TRAINING WITH CERTIFICATES: GLOBAL OPTIMALITY BOUNDS FOR NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

How much better could a neural network have been trained on the same data? A flat loss curve, a small gradient, or agreement across restarts does not provide a global bound on the improvement still possible. We develop *Neural MERIT*, a framework that returns a feasible neural network together with a certified interval $L_k \le p^\star(Y) \le U_k$ enclosing the globally optimal empirical objective, so that $U_k - L_k$ bounds the remaining global suboptimality of the returned solution. The central idea is to approach the target training problem Y through a sequence of globally solved or certified surrogate problems C_k, and to transport each surrogate's feasible solution and global lower information back to the fixed target. For squared-loss training, we show that exact surrogate minimizers along a straight continuation yield monotonically improving target objectives even for nonconvex neural classes whose global optimizer can change discontinuously. With globally valid directional evidence, the transported certificate obeys an exact quadratic identity; in particular, a zero-gap surrogate certificate at C gives target certificate width $\frac{1}{2}\|Y - C\|^2$. Inexact surrogate solves incur an explicit, vanishing allowance. For exact convex prediction formulations, both certificate endpoints tighten monotonically and meet at the target optimum. Globally solvable neural classes therefore provide bank-free Neural MERIT guarantees, while a *Neural Basis Bank* of previously certified surrogates can strengthen the same target interval through additional and cooperative evidence. Across 24 independently verified ReLU continuation paths, every final bank-free interval contracts below 4.91×10^{-7} (median 3.82×10^{-7}); a 64-record basis bank certifies 18 of 24 targets to width 10^{-2} before continuation begins. Exact finite-network and frozen-feature experiments further show how the certificate separates optimizer stasis from globally certified training quality.

1 INTRODUCTION

When a neural-training run stops improving, has it reached the best objective permitted by its architecture, or has the optimizer stopped short? A flat loss curve, a small gradient, or agreement among restarts cannot in general answer this question. They describe the training procedure's behavior, whereas global optimality concerns every feasible network. The practical uncertainty is how much better the same architecture could have been trained on the same data.

Resolving that uncertainty would support several decisions: whether to continue optimization, try a different optimizer, or change the model class. A feasible network establishes an achievable objective value. A global lower bound limits how far any competitor can improve on it. Together they give the desired output of training:

$$L_k \le p^\star(Y) \le U_k = F_Y(\widehat{\theta}_k), \qquad 0 \le F_Y(\widehat{\theta}_k) - p^\star(Y) \le U_k - L_k. \tag{1}$$

Here Y denotes the target values of the fixed training problem, $p^\star(Y)$ is its global empirical optimum, and $\widehat{\theta}_k$ is a feasible network. A small width certifies near-global training quality. If the lower bound exceeds a required objective level, further optimization within the declared class cannot meet that requirement.

Obtaining an informative lower bound is the difficult part. Neural training has substantial global complexity barriers (Blum & Rivest, 1992; Boob et al., 2022). A related problem whose global solution is known offers another source of evidence. Evaluating its solution on the original target and

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.

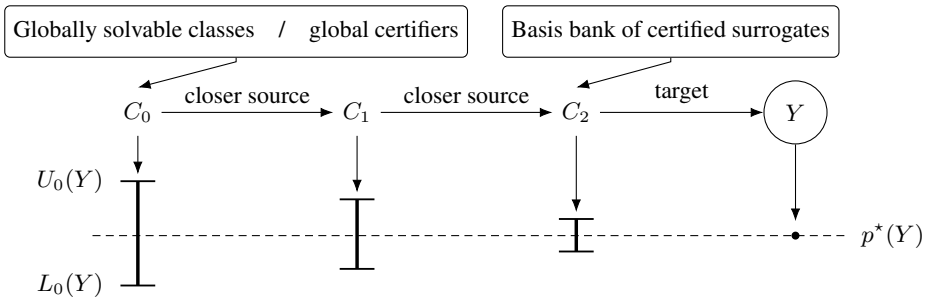

Figure 1: **A solved nearby problem can certify the quality of a target solution.** Each surrogate C_k supplies a feasible network evaluated on the fixed target Y and a transported lower bound. Global solvers generate these guarantees without a bank; stored surrogates strengthen them. The lower band depicts retained intervals, which remain nested with inexact source solves.

bounding the effect of the problem change yields a target guarantee. Repeating with closer certified surrogates produces approximate solutions with increasingly accurate global-suboptimality bounds.

We study squared-loss families: each surrogate replaces the target values Y by C, keeping inputs, architecture, constraints, regularization, and loss metric fixed. We call the framework *Neural MERIT*, extending the Monotonically ERror-bound Improving Technique for quadratic optimization (Soltanalian & Stoica, 2014a;b). Each source certificate bounds its returned network's quality on the original target.

Global surrogate information has two complementary sources. Solvable neural classes and sound relaxations provide bank-free certificates; a *Neural Basis Bank* adds previously solved problems, including earlier continuation sources (Soltanalian & Hassibi, 2015). Compatible records can cooperate (Figure 1). Proximity controls certificate error, not source-solve cost. Analytic solutions, reusable proof structure, and stored records make the route useful.

Contribution. We connect four steps. Scalar source certificates first yield nonconvex target bounds and monotone exact-surrogate continuation (Section 2). Global support evidence then sharpens the distance law (Section 3): a source pair with zero primal–lower gap, $g_C = 0$, gives certificate gap $\frac{1}{2}\|Y - C\|^2$; we call such a pair *closed*. Exact convex formulations yield a continuous quadratic path, with explicit allowances for inexact solves. Global neural engines supply these certificates (Section 4), and basis banks strengthen them (Section 5).

Relation to existing guarantees. Global-optimality conditions identify when particular neural stationary points or local minima are global (Haeffele & Vidal, 2017; Yun et al., 2018). Here we bound the suboptimality of a feasible target network using certified surrogate problems. Training Under Challenge constructs executable alternatives; global-gap conclusions additionally require challenge coverage (Yeganegi et al., 2026). Neural MERIT instead transports verified global lower information between training problems.

2 Certified approximation in problem space

To use a solved problem as evidence for another, we first need a precise measure of how they differ. We vary the training targets while keeping the neural class and all objective conventions fixed. Let Θ be a nonempty parameter domain, X the inputs, and $\mathcal{H}$ a finite-dimensional real Hilbert prediction space. Write $z(\theta) = f_\theta(X) \in \mathcal{H}$, let $C, Y \in \mathcal{H}$ be the source and target values, and consider

$$F_Y(\theta) = \tfrac{1}{2}\|z(\theta) - Y\|^2 + R(\theta), \qquad p^\star(Y) = \inf_{\theta\in\Theta} F_Y(\theta). \tag{2}$$

Assume $R \geq 0$ and a finite-cost feasible network; a known finite lower bound on R can be subtracted first. All distances use the fixed loss-normalized inner product. Transport uses infima; attainment is stated when needed. Outward evaluation gives $F_Y(\widehat{\theta}_k) \leq U_k$ in (1), preserving the suboptimality guarantee.

2.1 A SOLVED SURROGATE CERTIFIES THE TARGET

A source network immediately supplies a target upper bound by evaluation. The missing step is a lower bound that remains valid after the targets change. Let $0 \le \ell_C \le p^\star(C)$ be a verified source lower bound and θ_C a feasible network, with $u_C = F_C(\theta_C)$ and $\delta = \|Y - C\|$. To control the change in optimal value, combine prediction and regularization in the augmented reachable set

$$\mathcal{S} = \{(z(\theta), \sqrt{2R(\theta)}) : \theta \in \Theta,\ R(\theta) < \infty\}$$

which satisfies $\sqrt{2p^\star(Y)} = \operatorname{dist}((Y, 0), \mathcal{S})$. Distance to a nonempty set is 1-Lipschitz, whether or not that set is convex or closed. This gives a target certificate directly.

Theorem 2.1 (Metric Neural MERIT)**.** *For the family* (2)*,*

$$L_C(Y) = \tfrac{1}{2}[\sqrt{2\ell_C} - \delta]_+^2 \le p^\star(Y) \le U_C(Y) = F_Y(\theta_C), \tag{3}$$

and its root-width obeys

$$\sqrt{2U_C(Y)} - \sqrt{2L_C(Y)} \le 2\delta + \sqrt{2u_C} - \sqrt{2\ell_C}. \tag{4}$$

For an exact source, $u_C = \ell_C = p^\star(C)$, the sharper bound is $\delta + \min\{\delta, \sqrt{2p^\star(C)}\} \le 2\delta$. The factor two is sharp.

Proof. The distance inequality gives the lower bound. Feasibility gives the upper bound and $\sqrt{2U_C(Y)} \le \sqrt{2u_C} + \delta$. Subtract $[\sqrt{2\ell_C} - \delta]_+$ and distinguish whether $\sqrt{2\ell_C} \ge \delta$. The exact-source specialization follows by the same case split. Sharpness and the equivalent lifted formulation are given in Appendix A. □

The certificate closes as both problem distance and source uncertainty vanish. Exact scalar-source information gives a first-order objective-width envelope at positive residual: for $s = \sqrt{2p^\star(C)} \ge \delta$, the sufficient width is $2s\delta$.

2.2 CLOSER EXACT PROBLEMS GIVE BETTER TARGET SOLUTIONS

A shrinking distance bound is useful, but does each closer surrogate also provide a better target solution? This requires care because nonconvex optimizers can switch abruptly. Along the straight path $C(t) = C_0 + t(Y - C_0)$, $0 \le t \le 1$, the target objectives nevertheless improve for every choice of exact source minimizers.

Theorem 2.2 (Exact-surrogate continuation)**.** *At each considered $t < 1$, let $\theta_t \in \arg\min_\theta F_{C(t)}(\theta)$. Then $U(t) = F_Y(\theta_t)$ is nonincreasing for every choice of these global minimizers. The metric lower bound*

$$L(t) = \tfrac{1}{2}[\sqrt{2p^\star(C(t))} - (1 - t)\|Y - C_0\|]_+^2 \tag{5}$$

is nondecreasing. Along such a source sequence with $t \uparrow 1$, both endpoints converge to $p^\star(Y)$. No convexity, uniqueness, or continuity of the optimizer map is required.

Proof. Let $v = Y - C_0$, $A(\theta) = F_{C_0}(\theta)$, and $b(\theta) = \langle v, z(\theta) - C_0\rangle$. The source objective is $A(\theta) - tb(\theta) + \frac{1}{2}t^2\|v\|^2$. For $s < t$, the two source optimality inequalities imply

$$s\Delta b \le \Delta A \le t\Delta b, \qquad \Delta A = A(\theta_t) - A(\theta_s), \quad \Delta b = b(\theta_t) - b(\theta_s).$$

Hence $\Delta b \ge 0$ and $U(t) - U(s) = \Delta A - \Delta b \le 0$. The 1-Lipschitz root value and $\|C(t) - C(s)\| = (t - s)\|v\|$ give monotonicity of (5). Theorem 2.1 makes the root-width vanish at the endpoint. □

With inexact solves or nonuniform paths, (3) still gives valid endpoints. Retaining historical maxima of lower bounds and minima of feasible upper values makes the reported intervals monotone. Exact straight continuation gives this behavior intrinsically.

3 SHARPER GUARANTEES FROM GLOBAL LOWER EVIDENCE

The metric bound treats every target change of the same length alike. A tighter guarantee is possible when the source proof also records which prediction directions are costly. Represent that information by a vector $\xi \in \mathcal{H}$ and a finite global support bound a satisfying

$$a \geq \mathsf{H}(\xi) := \sup_{\theta \in \Theta} \{\langle \xi, z(\theta) \rangle - R(\theta)\}. \tag{6}$$

This inequality lower-bounds the regularization cost at every feasible network by $\langle \xi, z(\theta) \rangle - a$. Minimizing the resulting quadratic gives $D_Y(\xi, a) = \langle Y, \xi \rangle - \frac{1}{2}\|\xi\|^2 - a \leq p^\star(Y)$. Once a is globally verified, the same evidence is valid for every target in the family. Pairing it with a feasible network reveals exactly how the certificate changes with Y.

Theorem 3.1 (Exact and inexact quadratic transport)**.** *Let θ be feasible, $z = z(\theta)$, and $r \geq R(\theta)$ be finite. Together with* (6)*, define*

$$U_Y = \tfrac{1}{2}\|z - Y\|^2 + r, \quad \widetilde{C} = z + \xi, \quad \eta = r + a - \langle z, \xi \rangle \geq 0.$$

For every Y,

$$0 \leq U_Y - D_Y(\xi, a) = \eta + \tfrac{1}{2}\|Y - \widetilde{C}\|^2. \tag{7}$$

At a nominal source C, let $g_C = U_C - D_C(\xi, a)$ and $\rho = C - z - \xi$. Then

$$U_Y - D_Y = g_C + \tfrac{1}{2}\|Y - C\|^2 + \langle \rho, Y - C \rangle, \qquad \|\rho\| \leq \sqrt{2g_C}, \tag{8}$$

$$U_Y - D_Y \leq \tfrac{1}{2}\big(\|Y - C\| + \sqrt{2\bar{g}_C}\big)^2, \qquad \bar{g}_C \geq g_C. \tag{9}$$

If the source certificate closes, $g_C = 0$, the certificate gap equals $\frac{1}{2}\|Y - C\|^2$.

Proof. The support inequality at θ gives $\eta \geq 0$. Completing the square gives (7), and at C it gives $g_C = \eta + \frac{1}{2}\|\rho\|^2$. Expanding about C proves (8); Cauchy–Schwarz proves (9). □

The identity locates the effective source center $\widetilde{C} = z + \xi$, including for nonconvex classes. With inexact evidence, $\widetilde{C}$ may differ from the nominal source C, and (7) retains this directional information. The zero-gap quadratic law requires a closed source certificate. A global source minimizer alone need not close the chosen support dual: for $z \in \{-1, 1\}$, $R = 0$, and $Y = 0$, the optimum is $1/2$ while the best support lower bound is zero. Exact scalar-source values still yield convergent metric certificates. Appendix B gives the full comparison and sharp examples.

3.1 AN INTRINSICALLY TIGHTENING QUADRATIC PATH

The quadratic law is most useful when closed source pairs can actually be obtained. Exact convex prediction formulations provide such pairs while preserving the neural optimum and a feasible representation. Write

$$P_C(z) = \phi(z) + \tfrac{1}{2}\|z - C\|^2, \qquad D_C(\xi) = \langle C, \xi \rangle - \tfrac{1}{2}\|\xi\|^2 - \phi^*(\xi), \tag{10}$$

where ϕ is proper, closed, and convex, with $\phi(z(\theta)) \leq R(\theta)$ for every feasible θ. The quadratic is continuous everywhere, so strong duality holds; the unique optimal prediction is $z_C = \operatorname{prox}_\phi(C)$ and $\xi_C = C - z_C = \operatorname{prox}_{\phi^*}(C)$ (Moreau, 1965; Bauschke & Combettes, 2017).

Theorem 3.2 (Quadratic continuation)**.** *For every C on the segment $C(t) = (1-t)C_0 + tY$, assume* (10) *is exact for* (2) *and its minimizer z_C has a feasible representation θ_C with $z(\theta_C) = z_C$ and $R(\theta_C) = \phi(z_C)$. Then its exact primal–dual pairs satisfy*

$$\begin{gathered} L(t) = D_Y(\xi_{C(t)}) \uparrow p^\star(Y), \qquad U(t) = P_Y(z_{C(t)}) \downarrow p^\star(Y), \\ U(t) - L(t) = \tfrac{1}{2}(1-t)^2\|Y - C_0\|^2. \end{gathered} \tag{11}$$

Both endpoint paths are continuous. The statements concern exact surrogate pairs; finite computations report verified approximations to them.

To see the source of monotonicity, let $p(C) = \min_z P_C(z)$. Its Moreau envelope is convex, differentiable, and 1-smooth, with $\nabla p(C) = \xi_C$ (Parikh & Boyd, 2014). Thus $L(t) = p(C(t)) + \langle \nabla p(C(t)), Y - C(t) \rangle$. Convexity and monotonicity of the gradient make this tangent value rise along the segment. Smoothness bounds its rise by the decrease of $\frac{1}{2}\|Y - C(t)\|^2$, making $U(t)$ fall. Appendix C gives the complete finite-difference proof. Nonconvex exact paths may have optimizer-switch jumps; the continuous paths in (11) follow from the proximal structure.

How accurately should an intermediate source be solved? Intermediate solves need only be accurate enough for the target guarantee being sought. To balance their error with remaining problem distance, use $C_k = Y - q^k(Y - C_0)$, $0 < q < 1$, and $\delta_k = q^k\|Y - C_0\|$. Requiring $\bar{g}_k \le \kappa\delta_k^2$ gives

$$U_k(Y) - L_k(Y) \le \tfrac{1}{2}(1 + \sqrt{2\kappa})^2\delta_k^2. \tag{12}$$

A target tolerance ε is guaranteed once $\delta_k \le \sqrt{2\varepsilon}/(1 + \sqrt{2\kappa})$. This counts certified source stages; their cost depends on the solver. Numerical and reconstruction allowances enter $\bar{g}_k$ or the final endpoints.

4 Obtaining global surrogate information

Three exact neural classes and a sound relaxation construction supply global surrogate information without a bank. Each continuation stays within its declared training family.

Finite neural classes and deep linear networks. Finite parameter alphabets allow exhaustive source optimization, including all-layer nonconvex networks. Unregularized deep linear networks instead have an analytic optimum governed by input span and bottleneck rank. For $X \in \mathbb{R}^{n\times d}$, let $P_X = XX^\dagger$ and r be the effective bottleneck rank. The exact value is

$$p^\star(C) = \tfrac{1}{2}\|(I - P_X)C\|_F^2 + \tfrac{1}{2}\sum_{j>r}\sigma_j(P_XC)^2. \tag{13}$$

Projection and truncated singular-value decomposition reconstruct a global source network (Eckart & Young, 1936; Mirsky, 1960); Appendix D gives the factorization.

Regularized two-layer ReLU networks. Nonlinear sources with continuous weights can also be certified when their activation structure admits an exact convex formulation. For scalar targets, consider the width-N, bias-free problem

$$\min_{u_j,a_j}\frac{1}{2}\left\|\sum_{j=1}^N a_j(Xu_j)_+ - C\right\|_2^2 + \frac{\beta}{2}\sum_{j=1}^N(\|u_j\|_2^2 + a_j^2), \qquad \beta > 0. \tag{14}$$

Balancing each neuron converts its penalty to $\beta|a_j|\|u_j\|$. With $\mathcal{A}_X = \{\pm(Xu)_+ : \|u\|_2 \le 1\}$, $r_X = \dim\operatorname{span}\mathcal{A}_X$, and $N \ge r_X$, the exact prediction cost is $\phi = \beta\gamma_X$, where γ_X is the gauge of $\operatorname{conv}\mathcal{A}_X$ (Bach, 2017; Pilanci & Ergen, 2020; Ergen & Pilanci, 2021a). Its dual is

$$\max_\nu\ q_C(\nu) = \langle C, \nu\rangle - \tfrac{1}{2}\|\nu\|_2^2, \quad M_X(\nu) := \sup_{\|u\|_2\le 1}|\nu^\top(Xu)_+| \le \beta. \tag{15}$$

A master restricted to the available patterns supplies a feasible neural upper value. Global pricing checks omitted atoms: a certified $M \ge M_X(\nu)$ makes $\alpha\nu$ globally dual feasible for $|\alpha| \le \beta/M$. Maximize $q_C(\alpha\nu)$ on this interval, unbounded when $M = 0$. For a primal–dual optimal master, exact pricing either certifies global equality or returns a violated pattern block. Adding complete blocks and re-solving terminates finitely. We call this procedure *Pattern Completion*, a neural instance of restricted-master pricing (Desrosiers & Lübbecke, 2005; Mishkin et al., 2022). Patterns transfer between sources; Appendix D proves width-preserving reconstruction and the scaling edge cases. The source $C_0 = 0$ has the analytic pair $z_0 = \nu_0 = 0$.

Deep relaxations with feasible repair. When an exact source solve is unavailable, a sound relaxation supplies lower evidence, while a feasible repair supplies the upper witness. Let $D_Y^{\text{out}} = \langle Y, \xi\rangle - \frac{1}{2}\|\xi\|^2 - a$ be the target lower value from a globally valid support pair (ξ, a) for the outer relaxation (Appendix E). For bounded deep ReLU networks, retain the weights returned by the relaxed source solution and recompute the forward pass, converting output h into a neural output z. If $\|z - h\| \le \Delta$, the relaxed source gap is at most $\bar{g}_C$, and repair increases its regularization cost by at most χ, then

$$F_Y(\theta_{\text{repair}}) - D_Y^{\text{out}} \le \tfrac{1}{2}(\|Y - C\| + \sqrt{2\bar{g}_C})^2 + \|Y - h\|\Delta + \tfrac{1}{2}\Delta^2 + \chi. \tag{16}$$

The radius Δ propagates activation and weight–state product defects using ReLU hulls and McCormick envelopes (McCormick, 1976; Anderson et al., 2020); Appendix E gives the identity and global-scope conditions. Retained weights preserve a weight-only penalty, so $\chi = 0$. The experiments evaluate exact-source engines; the empirical scaling of this deep-relaxation extension is not studied here.

Algorithm 1 Neural MERIT for a fixed target Y

Require: Feasible target class, certified source engine, tolerance ε, optional bank $\mathcal{B}$
1: Obtain a feasible target network; initialize a globally valid lower bound.
2: Evaluate compatible bank certificates and retain their feasible endpoints.
3: **for** successively closer sources C_k **do**
4: Solve or certify C_k; retain source lower evidence and a feasible network or repair.
5: Transport the source certificate by (3), (9), or (16).
6: Verify numerical endpoints; update (19) and the associated neural witness.
7: **if** $U_{k,m} - L_{k,m} \le \varepsilon$ **then**
8: **return** witness and target certificate.
9: **return** the retained witness and interval if computation is interrupted.

5 Basis banks strengthen the same target guarantee

Retaining a surrogate's solution and proof makes its guarantee available without repeating the source solve. A Neural Basis Bank stores $\mathcal{B}_m = \{C_i, \theta_i, \text{lower evidence}_i\}_{i=1}^m$ from earlier paths or offline solves. The same transfer laws give $L_{\mathcal{B}_m}(Y) = \max_i L_i(Y)$ and $U_{\mathcal{B}_m}(Y) = \min_i U_i(Y)$.

Complementary dual directions and primal structures can strengthen the individual endpoints. Here $D_Y(\xi)$ is the convex-composite dual in (10), with support cost $\phi^*(\xi)$; inexact source solves change the recorded gaps, not this definition. Let ξ_i be globally dual feasible, $\bar{g}_i$ bound its source gap, and $\mathcal{P}_m$ unite the stored activation structures. Define

$$L_m^{\text{mix}}(Y) = \max_{\alpha \in \Delta_m} D_Y\left(\sum_i \alpha_i \xi_i\right), \qquad U_m^{\text{union}}(Y) = \text{optimal feasible union-master value}, \quad (17)$$

where $\Delta_m = \{\alpha \ge 0 : \sum_i \alpha_i = 1\}$. The dual hull preserves feasibility. In (14), the union reconstructs within $N \ge r_X$ (Appendix D); a smaller fixed-width class requires its own feasible synthesis rule.

Theorem 5.1 (Basis-bank improvement)**.** *If the union contains every stored representation at its recorded cost, then*

$$L_m^{\text{mix}}(Y) \le p^\star(Y) \le U_m^{\text{union}}(Y), \qquad U_m^{\text{union}}(Y) - L_m^{\text{mix}}(Y) \le \min_{i \le m} \tfrac{1}{2}\big(\|Y - C_i\| + \sqrt{2\bar{g}_i}\big)^2. \quad (18)$$

Nested banks cannot weaken either endpoint. The interval closes if the hull contains a target-optimal dual and the union contains a target-optimal representation.

Proof. Every stored dual is a hull vertex, and every stored primal is feasible in the union. Their optimized endpoints are therefore at least as strong as each individual pair. Apply Theorem 3.1. Nesting proves monotonicity; optimal containment proves closure. □

Composition can be strictly stronger than selecting a surrogate. For any $\alpha \in \Delta_m$ whose prediction mixture $\bar{z} = \sum_i \alpha_i z_i$ the union can realize at cost at most $\sum_i \alpha_i r_i$, the quadratic identities subtract $\frac{1}{2}\operatorname{Var}_\alpha(z) + \frac{1}{2}\operatorname{Var}_\alpha(\xi)$ from the averaged individual gap. Appendix F gives the exact inequality and examples. For interrupted computations, retaining the individual feasible endpoints as fallbacks preserves (18) even before the mixture or union is optimized.

The bank should strengthen the evidence already obtained for the current target. Let $L_k^{\text{class}}, U_k^{\text{class}}$ be retained bank-free bounds and U_k^{train} the trainer's incumbent. Their intersection with the bank gives

$$L_{k,m} = \max\{L_k^{\text{class}}, L_{\mathcal{B}_m}\}, \qquad U_{k,m} = \min\{U_k^{\text{train}}, U_k^{\text{class}}, U_{\mathcal{B}_m}\}. \quad (19)$$

Set $L_{\mathcal{B}_0} = -\infty$ and $U_{\mathcal{B}_0} = +\infty$ to recover the bank-free guarantee. Additional records strengthen the retained interval. Its upper endpoint identifies the network certified by (1); Algorithm 1 gives the procedure.

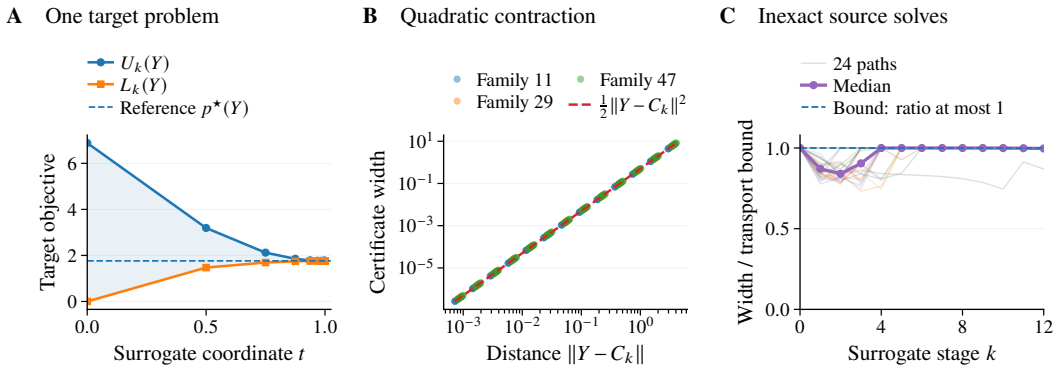

Figure 2: **Closer surrogates yield tighter guarantees for the same target problem.** Left: the first predeclared ReLU target, with every surrogate solution evaluated on it. Middle: computed certificates on all 24 tightly solved paths follow the theoretical quadratic law. Right: tolerance-controlled source solves preserve (9); the ratio uses the actual source gap and is bounded by one. Source uncertainty and outward rounding account for departures from the exact-source law.

6 Experiments: Solution Quality and Certification

Do closer certified surrogates improve target solutions and their guarantees? We test the distance law, nonconvex optimizer changes, and the bank's additional contribution. Independent target optima are used only for validation, never source selection or certificate acceptance. Protocols and verifiers appear in Appendices G–H.

Does the certificate gap follow the predicted distance law? We test quadratic contraction and controlled source accuracy on three 24×2 input families, each with 48 activation chambers, $\beta = 1/8$, and width 96 in (14). Eight three-dimensional target perturbations per family give 24 paths, disjoint from bank targets. From the exact zero source, take $C_k = (1 - 2^{-k})Y$, $k = 0, \ldots, 12$. One policy solves the full master tightly; the other warm-starts Pattern Completion until its verified source gap is at most $0.02\|Y - C_k\|^2$. Exact binary-rational arithmetic and 80-digit interval square roots verify the endpoints.

Every source meets its assigned accuracy (Figure 2; Appendix G.1, Figure 5). At stage 12, all 24 bank-free intervals are narrower than 4.91×10^{-7}, with median 3.82×10^{-7}; the largest tightly solved source gap is 4.22×10^{-9}. Intermediate solves can therefore stop before numerical closure while preserving verified target guarantees.

Do target solutions improve when the source optimizer changes? We test Theorem 2.2 in exact arithmetic by enumerating all $3^{10} = 59{,}049$ assignments of a 2–2–2–1 ReLU network with ternary weights, eight rational inputs, and penalty $\|\theta\|^2/32$. Deduplicating predictions while retaining their least penalty leaves 705 prediction–cost pairs. Twenty-four rational targets and 134 coordinates per path give 3,216 integer-arithmetic source minimizations. Every target objective is nonincreasing across optimizer changes, and every metric interval is valid and closes at the target. The median path visits three optimal witnesses. Using the source residual as the support direction leaves a positive terminal gap on all 24 targets, with median 0.3164; exact scalar-source information closes the metric certificate. Figure 3 and Figure 6 in Appendix G.3 compare these evidence types.

How much does a bank add to the same continuation? Holding the 24 ReLU paths fixed, we vary the prior bank from 0 to 4, 16, or 64 solved targets per family. Dual mixing and pattern-union optimization supply its endpoints. At width 10^{-2}, bank-free continuation first certifies at median stage five; 64 records certify 18 targets before continuation. At 10^{-3}, the median first stage falls from seven to five. All paths reach each tolerance (Table 1). Stage counts include the result of target-specific bank optimization; they measure certificate availability, not wall-clock speedup.

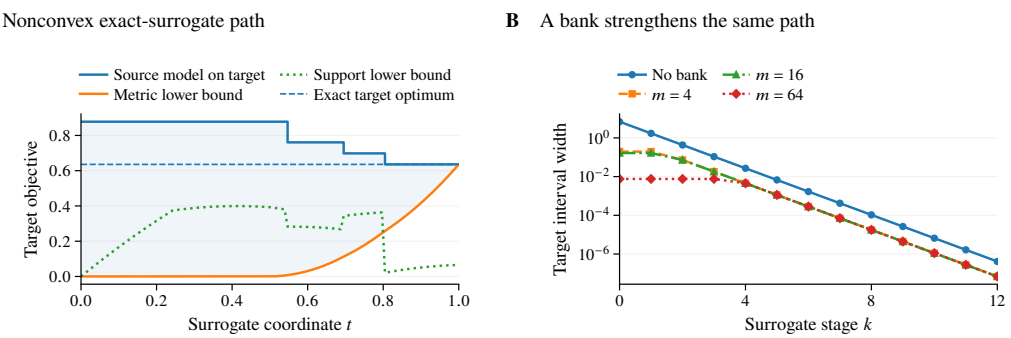

Figure 3: **Continuation and stored surrogates strengthen a fixed target certificate.** Left: exact nonconvex surrogate minimizers switch as t increases, yet their target objective never rises. The scalar metric certificate reaches the exact target value; the residual-direction support gap need not vanish. Right: on the first predeclared ReLU target, bank sizes $m = 0, 4, 16, 64$ strengthen the *same* continuation. A bank can improve the initial interval and the remaining stage requirement; bank-free continuation still converges independently.

Table 1: **Basis banks reduce the continuation still needed for a target guarantee.** Medians are over all 24 paths using the tight-source policy; every path succeeds at every displayed tolerance. Parentheses count targets already certified before continuation.

Target width	No bank	$m = 4$	$m = 16$	$m = 64$
10^{-2}	5 (0)	4 (0)	4 (5)	0 (18)
10^{-3}	7 (0)	5 (0)	5 (0)	5 (5)
10^{-5}	10 (0)	9 (0)	9 (0)	9 (0)

What does the certificate reveal when training has plateaued? We compare a trainer's achieved objective with the evidence available about its quality. A frozen ImageNet-pretrained ResNet-18 representation of 1,024 CIFAR-10 cat/dog images is reduced to four principal components (He et al., 2016; Deng et al., 2009; Krizhevsky, 2009). The trainable class is a width-three quantized ReLU head with 321 signed/scaled atoms. Exhaustive pair-plus-atom search evaluates 16,589,601 configurations. A safeguarded coordinate trainer uses 20 fixed restarts, evaluates all 321 slot replacements on the complete data, and accepts only strict decreases. Its incumbent reaches 0.448682 at update 29 and stays there through update 2,500; the exhaustive reference is 0.448645. Figure 4 intersects a bank-free convex outer bound with metric certificates from solved target perturbations.

Across 128 held-out perturbations (Appendix G.4, Table 2), median interval width decreases from 0.5306 without a bank to 0.02111 at 10,000 records; the maximum becomes 0.02457. Stored witnesses match the exhaustive reference within 2×10^{-9} on $28, 107, 126$, and 128 queries at bank sizes $10, 100, 1{,}000$, and $10{,}000$. These reference-optimal witnesses retain measurable lower-bound uncertainty. The certificate concerns the finite head on fixed features.

On 192 independently verified ReLU queries, 64 records reduce median width from 0.08275 for separately selected stored endpoints to 0.004914 with composition (Appendix G.5, Figure 7 and Table 3). A separate Pattern Completion study reduces new pattern additions from 6–11 to zero on four audited targets. Figure 8 in Appendix G.6 extends finite-class and identity-design calibration to 10^5 records, measuring how additional solved surrogates strengthen the same guarantee.

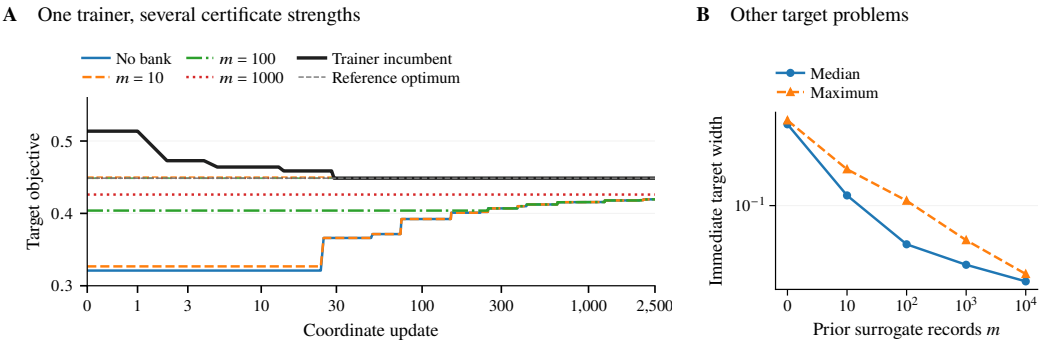

Figure 4: **A training plateau leaves uncertainty that a certificate quantifies.** Left: the complete 2,500-update trainer trace, independent reference optimum, and bank-free/bank-enhanced intervals for the finite head. Right: the same banks reduce immediate width on 128 other targets, relative to zero-model/zero-lower-bound evidence. A reference-optimal stored witness may still have a nonzero certificate gap, as the intervals show.

7 Relationship to Prior Work

Problem-space approximation and continuation. MERIT and basis banks transport case-dependent guarantees from nearby solved quadratic problems (Soltanalian & Stoica, 2014a;b; Soltanalian & Hassibi, 2015). Inverse optimization instead chooses perturbations that rationalize observed decisions (Ahuja & Orlin, 2001; Keshavarz et al., 2011). Gaussian continuation analyzes optimization through smoothed objectives (Mobahi & Fisher, 2015); recent neural predictor–corrector methods learn homotopy step and correction policies (Mai et al., 2026). Here continuation transports global evidence to one fixed target.

Global neural optimization and certification. Haeffele & Vidal (2017) give global-optimality conditions for positively homogeneous models and regularizers; Yun et al. (2018) characterize global critical points of deep linear networks and give sufficient conditions in a restricted nonlinear function space. Convex formulations provide exact engines for regularized classes (Bach, 2017; Pilanci & Ergen, 2020; Ergen & Pilanci, 2021a), with efficient solvers, broader formulations, and approximation guarantees (Mishkin et al., 2022; Bai et al., 2023; Kim & Pilanci, 2024; Prakhya et al., 2025). Deep and quantized classes also admit specialized exact constructions (Ergen & Pilanci, 2021b; Bartan & Pilanci, 2021). Relaxation-based verification supplies related proof techniques (Bunel et al., 2020; Anderson et al., 2020); incremental verification reuses conflicts across verification instances (Elsaleh et al., 2026).

Mathematical foundations. Fenchel duality and Moreau envelopes supply the analytic foundations (Rockafellar, 1970; Moreau, 1965; Bauschke & Combettes, 2017); safe screening also uses dual evidence for certified exclusions (Ndiaye et al., 2017). Here that evidence certifies a feasible neural solution after a target perturbation; closed convex primal–dual pairs give the exact quadratic path.

8 Conclusion

How much better could the same network class have been trained? Neural MERIT answers with a feasible network and a global-suboptimality bound. Exact straight continuation improves target objectives even for nonconvex classes; closed primal–dual pairs give quadratic certificates. Analytically solved or progressively certified sources provide bank-free guarantees, and stored surrogates strengthen them. Source distance and uncertainty control certificate quality; certification cost determines whether this route is preferable to a direct solve. These results use squared-loss target geometry: broader losses require their own transfer analysis, and source-certificate cost and tightness remain architecture-dependent.

REPRODUCIBILITY STATEMENT

Complete proofs and assumptions appear in Appendices A–F. Appendix G specifies the training classes, targets, source policies, and numerical tolerances. Appendix H documents arithmetic verification and the released scripts. The supplementary archive contains target and source arrays, actual primal/dual witnesses, source-stage records, and optimizer-free replay code. All plots are generated from these records with external LaTeX text rendering.

ETHICS STATEMENT

The certificates concern the empirical optimization problem and its declared feasible class. They do not establish population accuracy, fairness, privacy, robustness, or deployment safety. Reporting the class, objective normalization, numerical allowances, and unresolved interval is part of using the certificate responsibly. The experiments use public image data and synthetic numerical tasks; no human-subject study is performed.

AI USE STATEMENT

Generative AI assisted with mathematical formulation, derivations, literature retrieval, experimental code, manuscript preparation, and figure development. Wolfram Language and local symbolic software checked specified identities and finite algebraic reductions; independent arithmetic replay checked numerical certificates. These checks supplement the written proofs and do not constitute formal verification of every theorem or experimental assumption. The authors are responsible for the submitted claims, proofs, references, and artifacts.

REFERENCES

Ravindra K. Ahuja and James B. Orlin. Inverse optimization. *Operations Research*, 49(5):771–783, 2001. doi: 10.1287/opre.49.5.771.10607. URL https://doi.org/10.1287/opre.49.5.771.10607.

Ross Anderson, Joey Huchette, Will Ma, Christian Tjandraatmadja, and Juan Pablo Vielma. Strong mixed-integer programming formulations for trained neural networks. *Mathematical Programming*, 183:3–39, 2020. doi: 10.1007/s10107-020-01474-5.

Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

Yatong Bai, Tanmay Gautam, and Somayeh Sojoudi. Efficient global optimization of two-layer ReLU networks: Quadratic-time algorithms and adversarial training. *SIAM Journal on Mathematics of Data Science*, 5(2):446–474, 2023. doi: 10.1137/21M1467134.

Burak Bartan and Mert Pilanci. Training quantized neural networks to global optimality via semidefinite programming. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 694–704. PMLR, 2021.

Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. Springer, second edition, 2017. doi: 10.1007/978-3-319-48311-5.

Avrim L. Blum and Ronald L. Rivest. Training a 3-node neural network is NP-complete. *Neural Networks*, 5(1):117–127, 1992. doi: 10.1016/S0893-6080(05)80010-3.

Digvijay Boob, Santanu S. Dey, and Guanghui Lan. Complexity of training ReLU neural network. *Discrete Optimization*, 44:100620, 2022. doi: 10.1016/j.disopt.2020.100620.

Rudy Bunel, Jingyue Lu, Ilker Turkaslan, Philip H. S. Torr, Pushmeet Kohli, and M. Pawan Kumar. Branch and bound for piecewise linear neural network verification. *Journal of Machine Learning Research*, 21(42):1–39, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

Jacques Desrosiers and Marco E. Lübbecke. A primer in column generation. In Guy Desaulniers, Jacques Desrosiers, and Marius M. Solomon (eds.), *Column Generation*, pp. 1–32. Springer, 2005. doi: 10.1007/0-387-25486-2_1.

Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936. doi: 10.1007/BF02288367.

Raya Elsaleh, Liam Davis, Haoze Wu, and Guy Katz. Incremental neural network verification via learned conflicts. *arXiv preprint arXiv:2603.12232*, 2026. doi: 10.48550/arXiv.2603.12232.

Tolga Ergen and Mert Pilanci. Convex geometry and duality of over-parameterized neural networks. *Journal of Machine Learning Research*, 22(212):1–63, 2021a.

Tolga Ergen and Mert Pilanci. Global optimality beyond two layers: Training deep ReLU networks via convex programs. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2993–3003, 2021b.

Benjamin D. Haeffele and René Vidal. Global optimality in neural network training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7331–7339, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Haeffele_Global_Optimality_in_CVPR_2017_paper.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016. doi: 10.1109/CVPR.2016.90.

Arezou Keshavarz, Yang Wang, and Stephen Boyd. Imputing a convex objective function. In *2011 IEEE International Symposium on Intelligent Control*, pp. 613–619, 2011. doi: 10.1109/ISIC.2011.6045410. URL https://doi.org/10.1109/ISIC.2011.6045410.

Sungyoon Kim and Mert Pilanci. Convex relaxations of ReLU neural networks approximate global optima in polynomial time. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24458–24485. PMLR, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Jiayao Mai, Bangyan Liao, Zhenjun Zhao, Yingping Zeng, Haoang Li, Javier Civera, Tailin Wu, Yi Zhou, and Peidong Liu. Neural predictor-corrector: Solving homotopy problems with reinforcement learning. *arXiv preprint arXiv:2602.03086*, 2026. doi: 10.48550/arXiv.2602.03086.

Garth P. McCormick. Computability of global solutions to factorable nonconvex programs: Part i—convex underestimating problems. *Mathematical Programming*, 10:147–175, 1976. doi: 10.1007/BF01580665.

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

Leon Mirsky. Symmetric gauge functions and unitarily invariant norms. *The Quarterly Journal of Mathematics*, 11(1):50–59, 1960. doi: 10.1093/qmath/11.1.50.

Aaron Mishkin, Arda Sahiner, and Mert Pilanci. Fast convex optimization for two-layer ReLU networks: Equivalent model classes and cone decompositions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 15770–15816, 2022.

Hossein Mobahi and John W. Fisher, III. A theoretical analysis of optimization by Gaussian continuation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 29, pp. 1205–1211, 2015. doi: 10.1609/aaai.v29i1.9356.

Jean-Jacques Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965. doi: 10.24033/bsmf.1625.

Eugene Ndiaye, Olivier Fercoq, Alexandre Gramfort, and Joseph Salmon. Gap safe screening rules for sparsity enforcing penalties. *Journal of Machine Learning Research*, 18(128):1–33, 2017.

Art B. Owen. Randomly permuted (t, m, s)-nets and (t, s)-sequences. In Harald Niederreiter and Peter J.-S. Shiue (eds.), *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, volume 106 of *Lecture Notes in Statistics*, pp. 299–317. Springer, 1995. doi: 10.1007/978-1-4612-2552-2_19.

Neal Parikh and Stephen Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3): 127–239, 2014. doi: 10.1561/2400000003.

Mert Pilanci and Tolga Ergen. Neural networks are convex regularizers: Exact polynomial-time convex optimization formulations for two-layer networks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7695–7705, 2020.

Karthik Prakhya, Tolga Birdal, and Alp Yurtsever. Convex formulations for training two-layer ReLU neural networks. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=e0X9l4kecx.

R. Tyrrell Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

I. M. Sobol'. On the distribution of points in a cube and the approximate evaluation of integrals. *USSR Computational Mathematics and Mathematical Physics*, 7(4):86–112, 1967. doi: 10.1016/0041-5553(67)90144-9.

Mojtaba Soltanalian and Babak Hassibi. Beyond semidefinite relaxation: Basis banks and computationally enhanced guarantees. In *2015 IEEE International Symposium on Information Theory*, pp. 336–340, 2015. doi: 10.1109/ISIT.2015.7282472.

Mojtaba Soltanalian and Petre Stoica. Designing unimodular codes via quadratic optimization. *IEEE Transactions on Signal Processing*, 62(5):1221–1234, 2014a. doi: 10.1109/TSP.2013.2296883.

Mojtaba Soltanalian and Petre Stoica. MERIT: A monotonically error-bound improving technique for unimodular quadratic programming. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5656–5660, 2014b.

Hassler Whitney. Analytic extensions of differentiable functions defined in closed sets. *Transactions of the American Mathematical Society*, 36(1):63–89, 1934. doi: 10.1090/S0002-9947-1934-1501735-3.

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

Chulhee Yun, Suvrit Sra, and Ali Jadbabaie. Global optimality conditions for deep neural networks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJk7Gf-CZ.

A METRIC TRANSFER AND EXACT NONCONVEX CONTINUATION

A source guarantee is useful for the target only if the change in optimal value can be controlled. For the squared-loss family (2), distance geometry provides that control without restricting the neural parameterization. Predictions belong to a finite-dimensional real Hilbert space, and the effective feasible domain is nonempty. All distances use the same inner product as the loss. If $R \geq R_0 > -\infty$ is the available assumption, replace R by $R - R_0$ and shift every objective endpoint by R_0 afterward. The parameter domain itself can be disconnected, unbounded, and nonclosed.

A.1 DISTANCE REPRESENTATION AND ONE-SOURCE TRANSFER

The penalty must travel with the prediction when the target changes. Adding its square root as one coordinate turns the complete objective into squared distance.

For $\mathcal{S} = \{(z(\theta), \sqrt{2R(\theta)}) : R(\theta) < \infty\}$,

$$\operatorname{dist}((Y,0),\mathcal{S})^2 = \inf_{\theta: R(\theta)<\infty} \{\|z(\theta) - Y\|^2 + 2R(\theta)\} = 2p^\star(Y). \tag{20}$$

The infimum of squared nonnegative numbers is the square of their infimum. A finite-cost feasible parameter makes this distance finite for every Y. For any nonempty set S and every $s \in S$, $\|x - s\| \leq \|x - y\| + \|y - s\|$. Taking the infimum and then interchanging x, y proves

$$|\operatorname{dist}(x,S) - \operatorname{dist}(y,S)| \leq \|x - y\|. \tag{21}$$

No closest point or closedness is used.

At source C, let $a = \sqrt{2\ell_C}$ and $b = \sqrt{2u_C}$, where $0 \leq \ell_C \leq p^\star(C) \leq u_C = F_C(\theta_C)$. The two triangle inequalities yield

$$[a - \delta]_+ \leq \sqrt{2p^\star(Y)}, \qquad \sqrt{2F_Y(\theta_C)} \leq b + \delta.$$

Their difference is bounded by

$$b + \delta - [a - \delta]_+ = \begin{cases} b - a + 2\delta, & a \geq \delta, \\ b + \delta, & a < \delta, \end{cases} \leq b - a + 2\delta. \tag{22}$$

For $a = b = s$, the expression equals $\delta + \min(\delta, s)$. This proves every part of Theorem 2.1 except sharpness.

For sharpness, take the unregularized one-dimensional prediction class $\{-r, r\}$, source $C = 0$, and query $Y = \delta \in (0, r)$. The source has global value $r^2/2$; choose its globally optimal witness $z_C = -r$. Then the transported upper root is $r + \delta$ and the lower root is $r - \delta$. Their difference is exactly 2δ. Both optimality and the constant are independent of how this finite class is parameterized. The corresponding objective-width is exactly $2r\delta$, so scalar exact-source information cannot give a universal quadratic objective envelope at positive source residual.

More generally, an exact source with root value s gives the sufficient objective-width

$$U_C(Y) - L_C(Y) \leq \tfrac{1}{2}\{(s+\delta)^2 - [s-\delta]_+^2\} = \begin{cases} 2s\delta, & \delta \leq s, \\ \tfrac{1}{2}(s+\delta)^2, & \delta > s. \end{cases} \tag{23}$$

These are upper bounds on the width obtained from the stored witness; evaluating the witness can improve them.

A.2 AN EXACT LIFTED REPRESENTATION

Lower relaxations and feasible networks can be compared within a common geometry. The energy coordinate below preserves the original objective when taking a convex hull.

For unregularized squared loss and $\mathcal{M} = \{z(\theta) : \theta \in \Theta\}$, define

$$\mathcal{T} = \overline{\operatorname{conv}}\{(z, \tfrac{1}{2}\|z\|^2) : z \in \mathcal{M}\}.$$

Then

$$p^\star(Y) = \tfrac{1}{2}\|Y\|^2 + \inf_{(z,s)\in\mathcal{T}} \{s - \langle Y, z\rangle\}. \tag{24}$$

Expanding the square gives equality on the generators. An affine functional has the same infimum on a set and its convex hull; continuity preserves the infimum under closure. The infimum here is finite, being bounded below by $-\|Y\|^2/2$. The energy coordinate is essential: replacing the class $\{-1, 1\}$ by its ordinary convex hull would change the value at zero from $1/2$ to zero. With a regularizer, use the generators $(z(\theta), \|z(\theta)\|^2/2, R(\theta))$ and the affine objective $s - \langle Y, z\rangle + r$.

For scalar center intervals $a_i \le d(C_i) \le b_i$ of a nonnegative 1-Lipschitz function, the extremal compatible envelopes are

$$\underline{d}(Y) = \max\{0, \sup_i(a_i - \|Y - C_i\|)\}, \qquad \overline{d}(Y) = \inf_i(b_i + \|Y - C_i\|).$$

For a finite consistent bank these are 1-Lipschitz, satisfy every interval, and bound any other consistent extension (McShane, 1934; Whitney, 1934). A stored neural witness contains additional information: its evaluated target root is no larger than the corresponding scalar upper cone, and can be strictly smaller. Neural MERIT always retains the actual feasible witness.

A.3 Complete proof of Theorem 2.2

The target objective can improve even when the source optimizer jumps. Comparing two global source optimality inequalities establishes this directly, without differentiating an optimizer map.

Set $v = Y - C_0$ and $C_t = C_0 + tv$. For every feasible θ with finite cost,

$$F_{C_t}(\theta) = F_{C_0}(\theta) - t\langle v, z(\theta) - C_0\rangle + \tfrac{1}{2}t^2\|v\|^2 \tag{25}$$

$$= (1-t)F_{C_0}(\theta) + tF_Y(\theta) - \tfrac{1}{2}t(1-t)\|v\|^2. \tag{26}$$

The final term is independent of θ. Choose any exact source minimizers θ_s, θ_t at $0 \le s < t < 1$. With A and b as in the main proof, their optimality gives

$$A_t - tb_t \le A_s - tb_s, \qquad A_s - sb_s \le A_t - sb_t.$$

Consequently $s(b_t - b_s) \le A_t - A_s \le t(b_t - b_s)$. Since $t - s > 0$, this implies $b_t - b_s \ge 0$, and

$$F_Y(\theta_t) - F_Y(\theta_s) \le -(1-t)(b_t - b_s) \le 0. \tag{27}$$

No derivative, tie-breaking rule, or unique optimizer is used. This argument also applies to any common-domain affine objective homotopy, up to additive terms independent of the decision.

Let $d_t = \sqrt{2p^\star(C_t)}$ and $\delta_t = (1-t)\|v\|$. By (21), $d_t \ge d_s - (t-s)\|v\|$, whereas $\delta_s - \delta_t = (t-s)\|v\|$. Therefore $d_t - \delta_t \ge d_s - \delta_s$, proving monotonicity of the clipped, squared lower endpoint. The upper endpoint is nonincreasing by (27). Both bound $p^\star(Y)$, and Theorem 2.1 makes their root-width at most $2\delta_t \to 0$. Since the upper roots are bounded by their first finite value, the objective-width also tends to zero. Attainment at Y is unnecessary for this limiting argument; attainment is assumed only at the source parameters actually used.

For inexact sources, $C_k \to Y$ and $\sqrt{2u_k} - \sqrt{2\ell_k} \to 0$ imply closing target intervals by (22). For arbitrary source order, retaining the largest lower and smallest feasible upper endpoints makes the reported sequence monotone. The root-width is then no larger than the best earlier source envelope. Proximity by itself is not sufficient if the source uncertainty does not decrease.

B Global support evidence and quadratic transport

Directional lower evidence can give a tighter target guarantee than a scalar source value. The support inequality provides such evidence for every feasible network; completing the square identifies its exact interaction with the returned witness.

B.1 Weak duality without a convex neural class

For a finite support bound $a \ge \mathsf{H}(\xi)$ from (6), every feasible finite-cost network satisfies

$$F_Y(\theta) \ge \tfrac{1}{2}\|z(\theta) - Y\|^2 + \langle \xi, z(\theta)\rangle - a \tag{28}$$

$$= \tfrac{1}{2}\|z(\theta) - (Y - \xi)\|^2 + \langle Y, \xi\rangle - \tfrac{1}{2}\|\xi\|^2 - a \ge D_Y(\xi, a). \tag{29}$$

Taking the infimum proves the lower bound. A feasible network with cost upper $r \geq R(\theta)$ proves the upper bound $U_Y = \|z - Y\|^2/2 + r$. In particular,

$$\eta = r + a - \langle z, \xi \rangle \geq r - R(\theta) \geq 0.$$

Subtracting and expanding proves

$$U_Y - D_Y = \tfrac{1}{2}\|Y - z - \xi\|^2 + \eta.$$

At C, $g_C = \eta + \|C - z - \xi\|^2/2$. Thus the residual $\rho = C - z - \xi$ satisfies $\|\rho\| \leq \sqrt{2g_C}$. Replacing Y by $C + (Y - C)$ and expanding the square gives (8). For any upper bound $\bar{g}_C \geq g_C$,

$$U_Y - D_Y \leq \bar{g}_C + \tfrac{1}{2}\delta^2 + \sqrt{2\bar{g}_C}\,\delta = \tfrac{1}{2}(\delta + \sqrt{2\bar{g}_C})^2. \tag{30}$$

The actual residual identity can be substantially tighter than its worst-case radial envelope.

If $g_C = 0$, both nonnegative terms in its decomposition vanish: $\eta = 0$ and $C = z + \xi$. Hence the transported gap is exactly $\delta^2/2$. The exact statement concerns a *closed primal–lower pair*. Knowing only that a network minimizes the source objective does not imply that the chosen lower family attains that value.

B.2 Sharpness and the nonconvex evidence floor

Two examples identify what the quadratic guarantee can achieve: the first attains its radial bound, and the second shows why an exactly solved nonconvex source may still need scalar evidence in addition to a support dual.

The quadratic radial envelope is sharp. Use the singleton class $z = 0$, $R = 0$, which has exact support value zero for every ξ. Take $C = 0$, $\xi = -su$, and $Y = \delta u$, where $\|u\| = 1$ and $s, \delta \geq 0$. Then $g_C = s^2/2$, $\rho = su$, and the target gap is exactly $(\delta + s)^2/2$.

For the finite class $\mathcal{M} = \{-1, 1\}$ with zero penalty, $\mathsf{H}(\xi) = |\xi|$. At $Y = 0$, the exact neural optimum is $1/2$, while

$$\max_{\xi \in \mathbb{R}} \{-\xi^2/2 - |\xi|\} = 0.$$

Thus even exact support evaluation has a $1/2$ primal–dual floor. Scalar values obtained by solving nearby finite neural problems still give the convergent metric continuation of Theorem 2.2. The two forms of evidence have different strengths and can be intersected.

For a given source record with effective center $\widetilde{C}$, (7) also describes exactly where that pair meets a chosen tolerance: $U_Y - D_Y \leq \varepsilon$ precisely when $\eta \leq \varepsilon$ and $\|Y - \widetilde{C}\| \leq \sqrt{2(\varepsilon - \eta)}$. This is a property of the fixed evidence pair, not a claim that its source optimizer remains optimal on this ball.

B.3 Synthesis costs and numerical allowances

A certificate must cover the cost of the actual returned network, including reconstruction and numerical evaluation. The same cost convention must therefore be used at the source and the target.

In a convex prediction formulation, a returned neural representation may have cost $r > \phi(z)$. Keep that same representation cost at both source and target. The nonnegative slack $r - \phi(z)$ enters both gaps, so (30) holds with the recorded source gap. Using the smaller convex objective while returning an uncompressed higher-cost network would not certify the returned network.

If $\overline{U}_C \geq U_C$ and $\underline{D}_C \leq D_C$ are verified numerical endpoints, then $\bar{g}_C = \overline{U}_C - \underline{D}_C$ bounds the exact source gap. Target endpoint evaluations likewise need outward bounds. For an approximately feasible dual, first establish feasibility, weaken its support value by a proved allowance, or substitute a verified feasible dual; then apply any objective-evaluation allowance to the endpoint. The numerical experiments follow these one-sided rules.

C Convex-composite paths and intermediate accuracy

Closed primal–dual sources turn the quadratic identity into a complete continuation guarantee. We establish when such pairs exist, how their target endpoints evolve, and how source accuracy can be

chosen as the target approaches. To interpret this path as neural training, Theorem 3.2 assumes a fixed-ϕ exact reformulation and cost-preserving neural realization at every source on the segment, including Y.

C.1 Existence, uniqueness, and source identities

Let $\phi : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be proper, lower semicontinuous, and convex. It has an affine minorant, so $P_C(z) = \phi(z) + \|z - C\|^2/2$ is coercive. Its strong convexity gives a unique minimizer z_C. Since the quadratic is continuous everywhere, Fenchel duality gives equality with the attained dual optimum in (10). The optimality condition is

$$C - z_C \in \partial\phi(z_C).$$

Writing $\xi_C = C - z_C$, Fenchel equality gives $\phi(z_C) + \phi^*(\xi_C) = \langle z_C, \xi_C \rangle$. This is the zero-gap specialization of the support decomposition. Moreau decomposition identifies $z_C = \mathrm{prox}_{\phi}(C)$ and $\xi_C = \mathrm{prox}_{\phi^*}(C)$ (Moreau, 1965; Bauschke & Combettes, 2017).

For completeness, if z_C, z_B are the two proximal predictions, monotonicity of $\partial\phi$ gives

$$\langle (C - z_C) - (B - z_B), z_C - z_B \rangle \geq 0, \qquad \|z_C - z_B\|^2 \leq \langle C - B, z_C - z_B \rangle.$$

The proximal map is firmly nonexpansive, and its residual map is firmly nonexpansive by the same expansion. Consequently both are 1-Lipschitz. The Moreau envelope $p(C) = \min_z P_C(z)$ is convex, continuously differentiable, and has gradient ξ_C, with Lipschitz constant one (Parikh & Boyd, 2014).

C.2 Complete finite-difference proof of Theorem 3.2

A shrinking gap alone would allow both endpoints to move in the same direction. To prove separate improvement of the lower and upper bounds, we control the change in the lower tangent model and compare it with the shrinking quadratic remainder.

Let $v = Y - C_0$, $C_t = C_0 + tv$, $g_t = \nabla p(C_t)$, and define

$$L_t = p(C_t) + \langle g_t, Y - C_t \rangle, \qquad U_t = L_t + \tfrac{1}{2}\|Y - C_t\|^2.$$

The first equality is $D_Y(\xi_{C_t})$ because the target enters the dual linearly; the second is Theorem 3.1. Fix $0 \leq s < t \leq 1$, put $\Delta = C_t - C_s = (t - s)v$ and $q_t = Y - C_t = (1 - t)v$. Then

$$L_t - L_s = p(C_t) - p(C_s) - \langle g_s, \Delta \rangle + \langle g_t - g_s, q_t \rangle. \tag{31}$$

The first three terms are nonnegative by convexity. The final term is nonnegative by gradient monotonicity, since q_t is a nonnegative multiple of Δ. Thus $L_t \geq L_s$.

For the upper endpoint, 1-smoothness bounds the first three terms in (31) by $\|\Delta\|^2/2$. Nonexpansiveness bounds its last term by $\|\Delta\|\|q_t\|$. Therefore

$$U_t - U_s \leq \tfrac{1}{2}(t - s)^2\|v\|^2 + (t - s)(1 - t)\|v\|^2 + \tfrac{1}{2}\{(1 - t)^2 - (1 - s)^2\}\|v\|^2 = 0.$$

The algebraic cancellation is exact. At $t = 1$, the pair is target-optimal, and the width identity gives convergence. Continuity follows from the displayed formulas for $p, \nabla p$, not from continuity of arbitrary neural parameterizations. Parameter permutations or balancing can still make a chosen weight representation discontinuous.

C.3 Distance-dependent source tolerances

Source accuracy should decrease as the remaining target discrepancy decreases. The following stage bound quantifies the accuracy schedule in (12).

For $\delta_k = q^k\delta_0$ and $\bar{g}_k \leq \kappa\delta_k^2$, (30) gives (12). Let $A_\kappa = (1 + \sqrt{2\kappa})^2/2$. For $\delta_0 > 0$ and $\varepsilon > 0$, a sufficient stage index is

$$k \geq \max\left\{0, \left\lceil \frac{\log(A_\kappa\delta_0^2/\varepsilon)}{2\log(1/q)} \right\rceil \right\}. \tag{32}$$

If $\delta_0 = 0$, only source accuracy remains. If a verified irreducible allowance e_k is added, replace ε by $\varepsilon - e_k$ when positive. The bound counts successfully certified stages, not solver iterations. It remains conditional on the engine meeting the assigned source tolerance.

Exact equally spaced sources also satisfy the warm-start energy identity

$$\sum_{k=1}^{m} \tfrac{1}{2}\|C_k - C_{k-1}\|^2 \geq \frac{\|Y - C_0\|^2}{2m}, \tag{33}$$

with equality for equal collinear increments summing to $Y - C_0$. This is Cauchy–Schwarz applied to the increments. It quantifies the total transported initialization gap of exact neighboring pairs, without assigning a universal runtime to resolving those gaps.

D GLOBALLY SOLVABLE NEURAL SOURCE CLASSES

The transfer theorems require a feasible source solution and global lower information. The constructions here supply both without a basis bank, and specify the architecture and objective conditions under which they do so.

D.1 DEEP LINEAR EXACT SOURCES

Although the weights enter through a matrix product, the attainable predictions have a rank characterization. This permits an exact source solve followed by a neural factorization.

Use samples in rows, $X \in \mathbb{R}^{n\times d}$ and $C \in \mathbb{R}^{n\times q}$. An unconstrained depth-K linear network predicts XA, where A is a product of layer matrices and has rank at most the narrowest effective width r. Every matrix of rank at most r with compatible input/output dimensions admits such a factorization by padding a rank factorization through the hidden layers. Hence the reachable prediction set is

$$\{Z \in \mathbb{R}^{n\times q} : P_X Z = Z,\ \operatorname{rank} Z \leq r\}, \qquad P_X = XX^\dagger.$$

Indeed, Z in that set is attained by $A = X^\dagger Z$, whose rank is at most r, followed by a product factorization. Orthogonal decomposition gives

$$\|C - Z\|_F^2 = \|(I - P_X)C\|_F^2 + \|P_X C - Z\|_F^2.$$

Truncated singular-value decomposition of $P_X C$ attains the second minimum and proves (13) (Eckart & Young, 1936; Mirsky, 1960). Repeated singular values need not give a unique optimizer; the optimal value and Theorem 2.2 are unaffected. This formula has no weight regularizer or additional layer constraints. A different objective needs its own exact source characterization.

D.2 TWO-LAYER ATOMIC FORMULATION AND SUFFICIENT WIDTH

Positive homogeneity separates each neuron's direction from its contribution to the prediction. The resulting atomic formulation supplies a global source optimum, provided its representation fits the declared neural width.

For one neuron with $u \neq 0$, positive homogeneity gives

$$a(Xu)_+ = \operatorname{sign}(a)|a|\|u\|\,(X(u/\|u\|))_+, \qquad \tfrac{1}{2}(\|u\|^2 + a^2) \geq |a|\|u\|.$$

Equality is attained by balancing the norms. Zero neurons contribute neither prediction nor cost. The compact symmetric atomic set $\mathcal{A}_X = \{\pm(Xu)_+ : \|u\| \leq 1\}$ spans a finite-dimensional subspace E. Its convex hull is compact and contains zero in its relative interior. Its gauge is a norm on E and $+\infty$ outside E. Therefore unrestricted-width training has exact value

$$\min_{z\in E} \tfrac{1}{2}\|z - C\|^2 + \beta\gamma_X(z).$$

This minimum exists. For a nonzero z, the boundary point $z/\gamma_X(z)$ belongs to an exposed supporting face of $\operatorname{conv}\mathcal{A}_X$ of dimension at most $r_X - 1$, where $r_X = \dim E$. Carathéodory's theorem on that face gives at most r_X atoms. Balancing constructs a neural realization with cost exactly $\beta\gamma_X(z)$. The zero prediction uses zero neurons. Thus width $N \geq r_X$ suffices for equality with (14). The degenerate case $r_X = 0$ is the identically zero prediction class.

The conjugate of $\beta\gamma_X$ is the indicator of

$$\{\nu : \sup_{a\in\mathcal{A}_X} \langle \nu, a\rangle \leq \beta\} = \{\nu : M_X(\nu) \leq \beta\}.$$

The everywhere-continuous quadratic provides strong duality and yields (15). These representations follow the established convex neural formulations (Bach, 2017; Pilanci & Ergen, 2020; Ergen & Pilanci, 2021a); here they supply globally certified surrogate pairs.

D.3 FINITE PATTERN MASTER AND RECONSTRUCTION

To compute the atomic optimum and return an executable network, group directions by their activation pattern. Reconstruction then connects the cone-program objective to a feasible neural upper endpoint.

For $D = \mathrm{diag}(1_{Xu\geq 0})$, define $K_D = \{v : (2D - I)Xv \geq 0\}$. Boundary directions may belong to several cones, but their predictions agree wherever $Xu = 0$. The complete pattern formulation is

$$\min_{v_D, w_D \in K_D} \frac{1}{2} \left\| \sum_{D \in \mathcal{D}_X} DX(v_D - w_D) - C \right\|_2^2 + \beta \sum_D (\|v_D\|_2 + \|w_D\|_2). \tag{34}$$

For $v \neq 0$, a positive neuron $u = v/\sqrt{\|v\|}$, $a = \sqrt{\|v\|}$ predicts $(Xv)_+$ and has balanced penalty $\beta\|v\|$. A negative vector w is represented with $a = -\sqrt{\|w\|}$. This reconstructs the master as a neural network. Grouping neurons of the same sign in a cone preserves prediction and cannot increase the norm penalty, giving the reverse direction.

Restricting $\mathcal{D}_X$ to $\mathcal{P}$ gives a neural-feasible master. A direct reconstruction uses at most $2|\mathcal{P}|$ neurons. To maintain the smaller sufficient width r_X, hold the reconstructed directions fixed and minimize their nonnegative masses subject to preserving their prediction. This finite linear program is feasible, has a cost-minimizing basic solution with at most $\mathrm{rank}(A) \leq r_X$ nonzero coefficients, and cannot increase the original mass. Balancing that representation preserves the prediction and respects width $N \geq r_X$. The experiments use width 96 for 48 patterns, so direct reconstruction already fits the declared width.

Floating-point vectors with small cone violations are always evaluated through the true ReLU forward map when reporting the upper endpoint. This produces a feasible neural witness even if the cone-master prediction has a small numerical mismatch. Its actual cost, not an unverified master objective, is retained.

D.4 SAFE SCALING, EXACT PRICING, AND FINITE CLOSURE

A restricted master leaves some dual constraints unchecked. Global pricing either verifies them or supplies a scaling that makes the dual globally feasible, so an unfinished master can still provide a valid source certificate.

Continuity on the compact unit ball makes $M_X(\nu)$ finite. Let $0 \leq M < \infty$ be a verified upper bound. Positive homogeneity yields

$$M_X(\alpha\nu) = |\alpha| M_X(\nu) \leq \beta \quad \text{whenever } |\alpha| \leq \beta/M.$$

If $M = 0$, every real scale is feasible. If $\nu \neq 0$, the best scale is

$$\alpha^* = \mathrm{clip}_{[-\beta/M, \beta/M]} \frac{\langle C, \nu \rangle}{\|\nu\|^2}, \tag{35}$$

with the interval unbounded when $M = 0$. If $\nu = 0$, the objective is identically zero and choose $\alpha^* = 0$. Any feasible scaled dual gives a lower bound regardless of restricted-master optimality.

For finite termination, require an attained primal–dual optimum of the restricted master. If exact pricing gives $M_X(\nu) \leq \beta$, its dual is globally feasible, so weak duality sandwiches the full optimum between equal restricted endpoints. If $M_X(\nu) > \beta$, a maximizer exists and belongs to an omitted violated pattern block; a currently present block cannot violate a constraint of the exactly solved restricted dual. Adding this complete new block strictly increases the finite pattern set. Re-solving and repeating terminates after at most $|\mathcal{D}_X \setminus \mathcal{P}_0|$ additions. This is an exact-oracle finiteness statement, not a polynomial complexity bound. In computations, outward feasibility tests and a nonzero target tolerance replace exact equality.

For rank-two inputs, every homogeneous chamber is a sector. A linear functional over its intersection with the unit disk is maximized either at the normalized functional direction, if that direction lies in the sector, or at a boundary ray; zero handles a nonpositive maximum. Exact integer ray ordering and rational signs enumerate the chambers. Interval square roots enclose each support value, providing the pricing upper bound used by the independent verifier.

E RETAINED-WEIGHT REPAIR OF DEEP SURROGATE RELAXATIONS

A relaxed source can provide a lower bound even when its intermediate states do not describe a realizable network. Retained-weight repair constructs the upper witness. Bounding the change in output and penalty then connects the relaxed certificate to that witness's target quality.

E.1 DIMENSIONS AND EXACT IDENTITY

Let the network have widths $d_0, \dots, d_K$, input $H_0 = X \in \mathbb{R}^{n \times d_0}$, weights $W_k \in \mathbb{R}^{d_{k-1} \times d_k}$, and elementwise activations σ_k that are ReLU or the identity. In a sound relaxation, preactivations Z_k and states H_k are independent variables, each $n \times d_k$. Retain its weights and recompute

$$\widehat{H}_0 = X, \qquad \widehat{H}_k = \sigma_k(\widehat{H}_{k-1} W_k).$$

Define $A_k = H_k - \sigma_k(Z_k)$, $B_k = Z_k - H_{k-1} W_k$, and $E_k = H_k - \widehat{H}_k$. For each entry, the scalar secant slope between Z_k and $\widehat{H}_{k-1} W_k$ lies in $[0, 1]$; at equal arguments choose any admissible slope. Let $\mathcal{D}_k$ apply those entrywise slopes. Then

$$E_k = A_k + \mathcal{D}_k B_k + \mathcal{D}_k(E_{k-1} W_k). \tag{36}$$

Add and subtract $\sigma_k(Z_k)$, apply the secant identity, and substitute $Z_k - \widehat{H}_{k-1} W_k = B_k + E_{k-1} W_k$. This proves (36) exactly. The operator is a diagonal map after vectorization and acts separately on samples; it is not multiplication by one sample-independent activation matrix.

Write $\mathcal{A}_k(E) = \mathcal{D}_k(E W_k)$ and $\mathcal{P}_{K:k+1} = \mathcal{A}_K \circ \cdots \circ \mathcal{A}_{k+1}$, with the empty composition the identity. Since $E_0 = 0$,

$$E_K = \sum_{k=1}^{K} \mathcal{P}_{K:k+1}(A_k + \mathcal{D}_k B_k). \tag{37}$$

Any outward-valid induced operator bounds give

$$\|E_K\|_F \le \sum_k \left(G_k^{\mathrm{act}} \|A_k\|_F + G_k^{\mathrm{bil}} \|B_k\|_F \right) =: \Delta,$$

where $G_k^{\mathrm{act}} \ge \|\mathcal{P}_{K:k+1}\|$ and $G_k^{\mathrm{bil}} \ge \|\mathcal{P}_{K:k+1} \mathcal{D}_k\|$. A simple choice bounds both by $\prod_{j>k} \|W_j\|_2$. If the initial states differ, include $\|\mathcal{P}_{K:1} E_0\|_F$; this contribution is required in the repair bound.

These secants belong to the particular relaxed/repair pair. Uniform bounds over a branch domain replace them by verified stable gates or the safe norm bound one. A checkpoint's activation mask does not describe every competing network.

E.2 GLOBAL SOURCE LOWER EVIDENCE AND REPAIR COST

The layerwise identity controls output displacement. To convert that displacement into an objective guarantee, the relaxed lower evidence must cover the full neural class and any penalty increase must also be included.

A closed convex lifted feasible set must contain every exact network trajectory in the stated weight domain. With unregularized squared loss, its output projection gives a closed convex prediction set whenever the lifted set is compact. For a convex weight penalty $R(W)$, minimize the penalty over lifted points with a fixed output to obtain a convex prediction cost ϕ_{out}; compactness of the lifted set and continuity of the penalty give closedness. More generally, a solver may furnish a verified lower/support certificate directly without relying on this projection argument.

Let a bound $\sup\{\langle \xi, h \rangle - r : (h, r) \text{ is relaxed-feasible}\}$ globally, and use $D_Y^{\mathrm{out}} = \langle Y, \xi \rangle - \|\xi\|^2/2 - a$. This support bound is shared by all targets in the fixed relaxed family. Let a relaxed source point have output h and cost r_h, and its relaxed dual have value D_C^{out}. Its source gap is $g_C = \|h - C\|^2/2 + r_h - D_C^{\mathrm{out}}$. Suppose repair gives output $z = h + e$ with $\|e\| \le \Delta$, and actual neural cost $R(\theta_{\mathrm{repair}}) \le r_h + \chi$. Transporting the relaxed pair and then comparing squared losses gives

$$F_Y(\theta_{\mathrm{repair}}) - D_Y^{\mathrm{out}} \le g_C + \tfrac{1}{2}\|Y - C\|^2 + \langle C - h - \xi, Y - C \rangle - \langle Y - h, e \rangle + \tfrac{1}{2}\|e\|^2 + \chi \tag{38}$$

$$\le \tfrac{1}{2}(\|Y - C\| + \sqrt{2 g_C})^2 + \|Y - h\| \Delta + \tfrac{1}{2}\Delta^2 + \chi. \tag{39}$$

The first inequality is equality when the recorded and repaired costs agree exactly. Retained weights preserve any weight-only regularizer, so $\chi = 0$ for that case. If a relaxation uses a smaller cost proxy, its difference from the repaired penalty must be charged.

E.3 Source defects, global validity, and convergence

A closing relaxation certificate needs controllable bounds on its activation and product defects. Interval refinement supplies those bounds and identifies the extra uniformity conditions required for convergence.

For a relaxed product v in the McCormick hull of wh on a rectangle,

$$|v - wh| \le \tfrac{1}{4}(\overline{w} - \underline{w})(\overline{h} - \underline{h}).$$

For an unstable ReLU interval $\ell < 0 < u$, the maximum vertical graph defect in its exact convex hull is

$$\frac{-\ell u}{u - \ell} \le \frac{u - \ell}{4}.$$

The product extremum is attained at the normalized rectangle center; the ReLU extremum is at preactivation zero. Applying these entrywise and then (37) yields a verified output radius. Splitting bounded factor intervals controls these defects, provided the refinement actually contracts every needed factor interval. Stable activation intervals have zero ReLU hull defect.

A lower bound for one branch is valid for that branch only. The minimum of verified bounds over an exhaustive branch partition supplies the class-wide lower endpoint. Target-dependent pruning, cuts, and incumbent-based domains must be revalidated when C changes. Feasible repaired networks may always be evaluated on the target if they satisfy its fixed parameter constraints.

For uniform closing on a compact target set, it suffices that source distance, verified relaxed source gaps, repair radii, penalty repair charges, and outward numerical allowances vanish uniformly, while $\|Y - h\|$ remains uniformly bounded. A common compact lifted domain ensures this last bound and finite propagation constants. These are additional assumptions for uniform relaxation closure; none is needed for the basic metric theorem.

F Basis banks and hybrid Neural MERIT

A stored surrogate can strengthen the current target certificate through either selection or compatible composition. The guarantees below specify what information must be retained, when records can cooperate, and how interrupted computations preserve the existing bound.

F.1 Individual sources and compatible mixing

A source record contains its exact problem identity, a feasible neural representation, a verified lower proof, and any gap or reconstruction allowance. Scalar records use Theorem 2.1. Support records use Theorem 3.1. For a finite bank, maxima of valid lower bounds and minima of feasible upper values remain valid. Adding candidates can only strengthen these envelopes.

In (10), $\operatorname{dom}\phi^*$ is convex. The objective D_Y is concave and upper semicontinuous on the finite hull of stored feasible duals; the hull is compact, and the objective is finite there by convexity of ϕ^*. Hence the maximum in (17) exists. The union upper master is assumed to have an attained optimum in a realizable class containing every stored representation at its recorded cost. In the finite ReLU master with $\beta > 0$, bounded objective sublevel sets in the grouped coefficients give this attainment; reconstruction is established in Appendix D.

Each individual pair is available in both feasible sets. Thus $L_m^{\mathrm{mix}} \ge D_Y(\xi_i)$ and $U_m^{\mathrm{union}} \le \|z_i - Y\|^2/2 + r_i$. Subtraction followed by (30) proves (18). If a target-optimal dual and a target-optimal representation are present, weak duality forces both endpoints to equal $p^\star(Y)$.

F.2 A quantitative cooperative gain

The best individual record uses only one source pair. Combining realizable predictions and globally feasible duals can remove additional uncertainty; the variance terms below measure that improvement.

Assume that *for every* $\alpha \in \Delta_m$, the union admits a feasible neural representation of $\bar{z} = \sum_i \alpha_i z_i$ with regularization cost at most $\sum_i \alpha_i r_i$. Let $\bar{\xi} = \sum_i \alpha_i \xi_i$ and $g_i(Y) = \|z_i - Y\|^2/2 + r_i - D_Y(\xi_i)$. Define $\mathrm{Var}_\alpha(v) = \sum_i \alpha_i \|v_i - \bar{v}\|^2$. The variance identity gives

$$\tfrac{1}{2}\|\bar{z} - Y\|^2 + \sum_i \alpha_i r_i = \sum_i \alpha_i \big(\tfrac{1}{2}\|z_i - Y\|^2 + r_i\big) - \tfrac{1}{2}\,\mathrm{Var}_\alpha(z).$$

Convexity of ϕ^* and the quadratic dual term give

$$D_Y(\bar{\xi}) \geq \sum_i \alpha_i D_Y(\xi_i) + \tfrac{1}{2}\,\mathrm{Var}_\alpha(\xi).$$

The two inequalities hold for every $\alpha \in \Delta_m$. Since the bank endpoints optimize over sets containing each such feasible construction,

$$U_m^{\mathrm{union}} - L_m^{\mathrm{mix}} \leq \min_{\alpha \in \Delta_m} \left\{ \sum_i \alpha_i g_i(Y) - \tfrac{1}{2}\,\mathrm{Var}_\alpha(z) - \tfrac{1}{2}\,\mathrm{Var}_\alpha(\xi) \right\}. \tag{40}$$

Taking a simplex vertex recovers individual transport. Both variance terms can improve it. If the synthesis assumption is known only for a nonempty set $\mathcal{A} \subseteq \Delta_m$, the same proof gives (40) with $\inf_{\alpha \in \mathcal{A}}$ instead of $\min_{\alpha \in \Delta_m}$. No closedness of $\mathcal{A}$ is needed for this infimum bound. The full-simplex minimum exists because its displayed objective is continuous. Neural realizability and the cost condition are indispensable; averaging arbitrary network parameters does not establish them.

For an exact illustration, take $\phi(z) = |z|$ on $\mathbb{R}$ and sources $C_1 = 2$, $C_2 = -2$, with pairs $(z_1, \xi_1) = (1, 1)$ and $(z_2, \xi_2) = (-1, -1)$. At $Y = 0$, either transported pair has upper $3/2$, lower $-1/2$, and width two. Equal mixing gives $\bar{z} = \bar{\xi} = 0$, hence exact target value zero. Retaining the nonnegative lower floor would improve either individual interval to width $3/2$, but neither individual primal witness is target-optimal. This example isolates algebraic cooperation; the ReLU experiments test realizable pattern-union composition.

F.3 Bank-free fallback, interrupted solves, and scope

A bank computation may stop before its auxiliary optimizations finish. Retaining already verified source endpoints ensures that this additional computation cannot weaken the certificate available for the target.

For a finite bank and any partial numerical mixture/union computation, retain every independently evaluated stored pair. The candidate mixture provides a lower improvement only after feasibility verification. The union provides an upper improvement only after actual neural evaluation. Taking the best of these candidates preserves the per-record transport bound even before the auxiliary optimizations converge. It does not entitle a partial solve to the full variance-gain bound unless a corresponding feasible mixture and synthesis have been evaluated.

In (19), each class endpoint is already a historical extremum. Adding the corresponding bank extrema makes the reported interval monotone in both source event k and nested bank size m. A fixed upper witness travels with its objective. If the current trainer has value Q, its separate certificate is $Q - p^\star(Y) \leq Q - L_{k,m}$; $U_{k,m} - L_{k,m}$ certifies the selected best witness.

For compatible primal–dual records of (10) with *zero recorded source gap* $g_i = 0$, including any synthesis cost, let $\mathcal{Y}$ be a nonempty totally bounded target domain and define $h_m = \sup_{Y \in \mathcal{Y}} \min_i \|Y - C_i\|$ for $m \geq 1$. Theorem 5.1 gives a uniform bank-width bound $h_m^2/2$. If $h_m \leq K m^{-1/q}$ for some $K < \infty$ and $q > 0$, the sufficient width rate is $K^2 m^{-2/q}/2$. Exact scalar source optima alone do not imply this quadratic bound: their metric root-width is at most $2h_m$, and positive source residual can leave a first-order objective-width envelope. The identity-design ReLU calibration below tests the quadratic law; its finite nonconvex counterpart tests metric transfer. Both describe certificate quality, independently of source-solve cost.

G Experimental protocols and additional evidence

The studies address three questions: whether target certificates follow the predicted distance law, whether exact nonconvex source solutions improve on the target, and how much a basis bank adds to the same guarantee. The protocols separate these questions from source-solver accuracy and timing.

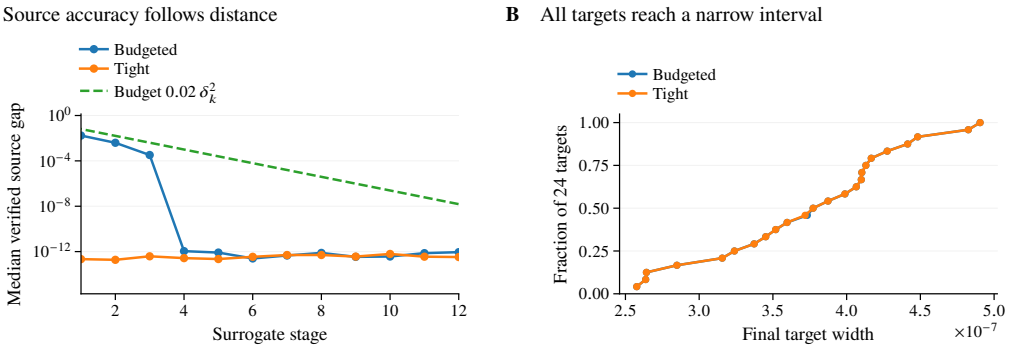

Figure 5: **Intermediate source accuracy and final target quality.** Left: medians over the 24 paths; the budgeted policy can accept larger early source gaps, then carries discovered structure forward. The dashed curve is its assigned source tolerance. Right: all 24 final target widths are shown for each policy. The nearly coincident curves show that permitting controlled intermediate error preserves a comparably tight final target certificate.

G.1 Continuation on three regularized ReLU families

To assess quadratic transport under both tight and tolerance-controlled source solves, each run follows one fixed target through the same geometric sequence. Independent families test whether the behavior persists across input and target realizations.

The input seeds are $11, 29, 47$. For each seed, 24 approximately circular two-dimensional inputs have perturbed angles and radii, and a four-neuron ReLU teacher supplies a noisy base target. Inputs and base targets are rounded to multiples of $1/256$. A Gaussian 24×3 matrix, seeded by the family seed plus 500, is orthogonalized and its columns are then rounded to multiples of $1/256$ to obtain V. The rounded V need not be exactly orthonormal; all distances use its actual values. The targets are

$$Y(c) = Y_0 + 1.5Vc, \qquad c \in [-1, 1]^3. \tag{41}$$

The fixed input matrices each have 48 empirical activation chambers. The optimization problem is (14) with $\beta = 1/8$, Euclidean squared loss, and width 96. The complete pattern master and its reconstructed networks belong to this same class.

Each family uses eight new scrambled Sobol query coordinates, with seed equal to the family seed plus 9,000, rounded to multiples of $1/1024$ (Sobol', 1967; Owen, 1995). None equals a retained bank center. The bank centers are 64 coordinates generated with the family seed plus 100 and rounded to multiples of $1/256$. Both source and query arrays are released. There are 24 distinct continuation targets, not 24 times the number of source stages independent tasks. The continuation sources $C_k = (1 - 2^{-k})Y$, $k = 0, \ldots, 12$, are represented exactly in the stored dyadic arithmetic. The zero source uses an analytic zero network and zero dual.

The tight-source policy solves the complete 48-pattern problem at each nonzero source and accepts a record only when its verified source gap is at most 5×10^{-8}. The actual maximum is 4.2181×10^{-9}. The dual finite cone formulation is optimized with SLSQP. Nonnegative masses on active pricing directions reconstruct a primal network; an additional SLSQP coefficient solve refines that same nonnegative quadratic problem when the initial coefficient solver is insufficient. Feasibility and the accepted gap are checked independently of solver status.

The distance-budgeted policy starts with an empty pattern set and carries its accepted set to the next source. At C_k, it solves the restricted master, globally prices its dual, and adds a new violated pattern when necessary. It stops as soon as the verified source gap is at most $0.02\delta_k^2$, $\delta_k = \|Y - C_k\|$. If a numerical restricted master supplies no new violated pattern while missing the requested gap, the implementation permits a full-master fallback; no such fallback was needed in the 24 reported budgeted paths. The actual maximum ratio of accepted gap to assigned tolerance is 0.99376.

Figure 5 shows source tolerances and the complete final-width distribution. The source policies produce 624 records in total, including the analytic zero sources. A separately certified target optimum

is computed for each path solely for validation, with largest reference gap 3.2254×10^{-10}. Source acceptance uses source evidence and distance, never the target reference. Figure 2 uses seed 11, query zero as its representative path; this is the first query in the declared ordering. The second and third panels include all 24 paths.

For each accepted source, the archive contains C_k, target Y, grouped primal vectors, a verified globally feasible dual, source and target endpoints, residual diagnostics, pattern set, master calls, and elapsed time. The root arithmetic and chamber checks are described in Appendix H. The transport identity diagnostic uses rounded displayed prediction vectors and can have floating-point residual; the reported lower and upper endpoints do not depend on that residual test for validity.

G.2 The bank comparison on the same continuation targets

Holding the target and source schedule fixed isolates the bank's contribution from changes in continuation difficulty. Only the amount of previously solved evidence varies in this comparison.

At each bank size $m \in \{0, 4, 16, 64\}$, stored duals form the finite simplex mixture in (17). SLSQP proposes a mixture. Clipping and normalization create a candidate combination, and complete pricing verifies its actual floating-point dual vector; any necessary inward scaling precedes its lower evaluation. The union master contains every active source pattern whose grouped coefficient mass exceeds 10^{-7}, while independently evaluated original source witnesses remain available as fallbacks even if a tiny pattern was omitted. The upper endpoint is the smaller of the verified union reconstruction and those stored witnesses.

Stage zero in the bank study includes these target-specific mixture and union-master computations. It excludes continuation but does not mean zero query computation. The source paths themselves are unchanged between bank sizes, and the hybrid intervals simply retain the stronger endpoints. Consequently Table 1 measures earlier availability of a certificate, not an equivalent-budget or time-to-solution comparison. All 96 bank/query pairs, all source pairs, and every independent reference were replayed without an optimizer.

The new continuation run uses one CPU thread. The total recorded time for source solving and arithmetic verification across the 624 source stages is approximately 162 seconds in the provided Linux/Python environment. Bank evaluation and target-reference construction have separate recorded times. These small-instance timings are reproducibility information; no runtime advantage over a direct target solve is claimed.

G.3 Exact all-layer finite-network continuation

Exhaustive integer-arithmetic source solves isolate the nonconvex continuation theorem from numerical solver error. The parameter class allows all layers to change, so discontinuous source-optimum changes remain visible.

The network has widths 2–2–2–1, ReLU after the two hidden layers, and a linear output. Ten weights take values in $\{-1, 0, 1\}$; all are trainable. The eight input rows are one half of

$$(-1,-1),\ (-1,0),\ (-1,1),\ (0,-1),\ (0,1),\ (1,-1),\ (1,0),\ (1,1).$$

The objective is $\|z - Y\|_2^2/16 + \|\theta\|^2/32$. Enumerating every parameter vector gives 59,049 states. Positive homogeneity and zero activations produce repeated predictions; each of 705 distinct predictions is retained with the minimum penalty among its realizations. This reduction preserves the global objective for *all* targets because a larger penalty with the same prediction is dominated.

A NumPy generator with seed 20260912 draws 24 integer target vectors in $[-16, 16]^8$, which are divided by eight. Source coordinates are the union of $j/128$ for $j = 0, \ldots, 128$ and $1 - 2^{-k}$ for $k = 0, \ldots, 12$, giving 134 distinct rational coordinates. All source objectives admit a common integer denominator. The verifier checks an explicit magnitude bound below 2^{63} before using vectorized 64-bit integer scores. It re-enumerates all parameters and checks every selected minimum and source witness independently.

The metric lower endpoint uses an outward interval square root. For comparison, the source-residual support certificate uses $\xi_k = C_k - z_k$ and exactly evaluates

$$\mathsf{H}(\xi_k) = \max_{\theta} \{\langle \xi_k, z(\theta) \rangle_n - R(\theta)\}, \qquad \langle u, v \rangle_n = u^\top v/8.$$

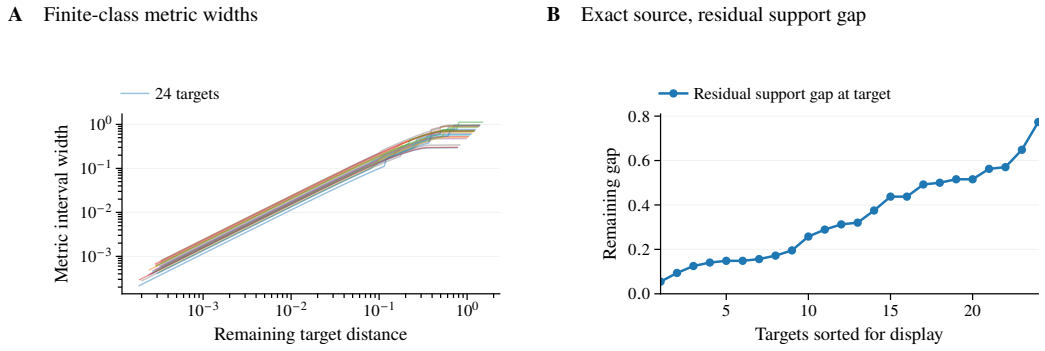

Figure 6: **Exact source values remain informative when directional evidence is incomplete.** Left: all 24 finite-class metric certificates contract with remaining target distance. The target endpoint itself, whose metric width is zero up to outward rounding, is omitted from the logarithmic plot. Right: source-residual support certificates retain positive gaps at the exact target solutions. The comparison uses the same finite network class and explicitly evaluates every global competitor.

It is not optimized over all possible support directions. Its positive terminal gap demonstrates that a particular exact source solution does not automatically close a support certificate. The analytic $\{-1, 1\}$ example in Appendix B separately establishes a genuine best-support-dual floor.

All 3,216 exact source problems pass target-upper monotonicity, metric-lower monotonicity, and interval validation. The median number of distinct source-winning networks is three, with a range from one to seven. Source minimizers can therefore change while their target performance remains monotone, as predicted by Theorem 2.2. Figure 6 shows the complete width and residual-gap distributions; the first target supplies the main-paper illustration.

G.4 Frozen-image-feature finite head and a separate local trainer

A trainer can plateau near an optimum without knowing how near it is. The finite-head experiment places the trainer's actual objective, its independent reference, and the available certificates on the same target, then measures what prior surrogates add.

The image experiment retains 1,024 balanced cat/dog images from CIFAR-10 and a fixed ImageNet-pretrained ResNet-18 feature extractor (Krizhevsky, 2009; He et al., 2016; Deng et al., 2009). The 512-dimensional feature vectors are centered and reduced to four principal components once. The released input matrix, image indices, labels, target directions, and preprocessing outputs define the numerical problem that is certified. No parameter of the feature extractor is part of the trainable class.

The finite head has at most three atoms. Each atom is $a[Xw/\|w\|]_+$ for $w \in \{-1, 0, 1\}^4 \setminus \{0\}$ and $a \in \{-1, -1/2, 1/2, 1\}$, with one additional zero option. Thus there are 321 listed atoms. The exhaustive engine uses 51,681 unordered atom pairs and a third atom, for 16,589,601 scored configurations. This includes 5,564,321 canonical multisets and redundant orderings; it is not a count of distinct prediction functions. The objective is half the mean squared residual, with no penalty on this finite head.

Targets have the form $Y(c) = Y_0 + \rho V c$ for three stored orthonormal data-derived target directions and $c \in [-1, 1]^3$. The selected query is $c = 0$, excluded from the nested source bank. A separate set of 128 target perturbations is used for held-out reporting. Every center and held-out target has an exhaustive numerical reference, with the reported matching tolerance 2×10^{-9}. These floating-point exhaustive references are distinct from the exact integer-arithmetic references in the preceding experiment.

The local trainer is a full-batch coordinate method with 20 fixed restarts. Each of 2,500 updates revisits one restart and one of its three slots, evaluates all 321 replacement atoms, and accepts only a strict decrease of the full objective. The incumbent across all restarts is plotted without smoothing. Thus the horizontal unit is one coordinate update, not a complete sweep over every restart and slot.

Table 2: **Additional solved surrogates strengthen immediate finite-head certificates.** The 128 held-out targets use half mean squared loss. A reference-matching witness is counted only by independent exhaustive evaluation at tolerance 2×10^{-9}; its online interval may remain wider.

Stored records	Median width	Maximum width	Reference-matching witnesses
0	0.530603	0.575095	0/128
10	0.123211	0.210755	28/128
100	0.045191	0.110517	107/128
1,000	0.029669	0.049190	126/128
10,000	0.021113	0.024567	128/128

Table 3: **Cooperation improves certificates across independently instantiated ReLU families.** Each row has 64 held-out queries and 64 stored sources. The loss is the Euclidean objective (14), so values are not pooled with the normalized finite-head experiment.

Family	Stored median	Cooperative median	Cooperative maximum	Width ≤ 0.01
11	0.0999	0.0057	0.1013	42/64
29	0.0706	0.0037	0.0719	48/64
47	0.0826	0.0042	0.1333	47/64
All	0.0828	0.0049	0.1333	137/192

The selected reference is 0.4486445059; the trainer reaches approximately 0.4486816 at update 29 and makes no further incumbent improvement through update 2,500.

For the selected trajectory, the convex outer prediction class $3\,\mathrm{conv}(\mathcal{A})$ supplies a bank-free dual lower trajectory, and the actual trainer supplies a feasible upper trajectory. Every stored exact source value supplies a metric exclusion certificate; every source head supplies a target-evaluated upper witness. Historical extrema and (19) form the plotted intervals. The selected query's bank width is 0.022624 at both $m = 1,000$ and $m = 10,000$: the larger bank is nonworsening but does not strictly improve this query.

The held-out *immediate* comparison starts with the zero network and lower bound zero at $m = 0$. Its initial state is therefore different from the selected trainer's already initialized class-level lower bound. Banks are nested, and the reported held-out intervals retain this baseline as well as all stored-source improvements. Table 2 reports the numerical outcomes. The source campaign recorded approximately 0.58 active A100 hours, including reused feature preparation; this is an offline provenance measurement, not a comparison of amortized solve costs.

G.5 Compatible composition and Pattern Completion

Certificate improvement from combining records should be distinguished from simply choosing a better record. This study compares those operations on identical bank prefixes; a separate pilot measures how stored patterns affect the proof work still required.

A separate 192-query experiment uses the same three input/teacher families, each with 64 bank centers and 64 held-out queries generated with a different Sobol seed. These queries are distinct from the 24 new continuation targets. All 384 source/reference pairs and 960 retained query-bank certificates have saved neural and dual witnesses. The arithmetic verifier uses the same global rank-two pricing procedure as the continuation study.

The nearest-pair baseline selects a source by Euclidean target distance. The stored-endpoint baseline independently selects the best stored upper and lower values. Dual mixing improves the lower candidate; union-pattern reoptimization improves the upper candidate. All methods retain the same finite nested bank prefixes. At $m = 64$, aggregate median widths are 0.083096 for nearest pairs, 0.082751 for stored endpoints, and 0.004914 for compatible composition. The last is 16.84 times smaller than the stored-endpoint median. Table 3 gives per-family results; Figure 7 shows the cardinality effect and the complete final-width distribution.

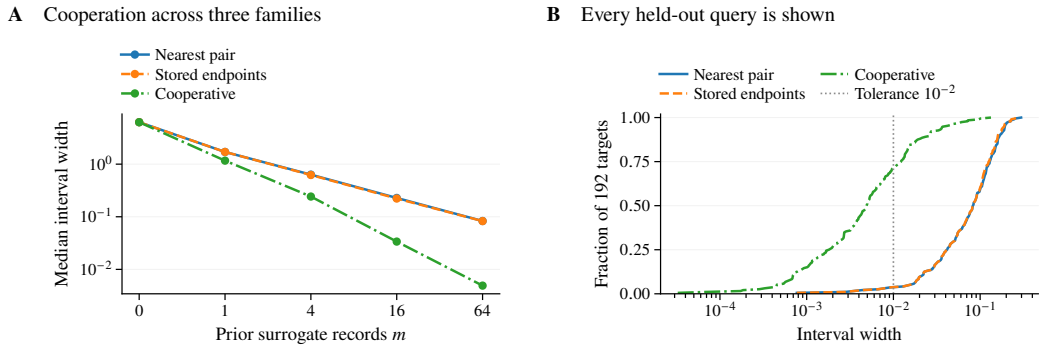

Figure 7: **Compatible records improve the same certificate beyond source selection.** Left: median widths over 192 ReLU queries using identical bank prefixes. Right: the full empirical width distributions at 64 records retain the difficult tail rather than only showing the median. The upper side is always a feasible neural reconstruction; the lower side is independently globally priced.

This conservative verifier scans every chamber and evaluates multiple candidates exactly. On these small problems, its cost need not be lower than directly solving the target. The experiment establishes the additional *certificate strength* supplied by cooperation. Solver and verification timings are separately recorded so that a stronger interval is not misreported as a speedup.

The Pattern Completion pilot fixes an 18×2 input matrix, $\beta = 0.12$, and 36 complete patterns. Independent scrambled Sobol sequences generate 100 nested sources and 16 held-out targets. Their largest independent reference interval is 3.17×10^{-9}. At bank sizes 10 and 100, the union contains 13 and 15 patterns; median immediate widths are 6.21×10^{-4} and 1.83×10^{-5}. Four predeclared cold queries require $6, 6, 9, 11$ new complete pattern blocks. Both nonempty banks require zero new blocks. This is a count of additions, not an absence of all pricing or feasibility checks. Median recorded closure times are 0.78 seconds cold, 0.22 at 10 records, and 0.24 at 100 records. The final residual gaps are below 3.4×10^{-9}.

G.6 Large-cardinality calibration within exact source classes

Two controlled families test the dependence on the number of globally solved surrogates through $m = 10^5$. The finite nonlinear family has six two-dimensional inputs and a width-three ReLU network with hidden vectors in $\{-1, 0, 1\}^2 \setminus \{0\}$ and signs ± 1. It has 4,096 ordered parameter settings and 664 distinct predictions. Every source is solved by exhaustive nearest-prediction evaluation. The no-bank lower bound is distance to the output bounding box; its upper is a fixed feasible prediction. The bank uses metric transfer on a two-dimensional affine target family, with 1,024 disjoint evaluation queries.

For the exact ReLU calibration, take $X = I_8$ and $\beta = 1/2$. The atom gauge is

$$\gamma_I(z) = \|z_+\|_2 + \|z_-\|_2.$$

The decomposition into positive and negative parts supplies an upper representation. In any positive/negative decomposition $z = p - n$, $p, n \geq 0$, the inequalities $p \geq z_+$, $n \geq z_-$ show that its masses are no smaller, proving equality. The source solution and dual are

$$z_C = \left(1 - \frac{\beta}{\|C_+\|}\right)_+ C_+ - \left(1 - \frac{\beta}{\|C_-\|}\right)_+ C_-, \qquad \xi_C = C - z_C,$$

with a zero term when its denominator is zero. Consequently all 10^5 source pairs are globally known without a numerical nonconvex search. This exact class tests the quadratic transport law at large bank cardinality. It is not evidence that arbitrary 10^5 neural training problems are comparably cheap.

Figure 8 reports medians, 90th percentiles, and maxima. At 10^5 sources, the finite-class median width is approximately 1.41×10^{-3} and the identity-design ReLU median is 3.13×10^{-6}. All 12,288 query/snapshot records across the two families and six bank sizes are supplied. These cardinality

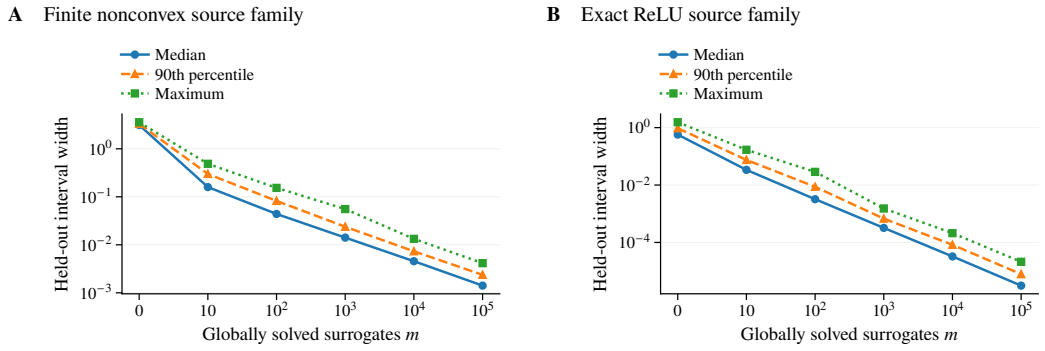

Figure 8: **More globally solved surrogates continue to tighten the target guarantee.** Both calibrations use 1,024 held-out targets and nested banks through 10^5 records. Left: scalar metric evidence in a finite nonconvex ReLU class. Right: exact primal–dual evidence in the identity-design ReLU class. The two families have distinct objectives and units; their separate panels show the applicable transfer laws, not a pooled neural scaling exponent.

experiments use previously solved sources for independent target queries; they do not alter the continuation claims.

H VERIFICATION, PROOF SCOPE, AND REPRODUCIBILITY

The reported interval is useful only when its lower endpoint is globally valid and its upper endpoint belongs to a feasible network. The following checks address these obligations separately and allow numerical results to be replayed without trusting the source optimizer.

H.1 PROOF OBLIGATIONS AND ARITHMETIC CONTRACTS

Every main theorem has a complete argument above. The metric theorem uses nonempty-set distance geometry. Exact-source monotonicity uses two global source inequalities on a common domain. Quadratic transport uses a global support inequality and an actual feasible witness. The continuous quadratic path additionally uses a proper closed convex prediction cost. Basis-bank improvement requires compatible dual domains and realizable primal synthesis. Deep repair adds an explicit output and penalty discrepancy. These assumptions are not exchanged between the results.

The arithmetic verifier interprets every saved binary64 input, weight vector, and dual vector as its exact binary rational value. Dot products and chamber-membership signs use Python `Fraction`. Norms and pricing ratios involving square roots use 80-decimal `mpmath.iv` enclosures. Lower objective values are rounded down and upper values up. A candidate dual is inward-scaled only when necessary and then checked again on the complete chamber family. The saved dual itself is the verified vector, so later target evaluation reuses a globally valid support certificate.

Upper values are computed from the actual ReLU forward pass of the stored grouped vectors. The balanced neural realization can contain square roots; its exact prediction and cost are evaluated through the algebraically equivalent grouped form. The declared width bounds this actual representation. Approximate solver stationarity or successful exit status never substitutes for global pricing or endpoint verification.

H.2 INDEPENDENT REPLAY AND SYMBOLIC CHECKS

Saved witnesses make it possible to check a result independently of the algorithm that proposed it. Algebraic checks test the identities used in transport; arithmetic replay tests the concrete source and target certificates.

`code/verify_continuation.py` calls no optimizer. It rechecks 624 source pairs, 24 target references, and 96 bank/query pairs. It also independently regenerates all 59,049 finite-network assignments and verifies the 3,216 exact source optima, target intervals, and monotonicity relations. Every check passes. The saved coefficients and duals are sufficient for this replay; the source optimizer need not reproduce an identical floating-point solution.

Wolfram Language checks the exact support/transport identities, residual decomposition, sharp radial envelope, continuation cancellation, quantified nonconvex monotonicity implication, variance identities, scalar dual scaling, a finite-class primal–dual counterexample, a finite-depth repair telescope, and quantified McCormick/ReLU hull bounds. Successful inputs and returned outputs are retained. A local SymPy check independently expands the central identities, and 1,912 randomized or boundary-case instances test nonconvex transport and paths, convex proximal paths, and compatible synthesis. These computational checks complement, rather than replace, proofs of the quantified functional statements.

The inherited experiments have additional standalone replay scripts. The frozen-feature finite-head results use exhaustive floating-point evaluation with stated tolerances; the rank-two ReLU certificates use rational/interval verification; the all-layer finite continuation uses exact integer global minimization. The report preserves these different levels of numerical evidence rather than treating all solver outputs as exact arithmetic.

H.3 Artifact map and timing boundaries

Reproducing a certificate requires its problem data and proof witnesses, while interpreting its computational cost requires the timing boundary. The archive supplies both.

The package includes LaTeX sources, numeric records, saved neural and dual witnesses, and independent verifiers. Its `README.md` lists every reproduction command. `build.sh` compiles the paper; `code/make_figures.py` rebuilds all numerical panels with external LaTeX text. The two `run_continuation.py` and `run_finite_continuation.py` scripts reproduce the new studies; the former resumes from atomically saved source records. Timing fields separate source generation, bank evaluation, and verification. Stage counts measure certificate availability after the stated bank computations.