7%. Star-set approx reaches the PGD ceiling at both radii and is strictly tighter than IBP and IBP+elide on all 50 queries (Tables 4 and 5), at 17.6 and 18.2 s per query on average at the two radii.

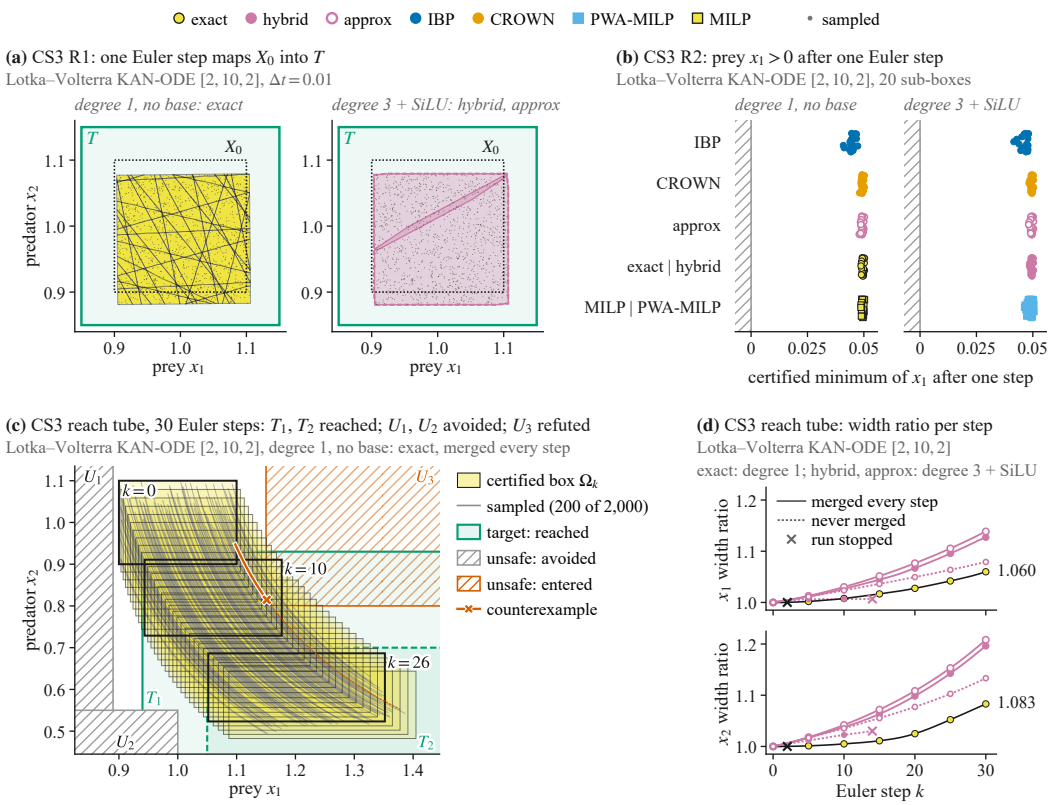


Figure 10: **One-step reachability and reach tubes on CS3**, the $[2, 10, 2]$ Lotka–Volterra KAN-ODE N with prey x_1, predator x_2 and Euler step $S(\boldsymbol{x}) = \boldsymbol{x} + \Delta t\, N(\boldsymbol{x})$, $\Delta t = 0.01$. Markers and colors are those of Figure 2, and CROWN shows β-CROWN. (a) R1: the certified enclosure of the image $S(X_0)$ of the initial box $X_0 = [0.9, 1.1]^2$ (dotted) lies in the target box $T = [0.85, 1.15]^2$ (green), and black dots are the images of 1,200 sampled initial states. On the degree-1 network the enclosure is the exact image, a union of 189 stars separated by thin lines. On the degree-3 network hybrid mode returns two stars (filled), which are relaxed separately and overlap in the darker band, and approx mode returns one star (dashed). At this scale the three enclosures are indistinguishable from the sampled image, with width ratios 1.0000, 1.002 and 1.003 for exact, hybrid and approx. (b) R2: dots show the certified minimum of x_1 after one step on each of the 20 sub-boxes of $[0.05, 0.15] \times [0.1, 4]$, with vertical jitter. Every member proves $x_1 > 0$ (hatched: $x_1 \leq 0$) on every sub-box in both regimes. (c) The 30-step reach tube of exact mode on the degree-1 network merges the stars into their box hull Ω_k (yellow) after every step. Grey lines are 200 of the 2,000 sampled network-Euler trajectories, all inside the tube. Thick outlines mark $\Omega_0 = X_0$ and the first boxes inside the targets, $\Omega_{10} \subseteq T_1$ (solid green) and $\Omega_{26} \subseteq T_2$ (dashed green). The tube avoids the unsafe boxes U_1 and U_2 (grey hatching) but meets U_3 (vermilion hatching) at steps 7 to 17, and a sampled trajectory, re-executed through the network, enters U_3 at step 8 (cross; dotted after entry), which refutes the avoid claim for U_3. All target and unsafe sets are boxes in the operating region $[0.05, 4]^2$, clipped to the view. (d) The width ratio (certified over sampled width) of each tube per step and coordinate, for exact mode on the degree-1 network and for hybrid and approx modes on the degree-3 network, with markers every five steps. Solid lines are runs that merge after every step and dotted lines are runs that never merge. Crosses mark where the two unmerged runs stopped: exact mode at step 2 with more than 400 stars, and hybrid mode after step 14, when step 15 aborted. The numbers are the ratios of the merged exact tube at step 30.

UCI tabular KANs (E12c). On MAGIC, Dry Bean and WDBC we compute a certified radius for 80 test points per dataset on a ladder that starts at 0.0025 standard deviations and grows by $\sqrt{2}$ per rung; the rungs closest to 0.01 and 0.02 form the UCI row of Table 2. Exact mode on the degree-

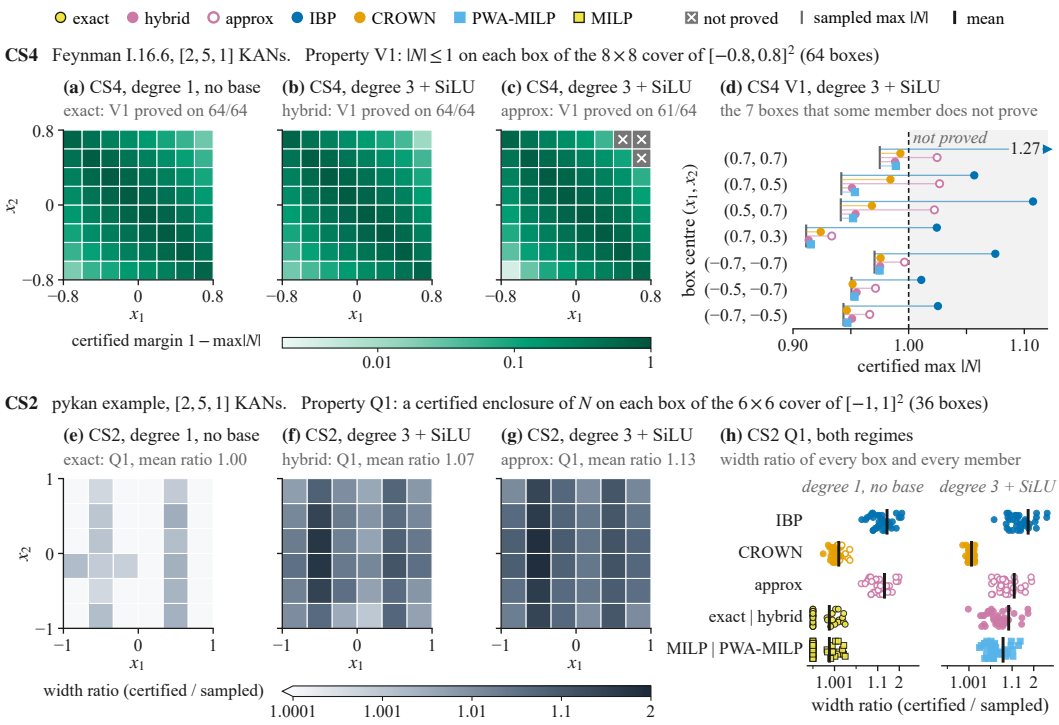


Figure 11: **The two output-range benchmarks, box by box.** Labels are as in Table 2, and CROWN shows β-CROWN. The top row is CS4, Feynman I.16.6, with property V1 ($|N| \leq 1$, velocities in units of c), and the bottom row is CS2, the pykan example, with property Q1. (a–c) The color of each box is its certified margin $1 - \max|N|$ on a logarithmic scale, and grey boxes with a white cross are not proved. (d) On the seven boxes of the degree-3 network that IBP does not prove, three of which approx does not prove either, a line joins the sampled maximum of $|N|$ (grey tick) to the certified maximum of each member. A certified maximum in the shaded region exceeds 1, and the arrow marks an off-scale IBP bound. On the degree-1 network every member proves all 64 boxes. (e–g) The color of each box is its width ratio, the certified over the sampled width, on a logarithmic scale of the excess over 1. Since exact mode computes the exact range (Theorem 1), the excess in (e) is the shortfall of sampling. (h) Markers give the width ratio of every box for every member, and black bars give the mean over the 36 boxes, which is the entry of Table 2. Ratios below 1.0001 take the lightest color in (e–g) and are drawn at 1.0001 in (h), and hollow CROWN markers are the 9 degree-1 boxes at which β-CROWN stopped at its time limit, with sound enclosures.

1 networks returns no UNKNOWN in 236 rung verdicts, and on each of the 39 points with both a certified and a falsified rung the two are adjacent, whereas hybrid mode on the degree-3 networks leaves 6 UNKNOWN in 245 and a gap of more than one rung on 5 of 10 such points. The certified-accuracy curve equals the attack ceiling on all 25 (dataset, radius) cells for exact mode and on 21 of 24 for hybrid mode. One data check of this experiment fails (Appendix C.12).

C.7 FAMILY COMPARISON DETAILS

Star-set exact against the exact MILP (E6). On 560 degree-1 queries of the annulus, two-moons and Gaussian networks the two agree on every verdict, and their margins differ by at most 7.3×10^{-5}, within the MILP solver's tolerance but above the 10^{-5} of our agreement check (Appendix C.12); an LP re-solve with the binaries fixed agrees with star-set exact to 2.7×10^{-14}. The median solver time per query of the MILP is $16\times$ smaller (0.0144 against 0.231 s), whereas star-set exact also returns the reachable set.

The degree-3+SiLU regime (E10). IBP, plain CROWN, the star-set modes and PWA-MILP (Appendix C.1) at $K \in \{2, 4, 8\}$ run on the same 120 robustness queries of the degree-3 annulus network

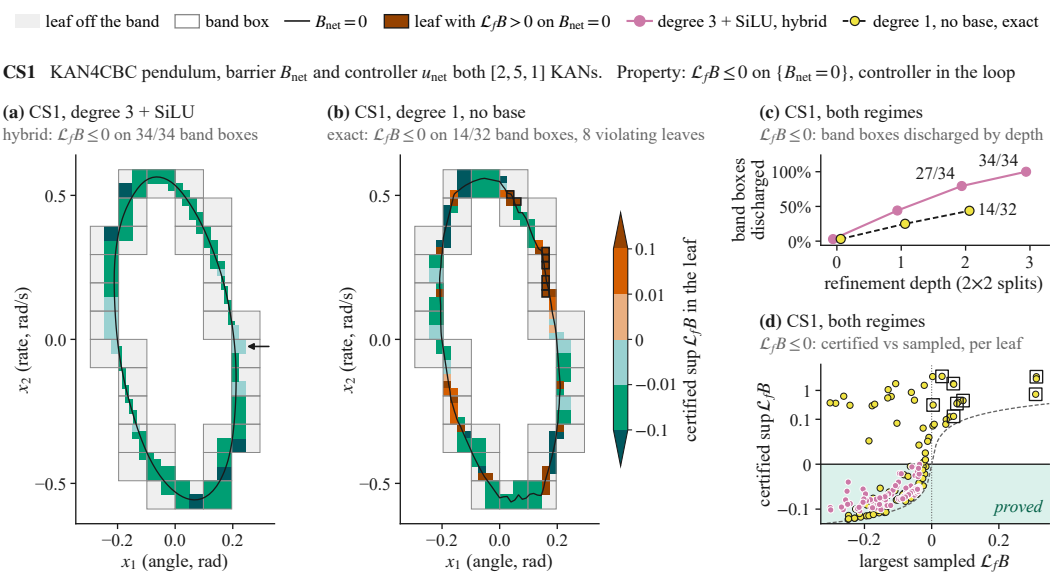


Figure 12: **The Lie-derivative condition on the CS1 pendulum barriers, leaf by leaf.** In each regime the barrier B_{net} and the controller u_{net} are $[2, 5, 1]$ KANs, and the interval test of Proposition 2 checks $\mathcal{L}_f B \le 0$ on $\{B_{\text{net}} = 0\}$ with the controller in the loop (Appendix C.5). (a, b) Outlined squares are the band boxes, the cells of the 16×16 grid on $[-\pi/4, \pi/4]^2$ whose certified B_{net} range contains 0, and the black curve is $B_{\text{net}} = 0$. A band box that is not proved is split $2{\times}2$ and retried; each final leaf is colored by its certified sup $\mathcal{L}_f B$ (green: proved; vermilion: not proved), or grey if its certified B_{net} range excludes 0 (off the band). (a) Hybrid mode discharges all 34 band boxes of the degree-3+SiLU barrier by depth 3; the arrow marks the largest certified bound, -8.7×10^{-5}. (b) For the degree-1 no-base barrier in exact mode, 39 leaves remain unproved at depth 2, where the run stops. Each of the 8 leaves outlined in black contains a point of $\{B_{\text{net}} = 0\}$ with $\mathcal{L}_f B > 0$ (up to $+0.312$), so the condition is violated there and no refinement can discharge these leaves. (c) The share of band boxes discharged (every leaf proved or off the band) at each refinement depth; the two series are offset slightly for legibility. (d) The certified bound of each colored leaf of (a) and (b) against the largest $\mathcal{L}_f B$ at 200 points sampled in the leaf, on a vertical axis that is linear within ± 0.01 and logarithmic beyond. Every leaf lies above the dashed curve, on which the two are equal; the shaded region is proved, and squares mark the 8 outlined leaves of (b). Regime colors and markers follow Figure 2.

(40 points at $\varepsilon \in \{0.05, 0.10, 0.20\}$), 129 pendulum boxes and 36 pykan boxes. These rows give the PWA row of Table 7 and, with the β-CROWN re-run of Appendix C.8, Figure 2(c), in which star-set approx and IBP stay under 25 ms per query. When every member gets star-set hybrid's time on each query, star-set hybrid is tightest on 82 of the 120 robustness queries and PWA-MILP on 84 of the 129 pendulum and 25 of the 36 pykan boxes; with unlimited time PWA-MILP at $K{=}8$ is tightest on 71 queries and on 129 of 129 pendulum and 33 of 36 pykan boxes, at $1.8\times$ star-set hybrid's time on robustness, and on the pykan boxes it is also faster. IBP+elide is tightest on all 200 queries of the depth-1 MNIST network, whose margin separates into 784 univariate problems, but at depth 2 star-set hybrid's certified margin exceeds that of IBP+elide on all 120 annulus queries, by a median factor of 1.3 and up to 23.6 where both are positive. At 784 inputs one class pair of the $K{=}2$ PWA-MILP took 18.7 s to reach a dual bound of 8.9235, against 8.9227 for star sets in 1.07 s; its cost grows steeply in K, but we do not claim that it fails.

The CROWN-style member (E13). On the 120 annulus queries of E10 the CROWN-style member certifies 85 (mean margin $+4.90$) against 82 ($+2.49$) for IBP, and at depth 4 and $\varepsilon{=}0.10$ it certifies 17 of 20, as many as star-set approx and hybrid, against 15 for IBP; star-set approx has the larger mean margin there only because of its tighten rows (Table 7). Of its three sound variants, the one with interval intermediate bounds never beats the better of plain and elided CROWN be-

Table 6: **Ablations**: one choice switched on the same networks and queries (A → B); #: queries or boxes per degree. Times: medians per query or box; gap: mean attacked minus certified margin; ×: ratio B/A. Further ablations are in Table 7.

Switch	Networks (#)	Effect on the guarantee	Cost
IBP → CROWN (plain)	annulus (660)	proved 250→322 (deg. 1), 407→423 (deg. 3)	5.3→5.3, 9.9→9.6 ms
tighten rows off → on	MNIST (200)	proved 128→129; gap 0.19→0.06	2.99→3.33 s
approx → exact / hybrid	annulus (660)	Unknown 91→0 (deg. 1), 16→1 (deg. 3)	26→316, 29→818 ms
IBP → IBP+elide	MNIST (200; depth 2: 50)	proved 99→130 (PGD ceiling); depth 2: 38→44	1.1→2.1 s
deg. 3+SiLU → deg. 1	MNIST; 9 tasks	exact mode applies; certified at 2/255 90%→34% (both at the PGD ceiling)	RMSE ×2.83; acc. −0.8 to +0.2 pts
KAN → ReLU conversion	26 nets (898)	same exact verdicts; IBP proves 317→0 (deg. 1), 340→77 (deg. 3); any method 402→394 (deg. 3)	exact star ×0.08–2.18, MILP ×0.13–2.02

yond floating-point rounding (3,570 robustness queries), so Tables 4 (CR. best) and 8 report that better one. No CROWN-style bound exceeds the exact one in 2,640 degree-1 comparisons, and all 2,934 FALSIFIED verdicts of the experiment re-execute, 1,884 of them from the candidate points of CROWN-style bounds.

The matrix run (E15). Re-running every member of Table 4 from the released files on every query with at most 30 inputs reproduces the 17,911 earlier records with 0 verdict differences. One of the 76 compared cells, the $K{=}4$ PWA-MILP on the pykan boxes, differs numerically, by up to 2.98×10^{-4}: the earlier run built 27 of the 36 boxes with arithmetic that differs by one ulp, and on these shifted boxes the MILP, which stops at the solver's default relative gap, returns a different sound dual bound; on the earlier boxes the difference is 0.0. One of the 20,100 enclosures misses a sampled value at solver tolerance (Appendix C.12, item (iv)).

The split cap (E17). In the split-cap ablation of Table 7 all 560 margins are sound, and $S_{\max}{=}4096$ reproduces the margins of E1 exactly. Three margins are non-monotone in $S_{\max}$, because the whole-range and the split relaxations are different sound over-approximations: one MAGIC point is ROBUST only at $S_{\max}{=}1$ (margin 0.0043, and -0.0075 from $S_{\max}{=}16$), and a second changes the other way.

Which member when. For degree-1 KANs we recommend the exact MILP or β-CROWN (depth ≤ 2) for verdicts, β-CROWN also at 784 inputs, and star-set exact for the reachable set and counterexamples. For degree-3 KANs, IBP and plain CROWN screen quickly, β-CROWN proves the most and gives tight enclosures, and star-set hybrid gives sets up to 30 inputs; at 784 inputs CROWN with final-layer elision, α-CROWN or star-set approx suffice, IBP+elide suffices for depth-1 margins, and at depth 2 plain CROWN and star-set approx match α-CROWN and β-CROWN (47 of 50) at a fraction of their root-bound time.

C.8 NATIVE α-CROWN AND β-CROWN

α-CROWN on a KAN. E19 implements α-CROWN (Xu et al., 2021) and β-CROWN (Wang et al., 2021) on the KAN layer itself, and we describe only what the KAN layer changes; CROWN denotes the member without slope optimization. α-CROWN bounds every edge by a lower line $ax + \underline{o}(a)$ and an upper line $a'x + \overline{o}(a')$ with separate slopes and the tightest offsets of Lemma 3(i), so any slopes are sound. Starting from the chord slopes, that is, from the CROWN bound, 100 Adam steps per objective optimize the bound computed with sampled offsets; the final offsets are then certified by Lipschitz sampling (Lemma 6), and the larger of this bound and the certified chord-slope bound is kept, which was the chord-slope bound on 15 queries. The last layer is relaxed through combined edges: a property row $\boldsymbol{h}^\top N(\boldsymbol{x})$ equals $\sum_p g_p(z_p)$ over the inputs z_p of the last layer, with $g_p = \sum_q h_q \phi_{q,p}$ univariate, which for a margin is the difference of two edges, as in the final-layer elision of IBP+elide. The hidden bounds of a depth-2 network are the certified per-

Table 7: **Further ablations** (reserve rows of Table 6; approx, exact, hybrid: star-set modes).

Choice (A → B)	Networks, #	Effect on the guarantee	Effect on cost
reach-tube merge: never → every step	Lotka–Volterra, degree 3, approx, 30 steps	tube width $+5.5\%$ (x_1), $+6.6\%$ (x_2); never merging, exact mode stops at step 2 (over 400 stars) and hybrid aborts without a verdict (NaN) at step 15	predicate rows 1,202→42; 186.6→1.9 s
split cap $S_{\max} \in \{1, 16, 256, 4096\}$	annulus, degree 3, 120; MAGIC, 20	Unknown 4, 0, 0, 0; cap reached 120, 88, 0, 0; MAGIC counts unchanged, but 2 points change verdict	0.03, 0.79, 1.01, 1.00 s
chord → optimal slope	annulus, degree 3, and pendulum: 185 queries/boxes	0 verdict changes; 0.35% of the relaxation slack removed	$\times 1.37$ total time
PWA $K{=}2{\to}4{\to}8$	annulus, degree 3, 120	proved 88→89→90 (attack ceiling 90)	32→121→907 ms
tighten rows off → on (approx)	depth 4, $[2, 4, 4, 4, 2]$, 20 at $\varepsilon = 0.1$	mean margin 7.58→10.26 (CROWN-style 7.94); proved 17→17 (CROWN-style 17)	–
approx → hybrid	Feynman, degree 3, 64 boxes	proved 61→64	0.02→0.19 s

Table 8: **The CROWN methods on the benchmark.** Entries count the queries or boxes proved out of #; on the CS2 row they are the mean width ratio of the certified enclosure (1: exact). The ceiling is # minus the queries with a re-executed counterexample, the most any sound method can prove, and **bold** entries reach it. The exact and hybrid columns repeat Table 2. On robustness rows CROWN keeps the better of its plain and its elided final layer (a third variant, with interval intermediate bounds, never improves on them beyond floating-point rounding), and on the case studies it is plain CROWN. β-CROWN is the best of CROWN, α-CROWN and β-CROWN on every row, and Table 2 shows it in its CROWN column. The median seconds per query of α-CROWN and of β-CROWN, the latter over the queries that the shared attack does not falsify (such a query closes before the root bound at almost no cost), are 0.08–0.27 and 0.08–0.21 up to 30 inputs; 1.7–6.5 and 1.8–7.2 on MNIST and Fashion-MNIST; 136 and 125 on MNIST, depth 2; and 0.10–0.13 and 0.16–1.2 on the case studies; the 60 s limit of β-CROWN is checked only after the root bound. $\times$: not run at 784 inputs; –: not applicable; ‡: 9 of the 36 boxes stopped at the limit, with sound enclosures.

		degree 1, no base				degree 3 + SiLU				
Task	#	ceiling	exact	CROWN	α-CROWN	β-CROWN	ceiling	hybrid	CROWN	α-CROWN	β-CROWN
Local robustness, 2–30 inputs											
Annulus	660	440	**440**	323	378	**440**	451	450	423	444	**451**
Two-moons	500	495	**495**	439	489	**495**	491	**491**	487	490	**491**
Gaussian-4	120	116	**116**	115	**116**	**116**	115	**115**	**115**	**115**	**115**
UCI×3	480	422	**422**	393	410	**422**	462	456	455	458	**462**
Local robustness, 784 inputs											
MNIST	200	57	×	**57**	**57**	**57**	130	×	129	**130**	**130**
Fashion-MNIST	200		*(degree-3 network only)*				98	×	**98**	**98**	**98**
MNIST, depth 2	50		*(degree-3 network only)*				47	×	**47**	**47**	**47**
Case studies											
CS1 barrier	129	129	**129**	127	128	**129**	129	**129**	**129**	**129**	**129**
CS2 enclosure	36	–	*1.00*	*1.47*	*1.11*	*1.00*‡	–	*1.07*	*1.31*	*1.08*	*1.00*
CS3 Euler step	21	21	**21**	**21**	**21**	**21**	21	**21**	**21**	**21**	**21**
CS4 range	64	64	**64**	63	**64**	**64**	64	**64**	60	**64**	**64**

edge interval box, which a backward pass cannot tighten because the first layer acts on each input separately; unlike the full α-CROWN we do not optimize them, and depth above 2 is not supported.

β-CROWN on a KAN. β-CROWN is best-first branch and bound with α-CROWN bounds (20 steps per subdomain). It picks an input or hidden node by the relaxation gap of the edges that read it and splits its range at the knot nearest to the midpoint, or at the midpoint if the range contains

no knot. An input split shrinks the box and recomputes every relaxation; a hidden split $z_h \leq t$ or $z_h \geq t$ enters the backward pass with a Lagrangian multiplier $\beta \geq 0$, optimized with the slopes and sound for every such β, and the last layer is then relaxed on the hidden box intersected with the split. On a degree-1 no-base network every edge is affine once its input range lies in one knot cell, after which the widest input is bisected, so the procedure is complete given enough time, up to the sampling margin of the certified offsets. Its counterexample candidates, the minimizing corners of open objectives in bounded subdomains and a short PGD at the root, count only once re-executed on the network. We omit the re-optimization of hidden bounds under splits, batched GPU bounding, the branching heuristics of the released tool and LP or MILP solves at the leaves, so the member is branch and bound with α-CROWN bounds and β multipliers in the style of β-CROWN, not the released tool.

Protocol and checks. Both members ran on the full query set of Table 2, $4{,}170$ queries and 500 boxes loaded from the released files. β-CROWN has a limit of $60\,\mathrm{s}$ per query and first tries the counterexample of the matrix run's attack, whose time it excludes, so the 845 queries that the attack falsifies close before the root bound. No certified bound exceeds the attack margin or the sampled value ($8{,}494$ bounds), all $1{,}691$ FALSIFIED verdicts re-execute on the network, and on the degree-1 rows β-CROWN agrees with star-set exact and the exact MILP on all $3{,}948$ verdict comparisons. A heavier attack finds no violation on the 242 queries and 7 boxes up to 30 inputs that α-CROWN or β-CROWN proves and CROWN does not, and the smallest positive certified margin of a robustness proof is 2.4×10^{-4}, far above the floating-point floor. The $55\,\mathrm{ms}$ per query of β-CROWN in Figure 2(c) is the median over the 120 queries of that panel of a single-process re-run of both members, which times the member call only and reproduces every verdict and bound of E19.

Results. β-CROWN reaches the ceiling on every row of Table 8. It decides every degree-1 query, including the 200 MNIST queries at 784 inputs, and on the degree-3 annulus and UCI rows no member without branch and bound reaches the ceiling. It branched on only 93 of the $4{,}170$ queries and 73 of the 500 boxes, so most of its proofs are root bounds, and its slowest query up to 30 inputs took $9.4\,\mathrm{s}$. α-CROWN closes about half of the gap between CROWN and the ceiling on the degree-1 annulus and most of it on two-moons, but leaves 93 queries and one box undecided. Its median cost per query is 0.08–$0.27\,\mathrm{s}$ up to 30 inputs and $1.7\,\mathrm{s}$ on the degree-1 MNIST network, against 0.006–$0.09\,\mathrm{s}$ and $0.08\,\mathrm{s}$ for the best CROWN variant.

Limits. On depth-2 MNIST the root bound alone takes a median of $136\,\mathrm{s}$ for α-CROWN and $125\,\mathrm{s}$ for β-CROWN, against $9.6\,\mathrm{s}$ for CROWN. Because the $60\,\mathrm{s}$ limit is checked only after the root bound, 41 of the 50 β-CROWN queries there exceed it, yet all 50 are decided: 47 are proved at the root without branching and 3 are falsified by the attack. Nine of the 36 degree-1 CS2 boxes stopped at the limit with sound enclosures of width ratio 1.0014–1.0052 (‡ in Table 8).

C.9 VERIFYING THE EQUIVALENT RELU NETWORK

Conversion (E18). A degree-1 no-base edge is a sum of hinges $\sum_g a_g \max\{x - t_g, 0\}$ over its $G{+}3$ knots (Appendix B.12), so a KAN layer converts exactly into one hidden ReLU layer whose hinges are shared by the m edges leaving each node, with $n(G{+}3)$ ReLUs instead of $nm(G{+}3)$. On the 12 degree-1 networks this gives 56 to $6{,}272$ hinges and $1.6\times$ to $15.2\times$ the nonzero parameters, and the converted networks match the KANs to 1.7×10^{-12} on 10^4 points on and beyond the query boxes. A degree-3+SiLU edge has no exact ReLU form, so we convert it soundly into its piecewise-linear interpolant with $s \in \{1, 2, 4, 8\}$ pieces per knot cell plus a per-edge error, an extra input ranging over a certified error box, which contains the true error at every sampled point. The converted networks are large: at $s{=}8$ the WDBC network has $950\times$ the KAN's parameters, and the 784-input networks have up to $13{,}446$ hinges at $s \in \{1, 4\}$.

Protocol. In this subsection α-CROWN and α, β-CROWN denote the released tools run on the converted network, not the native members of Appendix C.8: α-CROWN (Xu et al., 2021) through auto_LiRPA (Xu et al., 2020) and α, β-CROWN (Wang et al., 2021) with branch and bound, both in float64. The 898 paired queries and boxes cover all 26 instances, 40 per instance up to 30 inputs (or all, if fewer) and 16 per 784-input instance, evenly spaced over the released queries. The native side runs the members of Table 4, and its α-CROWN and β-CROWN counts are the verdicts of Appendix C.8 on the same queries; the converted network is verified by IBP, CROWN, the ***ReLU exact star*** (exact per-neuron star-set reachability (Tran et al., 2019)), the ***ReLU triangle star*** (star-

set reachability under the triangle relaxation) and a big-M MILP, all our implementations, and by the two released tools. IBP omits the final-layer elision on both sides, and native CROWN is plain CROWN. Bound methods run at every s, the complete methods and the 784-input networks at $s \in \{1, 4\}$, with a limit of 60 s per query and method. Both sides of a query run in the same process, one after the other, in an order that alternates from query to query. Every converted counterexample is re-executed on the KAN (929 in all), every converted bound is checked against the attack and 20,000 samples, and the complete methods agree on all 397 decided degree-1 queries.

Degree 1. Up to 30 inputs (381 queries on 10 networks), star-set exact, the native MILP, the ReLU exact star, the converted MILP and α, β-CROWN decide every query alike (360 proved, 21 falsified). On all 419 degree-1 queries and boxes that both star-set exact and the ReLU exact star decide, at any input width, the two return the same number of stars. The conversion is mostly faster: counting the pykan boxes, the ReLU exact star is faster than star-set exact on 8 of 11 instances, all but the three G=16 two-input networks with the most stars (pooled medians 0.17 and 0.60 s per query), and the converted MILP is faster than the native MILP on 9 of 11 (0.020 and 0.046 s). The incomplete methods separate the two views: native IBP proves 317 queries and converted IBP none (pykan mean width ratio 1.27 against 1,627), and native plain CROWN proves 332 against 320 for converted CROWN, which costs 7 to 39× less time per instance. On the conversion, α-CROWN (345, at 13 to 47× the time of the better of plain and elided native CROWN) and the ReLU triangle star (349) overtake native CROWN and star-set approx (340), and native slope optimization restores the lead (351 for the native α-CROWN, all 360 for the native β-CROWN). At 784 inputs (MNIST, 16 queries), IBP+elide attains the MILP optimum of the margin to a relative 1.2×10^{-4}, the order of the MIP gap, in a median of 0.083 s, against 0.88 s for the converted MILP and 4.35 s for α, β-CROWN, which return the same 4 proofs and 12 falsifications. Star-set exact decides only 2 of these queries within 60 s and the ReLU exact star 11; since the native exact MILP was not run at this width, only the converted MILP and α, β-CROWN decide every query of this comparison; outside it, the native β-CROWN of Appendix C.8 decides all 200 queries of this network.

Degree 3+SiLU. Up to 30 inputs (381 queries on 10 networks) every native member proves more than the best refinement of its converted counterpart: IBP 340 (converted: 77), plain CROWN 358 (353), star-set approx 361 (ReLU triangle star: 357), star-set hybrid 364 (ReLU exact star: 307), the native β-CROWN 365 and α-CROWN 364 (α, β-CROWN: 356) and PWA-MILP 362 (big-M MILP: 356), although α, β-CROWN is faster than star-set hybrid on 7 of 10 networks at s=1 and on 4 at s=4. On the three 784-input networks (48 queries), native plain CROWN and star-set approx prove 11, 9 and 16 of 16 on MNIST, Fashion-MNIST and depth-2 MNIST (native α-CROWN and β-CROWN: 12, 9 and 16) against 10, 6 and 11 for the best converted method at s=4, but take 7 to 21 s per query against 0.11 to 0.28 s for converted CROWN. Over the 13 degree-3 networks with proof properties (429 queries), the native members together prove 402 and the converted methods together, at any refinement including the finer conversions below, 394; no query is proved only on the conversion, and the native side leads on 7 networks by one or two queries each.

Per-edge tightness. On all 56,460 degree-1 edges and query intervals the native exact range lies inside the hinge-wise interval sum of Proposition 3(i), which, where the native range is nonzero, is wider by a median factor of 1.3 to 66 per network and layer and by up to 4.5×10^5. On the 6,082 cases with an unstable hinge, the CROWN relaxation of the conversion (the chord above and one line below each unstable hinge) has a smaller area than the native parallel strip on 5,333, but a smaller maximum vertical width than the native gap on only 99, all with unstable hinges of opposite sign. For degree-3+SiLU edges the certified conversion error exceeds the native relaxation gap in the median on every network and layer at s=1, by 1.1 to 1,300×, and still at s=4 on the 784-input networks (29 to 65×) and at s=8 on the first tabular layer (13 to 20×); at s=8 it falls below the gap on the two-input networks, except the first KAN-ODE layer (1.3×).

Depth. The looseness of converted interval bounds compounds with depth: on untrained degree-1 networks of hidden width 4 and depth 1 to 6 (24 queries each) the median ratio of converted to native interval width grows from 6.1 to 143,000, and on trained ladders of depth 1 to 4 from 7.8 to 3,180 (degree 1) and from 3.4 to 262 (degree 3, s=8). For CROWN the ratio on the untrained networks stays between 0.95 and 1.00 up to depth 3 and reaches 4.66 at depth 6, and α-CROWN and the ReLU triangle star are at least as tight as native CROWN and star-set approx up to depth 5 but not at depth 6. On the trained ladders, however, the CROWN ratio is already 1.13 at depth 2 for degree 1, reaching 2.10 at depth 4, and lies between 1.04 and 1.50 at every depth for degree 3 at s=8. Every benchmark network has depth at most 2.

Finer conversion at 784 inputs. Re-converting the one-layer degree-3 networks at s=8 (17,610 and 26,707 hinges) lets converted CROWN prove 11 of 16 queries on MNIST, as many as native plain CROWN (native α-CROWN, β-CROWN and IBP+elide: 12), and 7 on Fashion-MNIST against 9 natively (α-CROWN on the conversion: 8); refining layer 2 of the depth-2 network to s=16 brings converted CROWN, α-CROWN and α, β-CROWN to 15 of 16, against 16 natively. Eight α, β-CROWN runs on Fashion-MNIST at s=8 exceeded the time limit and count as timeouts (Appendix C.12).

C.10 The accuracy price of the verifiable regime

Over nine tasks (the eight case-study tasks and the $[2, 2, 1]$ Feynman shape), with 135 fits across the degree-3 regime, the degree-1 regime at $G \in \{5, 8, 16, 24\}$ and three seeds, the degree-1 no-base restriction costs a median $2.83\times$ in held-out RMSE on the six regression tasks (range $0.87\times$ to $6.04\times$) and between -0.8 and $+0.2$ accuracy points (degree 1 minus degree 3) on the three classification tasks, at comparable parameter counts. Without the base branch, the bounded spline support gates the gradients and dead-grid seeds occur, which widens the worst-to-best spread over seeds from at most $1.5\times$ to $3.4\times$.

C.11 What was not certified

The degree-1 pendulum barrier violates the Lie-derivative condition, with re-executed counterexamples in 8 leaf boxes ($\mathcal{L}_f B$ up to $+0.312$) and only 14 of 32 band boxes discharged (Appendix C.5). In the degree-1 regime the relaxed members leave up to 189 verdicts per entry of Table 2 undecided. Star-set approx and CROWN without slope optimization are one image short of the MNIST PGD ceiling at $8/255$, and no member attempts the $32/255$ radius attacked by Alter et al. (2025) (Appendix C.6). Without star merging only the approx reach tube completes, at about $100\times$ the merged cost: the exact tube stops at step 2 with over 400 stars and the hybrid tube aborts at step 15 on a float64 NaN; the continuous-time certificate covers zero steps of the degree-3 hybrid configuration (Appendix C.4). PWA-MILP hit its node limit on 405 of its 2,010 queries and boxes, all on the UCI tasks, which keeps it below the ceiling on MAGIC and Dry Bean, and 5 capped solves returned no finite dual bound (Appendix C.2). The native β-CROWN exceeds its 60 s limit on 41 of the 50 depth-2 MNIST queries, because the limit is checked only after the root bound (median 125 s), although it decides all 50, and it stops 9 of the 36 degree-1 CS2 boxes at the limit, with sound enclosures. The native α-CROWN leaves 93 queries and one box undecided, does not optimize the hidden bounds and supports depth at most 2 (Appendix C.8). Star-set exact gives the native view no speed advantage: the ReLU exact star and the converted MILP are faster than star-set exact and the native MILP on 8 and 9 of 11 degree-1 instances up to 30 inputs, and at 784 inputs star-set exact decides only 2 of 16 queries within 60 s. On the conversion, 8 α, β-CROWN runs exceeded the time limit and count as timeouts (Appendix C.9). The end-to-end slope of the relaxation-gap sweep over $G \geq 4$ is 1.46, outside the registered window, because the fits at G=4 and 6 had not converged (Appendix C.3). Finally, PWA-MILP, which may be looser than the original (Appendix C.1), and the ReLU star and MILP baselines are our reimplementations, and the ReLU comparison uses query subsets (40 per instance, 16 at 784 inputs).

C.12 Disclosed failed checks

Every experiment registers its soundness and methodology checks before it runs, and the following checks fail. (i) Relaxation-gap sweep (E2): at the finest widths the gap reaches the Lipschitz-sampling floor, which fails the within-knot-cell margin and floor-ratio checks, and the end-to-end slope over $G \geq 4$ is 1.46, outside the registered window $[1.7, 2.3]$ (Appendix C.3). (ii) Degree-1 MILP comparison (E6): margin agreement fails on 12 of 560 queries (worst relative difference 5.2×10^{-5}) and counterexample reproduction on 69 (worst 9.8×10^{-5}), both at MIP tolerance. (iii) Tabular experiment (E12c): the standardized Dry Bean training features miss unit standard deviation by 1.7×10^{-9}, against a tolerance of 10^{-9}, because of the 10^{-12} epsilon added to a feature with standard deviation 0.000594. (iv) Matrix run (E15): the enclosure check fails on one Lotka–Volterra positivity box, where the exact MILP's upper end lies 1.0×10^{-7} below a sampled value, at solver tolerance and with the verdict unchanged (the ReLU comparison flags the same box); the reproduction of the earlier experiments' records fails numerically on one cell, the pykan

PWA-MILP, by up to 2.98×10^{-4} without a verdict difference (Appendix C.7); and the node-limit check fails because 5 capped PWA-MILP solves returned no finite dual bound, so four of their rows report UNKNOWN and the fifth is falsified by the attack. (v) First degree-1 ReLU comparison (E16; 280 queries and boxes on seven networks, with the same exact verdicts): the MILP-margin and enclosure checks fail on one controller box, where the converted big-M MILP reported as optimal an unsound maximum of -8.013 although samples and star-set exact reach -7.667, because its big-M constants came from interval bounds of the converted network that reach ± 2793 there; with presolve off the re-solve returns -7.667, and in E18 every converted enclosure contains its samples. (vi) ReLU comparison (E18): the time-limit check fails on 8 α, β-CROWN runs on Fashion-MNIST in the finer conversion, which took 72.6 to 80.0 s against the 65 s allowed because the model build and the α-CROWN initialization lie outside its branch-and-bound timeout; they count as timeouts and change no claim (Appendix C.9).