POINTWISE WASSERSTEIN VERIFICATION OF NEURAL NETWORKS VIA CLOSED-FORM BOUND CONCRETIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Affine-bound verifiers such as CROWN, DeepPoly, and DeepZ certify neural networks against ℓ_p perturbations by producing a valid affine bound $A \cdot x + b$ and minimizing it over the ball through the dual norm $\|A\|_q$. We show that this last step—the concretization—admits a closed form for Wasserstein balls: the worst case of any valid affine bound over a W_1 ball is the budget ε times the *maximal edge drop* $\|B^\top A\|_\infty$, the W_1-Lipschitz constant of the affine envelope, computed in $O(|E|) = O(n)$ tensor operations on the pixel graph under the L1 ground metric, with no transport-side optimization. The form is exact on the relaxed (signed-measure) ball and is therefore a sound certificate on its own; an optional LP tier, gated by a zero-cost test that never changes a certification decision in our benchmarks, restores exactness on the probability-measure ball. Because the closed form depends only on the envelope, not on the domain it was computed on, every ℓ_p verifier is a ready-made envelope generator; and because the interface consumes validity rather than provenance, even learned envelopes are admitted by a check built from the same closed form, which certifies or abstains but never issues a false certificate. The method yields per-sample certified radii with an exact two-term gap decomposition that localizes all conservatism to the envelope, and the closed form matches explicit LPs to solver tolerance at roughly three orders of magnitude lower concretization cost.

1 Introduction

Neural network verification asks for a mathematical guarantee that a property holds over a whole input region. For a classifier f and correctly classified input x_0 with label y_0, the standard region is an ℓ_p ball and the property is $f_{y_0}(x) > f_j(x)$ for all $j \neq y_0$. Scalable verifiers—reverse-mode bound propagation (IBP, CROWN, DeepPoly, α, β-CROWN; Gowal et al., 2019; Zhang et al., 2018; Singh et al., 2019b; Wang et al., 2021), symbolic-interval methods (Neurify, VeriNet Wang et al., 2018; Henriksen & Lomuscio, 2020), zonotope-domain verifiers (AI^2, DeepZ, RefineZono Gehr et al., 2018; Singh et al., 2018; 2019a)—all share the same final step: given a valid affine bound $A \cdot x + b$ on the target quantity, find its worst case over the perturbation ball. We call this step the *concretization*; for an ℓ_p ball it has a closed form, the dual norm $\|A\|_q$ (Section 2.1).

An ℓ_p ball is blind to geometric change. An ℓ_p budget prices a rigid one-pixel translation at the *maximum* possible norm, so no small ℓ_p ball can express "small translations only." The same blindness appears in text, where no ℓ_p or substitution-set budget prices good $\to$ great cheaper than good $\to$ bad. The fix is a *metric ball*: the 1-Wasserstein distance prices translations, deformations, and thickness changes as cheap, and covers every mass-preserving perturbation. By Kantorovich–Rubinstein duality, certifying a W_1 ball *is* computing the sharp W_1-Lipschitz constant of the quantity under consideration. This duality is classical in optimal transport (Peyré & Cuturi, 2019), and W_1-Lipschitz estimation of networks is well-studied in the ℓ_p literature (Fazlyab et al., 2019; Shi et al., 2022; Chiu & Zhang, 2025). What is new here is not the duality but its *use*: we identify this constant, on the pixel graph, with the very concretization step that every affine-bound verifier must perform, and thereby obtain a closed-form, domain-independent, $O(n)$ concretization (Section 2).

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

This paper: a closed-form Wasserstein concretization at the affine-bound interface. We give that concretization in closed form for the W_1 ball (Section 2); it is a universal drop-in—it requires only that the bound be valid, not how it was produced—and is exact no matter on which enclosing domain the bound was computed (Proposition 2.2). Prior work falls short of this target: randomized smoothing (Levine & Feizi, 2020) certifies a smoothed classifier probabilistically, and Wegel et al. (2021) give a deterministic certificate via flow-space lifting (Section 3.3) that requires an affine first layer and a per-sample reference flow—restrictions our coordinate-free form removes.

Contributions.

1. **Closed-form Wasserstein concretization** (Section 2): the worst case of any valid affine bound over a W_1 ball is $\varepsilon\|B^\top A\|_\infty$ under the L1 ground metric, computed in $O(n)$ tensor ops with no transport-side optimization, and a universal drop-in for every affine-bound verifier. An optional LP tier restores exactness on the probability-measure ball and changes no decision across 2,176 tests.
2. **Envelopes are a free choice** (Proposition 2.2): the concretization depends only on the validity of the bound, not on the domain it was propagated on, so every ℓ_p verifier plugs in as an envelope generator without modification; the closed form is also the reference-free limit of flow-space lifting (Wegel et al., 2021), reproducing their certificates without a per-sample decomposition, reference distribution, or affine first layer (Appendix G).
3. **Per-sample radii, and envelopes beyond bound propagation** (Sections 2.3, 3.4, 3.5): an exact two-term gap decomposition localizes all conservatism to the envelope, and the interface admits generators that bound propagation cannot supply (Proposition 2.3): on the theorem-proved subset a learned (RL) envelope certifies radii $1.4\times$ larger than the per-sample α-CROWN teacher at a matched budget and 10–$100\times$ larger as the budget widens, every reported radius sound.

Section 2 develops the method; Section 3 evaluates it through four questions (Q1–Q4).

2 Method

The method is a drop-in replacement for the concretization step of an affine-bound verifier. Full derivations, the feasibility analysis with its exact LP, and implementation details are in Appendices A–C.

2.1 The valid-bound interface

Bound propagation produces affine lower and upper bounds on the network output, $\underline{A}\,x + \underline{b} \le f(x) \le \overline{A}\,x + \overline{b}$ for all x in a perturbation region $\mathcal{S}$. For classification, verifying a correctly classified input x_0 with label y_0 means showing that, for each margin $g_j(x) = f_{y_0}(x) - f_j(x)$, the affine lower bound supplied by the propagator (Definition 2.1) stays positive over $\mathcal{S}$:

$$\min_{x\in\mathcal{S}} A\cdot(x - x_0) + b \;>\; 0. \tag{1}$$

Definition 2.1 Valid-bound interface.
A bound propagator meets the interface at x_0 if, for every margin g_j, it supplies coefficients $A^{(j)}$ and an intercept $b^{(j)}$ with

$$g_j(x) \;\ge\; A^{(j)}\cdot(x - x_0) + b^{(j)} \qquad \text{for all } x \in \mathcal{S}.$$

Our method consumes only the pairs $(A^{(j)}, b^{(j)})$; how the propagator produced them is irrelevant.

Once such a bound is in hand, certification reduces to the minimization in equation 1, the *concretization* of the bound over $\mathcal{S}$. For an ℓ_p ball, Hölder's inequality closes it:

$$\min_{\|\delta\|_p\le\varepsilon} A\cdot\delta \;=\; -\varepsilon\,\|A\|_q, \qquad \frac{1}{p} + \frac{1}{q} = 1. \tag{2}$$

Proposition 2.2 The concretization is independent of the enclosing domain.
Let $D_0 \supseteq \mathrm{Ball}_{W_1}(\varepsilon)$ be any domain on which the propagator supplies a valid affine bound (e.g.

the per-coordinate box). Because the W_1 ball is a subset of D_0, a bound valid throughout D_0 remains valid throughout the ball; crucially, we then minimize this bound over the W_1 ball itself, not over D_0. The minimizer and its value therefore do not depend on which D_0 was used to obtain the bound—a larger domain only loosens the envelope (A, b) (tightness), never the validity of the ball certificate (soundness). Any affine-bound propagator can thus serve as an envelope generator without modification.

Proposition 2.3 Admission consumes validity, not provenance.
Let (A, b) be a candidate pair from any source and $h(x) := g_j(x) - A \cdot (x - x_0)$. If $\min_{\text{ball}} h \geq b$, then $g_j(x) \geq A \cdot (x - x_0) + b$ throughout the ball, so (A, b) meets the valid-bound interface (Definition 2.1) and the certificate is sound by the same transitivity. Moreover h is still a network—subtracting the affine term edits only the last layer of g_j—so the admission test *is* the concretization problem equation 1, applied to h.

A generator outside bound propagation—e.g. a learned policy—feeds the interface through Proposition 2.3: its candidate is admitted only after the check certifies validity (Section 2.2).

2.2 Closed form concretization for the W_1 ball

The Wasserstein perturbation model lets an adversary relocate mass between pixels, paying the ground-metric distance per unit mass moved, with total budget ε.

View an $h \times w$ image as the 4-neighbor grid graph $\mathcal{G} = (V, E)$, with B the vertex–edge incidence matrix, $(B^\top A)_e = A_v - A_u$. Perturbation $\delta = x' - x_0$ is a transport when it conserves mass, $\sum_i \delta_i = 0$; on the connected grid this is equivalent to $\delta = B\phi$, where ϕ_e is the signed net flow across edge e. The 1-Wasserstein distance is the cheapest such routing,

$$W_1(x_0, x') \;=\; \min_{\phi:\, B\phi = x' - x_0} \sum_{e \in E} c_e |\phi_e|,$$

where c_e is the edge price under the ground metric: $c_e = 1$ on neighboring pixels for the L1 (Manhattan) metric, while the L2 construction below uses Euclidean edge costs. The constraint $B\phi = x' - x_0$ distinguishes a W_1 ball from the per-coordinate box in Figure 1: mass is only relocated, never created or destroyed.

Theorem 2.4 The worst case over a W_1 ball is the budget times the steepest edge.
Let $\mathcal{G} = (V, E)$ be the 4-neighbor grid. For any $A \in \mathbb{R}^n$,

$$\min_{\{x' = x_0 + B\phi :\, \|\phi\|_1 \leq \varepsilon\}} A \cdot (x' - x_0) \;=\; -\varepsilon \|B^\top A\|_\infty. \tag{3}$$

On the 4-neighbor grid with unit edge costs, $(B^\top A)_e = A_v - A_u$ for $e = (u, v)$, so this specializes to the steepest-edge form

$$\min_{\{x' = x_0 + B\phi :\, \|\phi\|_1 \leq \varepsilon\}} A \cdot (x' - x_0) \;=\; -\varepsilon \max_{(u,v) \in E} |A_u - A_v| \,. \tag{4}$$

The ball $\{x_0 + B\phi : \|\phi\|_1 \leq \varepsilon\}$—the *flow ball*—is the W_1 ball on signed measures, a sound relaxation of the probability-measure W_1 ball, which additionally requires $x' \geq 0$ (equivalently, the pixel box $-m/s \leq x' \leq (1 - m)/s$ for standardized inputs). The flow-feasibility LP of Proposition A.2 projects the relaxed value onto that box-feasible subset.

Proof sketch. Any feasible perturbation decomposes into edge flows, $\delta = B\phi$ with $\|\phi\|_1$ equal to the transport cost. Hence $A \cdot \delta = (B^\top A) \cdot \phi \geq -\|B^\top A\|_\infty \|\phi\|_1 \geq -\varepsilon \|B^\top A\|_\infty$ by Hölder's inequality. Equality holds by pushing all ε across a steepest edge $e^* \in \arg\max_{(u,v) \in E} |A_u - A_v|$. The full proof is in Appendix B. □

The closed form costs $O(|E|) = O(n)$ tensor ops ($|E| = 2hw - h - w < 2n$ on the 4-neighbor grid), is GPU-parallel, and handles multi-channel images channel-wise; it is the default, solver-free concretization. An optional exactness tier adds a zero-cost per-sample flow-feasibility test checking whether the steepest-edge plan stays inside the pixel box (its source can give and its sink can receive ε mass) and, only when the test fails, solves a small flow-feasibility LP (Proposition A.2, Algorithm 1); the tier is off by default and changes no certification decision in our benchmarks.

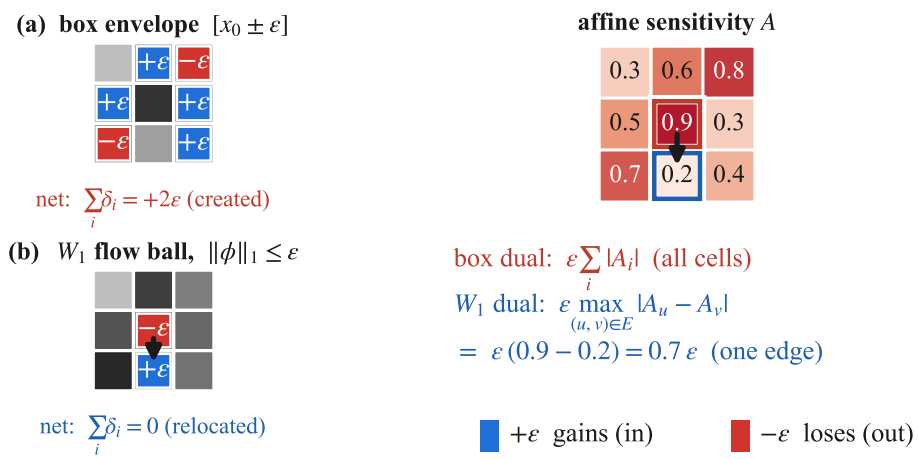


Figure 1: Why geometry matters, on one image and one affine sensitivity A. **Left:** (a) the per-coordinate box moves each cell independently, creating net mass $\sum_i \delta_i = +2\varepsilon \neq 0$; (b) the W_1 flow ball relocates a single adjacent pair, so $\sum_i \delta_i = 0$. **Right:** the box dual $\varepsilon \sum_i |A_i|$ charges every cell, while the W_1 dual $\varepsilon \max_{(u,v)\in E} |A_u - A_v|$ charges only the steepest edge ($0.9 \to 0.2$ here).

From closed form to admission. The admission test of Proposition 2.3 reduces to the same concretization: run any bound propagator on a superset of the ball (Proposition 2.2) to obtain an affine lower bound $L \cdot (x - x_0) + c$ on the residual h; by Theorem 2.4, $\min h \ge c - \varepsilon \|B^\top L\|_\infty$ on the ball, so the candidate is valid whenever $c - \varepsilon\|B^\top L\|_\infty \ge b$, and the certificate $b - \varepsilon\|B^\top A\|_\infty$ is then sound by the transitivity of Proposition 2.3. The cost is one propagation plus $O(n)$. Section 3.5 instantiates this check for learned envelopes, in its complete branch-and-bound form (Appendix F).

L2 ground metric. For the Euclidean ground metric the exact closed form is the all-pairs quantity $\max_{u \neq v} |A_u - A_v| / \|u - v\|_2$, costing $O(n^2)$; the 8-neighborhood stencil $\cos(\pi/8)^{-1} \max\left(\max_{\text{hv}} |A_u - A_v|,\ \max_{\text{diag}} |A_u - A_v| / \sqrt{2}\right)$ is a sound $O(n)$ upper bound on it, at most 8.2% looser.

2.3 Gap decomposition and certified radii

The closed form is exact on the signed-measure flow ball of Theorem 2.4; what makes that ball a relaxation is that the extremal plan—pushing the entire budget ε along the steepest edge—may accumulate into negative mass at a vertex, whereas a probability image must keep every pixel in the legal box (under mass conservation a deficit $x' < 0$ somewhere forces an excess $x' > 1$ elsewhere). For the standardized input $x_{\text{std}} = (x_{\text{raw}} - m)/s$ define

$$\mathcal{S}_{\text{orig}} = \{x' : -m/s \le x' \le (1-m)/s,\ W_1(x_0, x') \le \varepsilon\}, \qquad \mathcal{S}_{\text{relax}} = \{x_0 + B\phi : \|\phi\|_1 \le \varepsilon\},$$

so that $\mathcal{S}_{\text{orig}} \subseteq \mathcal{S}_{\text{relax}}$. The single-edge plan leaves $\mathcal{S}_{\text{orig}}$ exactly when an endpoint lacks box slack—the source is a background pixel that cannot give ε mass, or the sink is saturated and cannot receive it. The default solver-free concretization stays on $\mathcal{S}_{\text{relax}}$; when the test fails, one optional flow-feasibility LP—the same feasibility condition as flow-lifting verifiers—projects onto $\mathcal{S}_{\text{orig}}$ exactly.

Lemma 2.5 The certification gap decomposes into exactly two nonnegative terms.
If $A \cdot (x' - x_0) + b \le g_j(x')$ on $\mathcal{S}_{\text{orig}}$ (the valid-bound interface), then

$$\min_{\mathcal{S}_{\text{orig}}} g_j - \left[b - \varepsilon\|B^\top A\|_\infty\right] = \underbrace{\min_{\mathcal{S}_{\text{orig}}} g_j - \min_{\mathcal{S}_{\text{orig}}} \left[A \cdot (x' - x_0) + b\right]}_{\ge 0 \text{ (envelope gap)}} + \underbrace{\min_{\mathcal{S}_{\text{orig}}} A \cdot (x' - x_0) - \min_{\mathcal{S}_{\text{relax}}} A \cdot (x' - x_0)}_{\ge 0 \text{ (flow-feasibility gap)}}. \tag{5}$$

The two terms are the only two: the first is nonnegative by validity of the affine bound, the second by $\mathcal{S}_{\text{orig}} \subseteq \mathcal{S}_{\text{relax}}$, and the relaxed minimum is the closed form itself (Theorem 2.4). The flow-feasibility gap is zero whenever the single-edge transport stays inside the box (sufficient condition, Proposition A.1); otherwise the flow-feasibility LP (Proposition A.2) re-solves the extremal problem on the box and recovers the exact value.

From bound to certified radius. Propagating each margin g_j and concretizing gives the affine lower bound $m_j(\varepsilon) = b_j(\varepsilon) - \varepsilon \cdot \max_{(u,v)\in E} |A_u^{(j)}(\varepsilon) - A_v^{(j)}(\varepsilon)|$ at budget ε. The certified radius is

$$\varepsilon^* \;=\; \sup\big\{\varepsilon : \min_{j\neq y_0} m_j(\varepsilon) > 0\big\}. \tag{6}$$

When the envelope is fixed (e.g. a linear network), this collapses to $\varepsilon^* = \min_j b_j / \max_{(u,v)\in E} |A_u^{(j)} - A_v^{(j)}|$; when b_j and $A^{(j)}$ depend on ε (as with iterative bound propagation), we bisect the predicate $\varphi(\varepsilon) \triangleq \min_j m_j(\varepsilon) > 0$. The formula reuses the concretization's Lipschitz constant: the steepest-edge factor is the W_1-Lipschitz constant of the *envelope*, coinciding with the network's own W_1-Lipschitz constant only when the bound is exact. Soundness is immediate. The predicate is monotone at a fixed envelope (each m_j is then linear nonincreasing in ε, hence so is $\min_j m_j$); when the envelope is re-propagated at each ε monotonicity is not guaranteed, but every accepted ε is certified by its own evaluation, so bisection can only undershoot, never over-certify.

Radius consequence of the floor. The radius ε^* is the zero-crossing of $\min_j m_j$, so it inherits the flow-feasibility gap: if the binding margin's steepest-edge plan runs out of box slack at the crossing, the relaxed curve reaches zero *before* the LP curve does—the certified radius is then a strict underestimate of the probability-ball radius. The typical trigger on MNIST is a zero background pixel at the steepest edge's source, which cannot give mass; on dense CIFAR images the test rarely fires and the two radii coincide.

Completeness. Because the concretization is exact over the W_1 ball rather than an enclosing superset, fixing every ReLU phase makes the envelope exact and branch-and-bound over phases converges to the true Wasserstein worst case in the splitting limit, leaving no residual gap—unlike covering-based verifiers, whose gap survives exhaustive splitting. The argument is architecture-independent, and our splitting routine (`complete_radius`) implements it.

3 Experiments

3.1 Setup

Models and data. MNIST: MLP2/4/6, LeNet5, ConvNet3, DeeperCNN, PatchEmbedMLP, ConvLarge; CIFAR-10: CIFAR_MLP2/4, CIFAR_LeNet, CIFAR_PatchEmbedMLP, CIFAR_ConvLarge; plus two Wong–Kolter cert-trained MLP2 variants (cert-ε30/60), an SDP-CROWN Lipschitz-trained ConvLarge (*lip*) (Chiu & Zhang, 2025), and the three official VNN-COMP 2022 `mnist_fc` ONNX networks (Brix et al., 2023). Clean accuracies and evaluation pools are tabulated in Appendix E; all certified radii use n=100 correctly classified samples from a fixed pool unless noted.

Perturbation and algorithm. The threat model moves image mass along the 4-neighbor pixel grid (unit-cost horizontal and vertical edges, L1 ground metric; Theorem 2.4). Bounds come from α-CROWN (Xu et al., 2021) through auto-LiRPA (patches mode, final-specification α optimization) on one NVIDIA RTX 5090 (24 GB), with 15 cold-start and 5 warm-start α iterations per bound, warm-started across a 14-step radius bisection. Radii are medians with paired-bootstrap 95% CIs (10^4 resamples). Code and data are in the anonymized repository (Reproducibility statement).

Units. Radii are certified W_1 budgets ε^* in each model's own input space; ℓ_∞ and W_1 budgets use different units and are never compared numerically.

3.2 Q1: the closed form matches an exact LP at roughly $10^3\times$ lower cost

(a) Exactness. The closed form agrees with the primal LP $\min_\phi (B^\top A)\cdot\phi$ s.t. $\|\phi\|_1 \le \varepsilon$ (mass conservation, no box), solved per margin by HiGHS (Huangfu & Hall, 2018), to within 4.5×10^{-13}.

(b) The optional LP tier never changes a decision. The default certificate is the solver-free closed form, sound on its own; the LP upgrades only the optional *exactness tier* on the probability ball. Does the upgrade ever change a decision? We sweep $\varepsilon \in \{0.005, 0.02, 0.05, 0.1\}$ deliberately past

where certificates start failing, so that margins cross the decision boundary: across all 13 models and 2,176 (sample,ε) decisions the two tiers give *zero sign flips*, and the largest gap anywhere (5.4×10^{-2} on MNIST, 0.36 on CIFAR_ConvLarge) stays far below the margin it would overturn.

When the LP fires. It is reached only through a zero-cost single-edge test (Proposition A.1, Algorithm 1). The test fails only when the steepest-edge plan needs an endpoint without box slack—a pixel that cannot supply the mass or absorb it, on MNIST typically a zero background pixel—so the trigger rate tracks the share of such pixels (82% of MNIST pixels are zero vs. $< 0.4\%$ on CIFAR), and dense images stay solver-free end to end (per-ε tables: Appendix E.1).

Relaxed-ball radii. All radii in this paper are relaxed-ball radii, hence conservative by Section 2.3. A paired run on plain MLP2 (n=3) recovers 1.30–1.42$\times$ under the LP tier (Appendix E.1) without flipping any tested decision, and the gap is shared by every relaxed-ball W_1 certificate (Wegel et al., 2021) (Proposition G.1).

Table 1: Cost of one certification on the 4-neighbor grid: closed form (GPU figures amortize a 10-margin batch) against the exact LP tier. Each sample certifies all ten class margins.

Method	per margin	per sample (10 margins)
Closed form (CPU / GPU)	5–9 μs / $\approx$7.7 μs	50–90 μs / $\approx$77 μs
HiGHS LP	5–24 ms	50–240 ms

(c) Cost. Per margin the closed form is roughly $10^3\times$ cheaper than the LP tier (5–9 μs vs. 5–24 ms; Table 1), and end-to-end verification—bound propagation plus concretization—stays within 1.45$\times$ the wall-clock of an ℓ_∞-CROWN run—only the entry fold and the input specification change; the propagation sweeps that dominate the runtime are unchanged.

3.3 Q2: certified radii—flow geometry wins

Q2 asks how large the certified radii are and which design choices deliver them—envelope, concretization dual, and pipeline. Figure 2 starts with the envelope: the legacy per-coordinate box vs. the flow zonotope. A negative control first: the envelope is immaterial on MLPs (exact linear fold), which certify at the top of the bisection range.

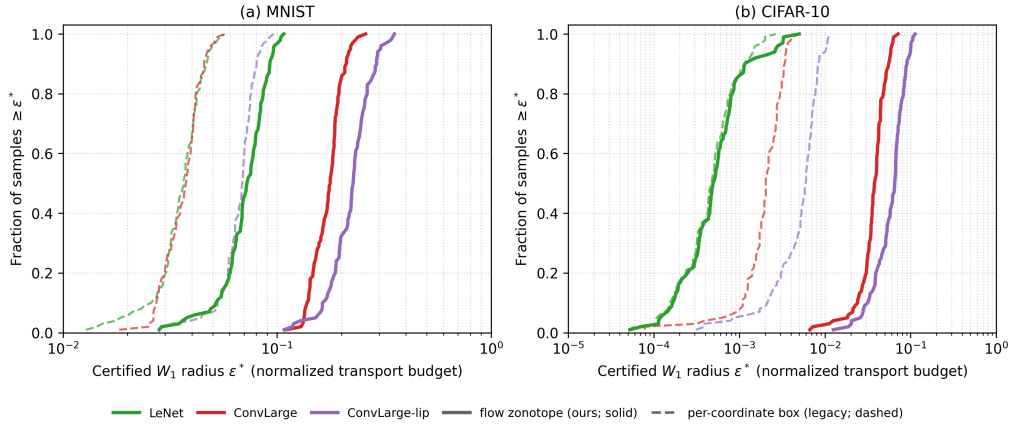


Figure 2: Certified-radius CDFs: flow-zonotope envelope (solid) vs. the legacy per-coordinate box envelope (dashed), same 100 correctly-classified samples per model, identical bound settings; rightward curves certify larger radii. Under the box envelope CNN radii collapse to a median 2×10^{-3}–7×10^{-2}; the flow zonotope lifts the medians by 2–19$\times$ on all but CIFAR_LeNet ($4.8\times10^{-4}\rightarrow5.1\times10^{-4}$) at no change to the concretization. Both curves certify the same W_1 threat model on the same samples; only the propagation envelope differs (the dashed curve is not an ℓ_∞ certificate), so the read is within-model—same color, solid vs. dashed. All curves are relaxed-ball radii (Section 3.2); floor-LP-tier counterparts and the discounted-but-intact lift are in Table 5.

The flow zonotope fixes the CNN box-collapse. Under the box envelope, CNN radii collapse to medians of 2×10^{-3}–7×10^{-2} (dashed, Figure 2). The collapse is an envelope artifact: covering each intermediate neuron by an independent coordinate box $[x_0 \pm \varepsilon]$ discards the mass-conservation coupling between adjacent pixels, and propagation returns correspondingly looser affine bounds. Switching the envelope to the flow zonotope $\{x_0 + B\phi : \|\phi\|_1 \leq \varepsilon\}$—the enclosing-domain freedom of Proposition 2.2—leaves everything else untouched and removes the collapse: on every deep CNN, in MNIST and CIFAR-10 alike, the certified-radius CDF shifts right by 2–19$\times$ at the median, with per-sample wins on 70–100 of 100 paired samples. The one near-exception is CIFAR_LeNet ($1.1\times$), where both envelopes already sit at $\approx 5 \times 10^{-4}$. The lift is not a relaxed-ball artifact: re-certified on the exact probability ball (floor-LP tier), it is unchanged on CIFAR ($17.8 \to 17.3\times$) and discounted but intact on MNIST ($4.8 \to 2.6\times$ on ConvLarge, $2.0 \to 1.6\times$ on LeNet5), where the floor recovers 1.4–$1.9\times$ for the box arm but only 1.00–$1.16\times$ for the flow zonotope (Table 5, Appendix E.1).

ℓ_p-reduction baselines. Since $\mathrm{Ball}_{W_1}(\varepsilon) \subseteq \mathrm{Ball}_\infty(\varepsilon)$, why not certify an ℓ_p ball and convert? Through the *same* α-CROWN pipeline, the W_1-native certificate wins on 100% of samples once every arm uses the flow-zonotope envelope: the margin grows to $12\times$ vs. the ℓ_∞ reduction on MNIST and 30–$50\times$ vs. the ℓ_∞ reduction on CIFAR (3.3–$3.7\times$ vs. ℓ_2 there). The ordering never inverts on any architecture we test: in this pipeline, only the W_1-native dual scales with envelope quality.

Flow-space lifting (Wegel et al., 2021), reimplemented and plugged in. The flow zonotope is the same geometry that flow-space lifting propagates through an off-the-shelf ℓ_1 verifier—per-sample flow decomposition relative to a reference, then first-layer folding—so the lifted arm is the natural next comparison. No public implementation exists, so we reimplement their pipeline in the same backend; the same closed form concretizes its bounds (its terminal value, worked out term by term, is our edge-drop dual: Appendix G, Corollary G.2). The two arms agree on ConvLarge to a median ratio of $1.004\times$ and on MLP2 to a median relative difference of $\sim 10^{-7}$; the box and SDP-CROWN baselines stay 1.8–$2.7\times$ looser. What lifting cannot supply is an envelope from outside bound propagation—that requires the admission check (Q4).

3.4 Q3: the radius is envelope-bound, and the envelope is pluggable

Two independent probes support one diagnosis: the certificate fails from bound looseness, not from true counterexamples. *(i) Attacks find no counterexample:* the projected-Sinkhorn attack of Wong et al. (2019) at matched budgets finds no adversarial example on any of the 34 CIFAR_LeNet samples tested (95% Clopper–Pearson upper bound $\leq 8.4\%$; budgets and protocol: Appendix E.4).

(ii) Bound quality drives the rate: through the same W_1 concretization, the certified radius tracks the tightness of the propagated envelope, and median radii across the deep-CNN family span two orders of magnitude (2×10^{-3}–0.36).

Where is the looseness? Lemma 2.5 splits the shortfall of the certified radius below the true robust radius into a concretization term (the relaxed-ball approximation of Section 2.3) and an envelope term (looseness of the propagated affine bounds). Probe (i) shows the shortfall is real: on samples we fail to certify, no attack succeeds, so a gap exists to attribute. The concretization term is measured directly by the flow-feasibility LP (Proposition A.2) and is $\leq 3 \times 10^{-7}$; what remains is the envelope. Nor is the gap a tuning artifact: exact bound injection and $20 \to 200$ α-iterations each move a margin by $< 0.03\%$, against the 2–$19\times$ lift from switching the envelope alone (Figure 2).

The envelope is a design choice. The certified radius does not depend on which enclosing domain it is verified over: with a complete branch-and-bound predicate, the W_1 ball, its box, and the ℓ_1-2ε ball all certify identical radii (Section 2.3). The domain only gates how tightly an *incomplete* propagator can close the gap (Q2)—so the envelope is a design freedom, not a commitment, and the generator behind it need not be a bound propagator at all: *learned* envelopes are admitted by the same validity check (Section 3.5).

3.5 Q4: certifying learned input-level envelopes

The closed form depends only on the bound, not on how it was produced (Proposition 2.2), so the envelope source is a free choice. We instantiate it with a *learned* envelope: a small policy network,

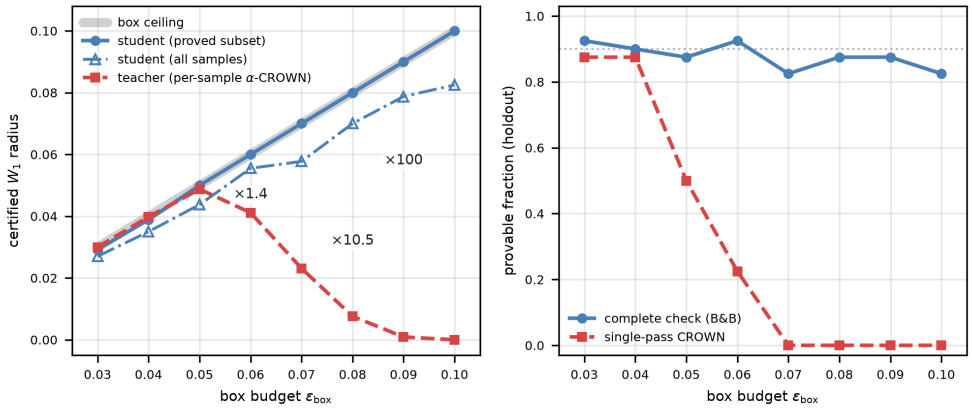


Figure 3: Learned envelopes keep the radius up and the certificate provable as the box grows (ConvLarge, n=40 holdout). **Left:** the per-sample α-CROWN radius (dashed red) collapses to zero by ε_{box}=0.10 while the learned radius on the theorem-proved subset (solid blue) rides the box ceiling (grey); counting abstentions as zero, the all-sample mean (dash-dot blue) keeps 82.5–92.5% of the ceiling. **Right:** the fraction of held-out samples the validity check *proves* sound: the complete branch-and-bound checker (solid blue) sustains 82.5–92.5% at every budget while a single CROWN pass (dashed red) collapses to 0 by ε_{box}=0.07. Reported radii are sound by the closed form.

trained with reinforcement learning to maximize the certified radius, that outputs the per-margin affine lower bounds $(A_m^{(j)}, b_m^{(j)})$ in one forward pass. It replaces the per-sample envelope-*generation* solve of the α-CROWN *teacher*; what remains per sample is only the admission check and the closed form.

Validity is measured at three levels: an empirical box-point audit (detects violations but cannot prove validity), a single CROWN pass, and the complete branch-and-bound check, which alone proves. Admission uses only the third: a candidate pair (A, b) is certified when this check proves it valid over the full ball—a branch-and-bound over ReLU phases on the residual network $g_j - A \cdot (x - x_0)$, the complete form of a single bound propagation (Algorithm 2, appendix). Any other outcome (a witnessed violation or budget exhaustion) is *abstention*, never a false certificate, and the sample falls back to the teacher. Learner quality therefore moves only the admission rate and the certified radius—what follows is a floor, not a ceiling.

Figure 3 shows the central effect on ConvLarge. As the enclosing-box budget ε_{box} grows (the pixel box of Section 2.3), the teacher radius collapses to zero by ε_{box}=0.10 while the learned radius rides the box ceiling, and the complete check *proves* the learned envelope valid on 82.5–92.5% of held-out samples at every budget—no crossing point where a larger budget forces an admission-rate collapse.

The effect is not an artifact of one architecture (Table 2): across three MNIST architectures the learned envelope tops the box ceiling with audit rates 0.80–0.97, and the complete check proves

Table 2: Learned envelopes: held-out audit rate (empirical box-point audit—detects but never proves validity), certified radius (all-sample mean, normalized ε), full-box fraction (share of radii reaching the ceiling ε_{box}), and proof-level admission rate (B&B-*proved*; ε_{box}=0.06; Appendix E.5). All radii on the proved subset are sound by the closed form. On CIFAR-10 every proved certificate is vacuous (radius 0): 10 witnessed over-claims and 19 abstentions of 50, consistent with its lower audit rate.

Architecture (dataset)	#params	held-out audit	cert. radius	full-box	provable (B&B)
LeNet5 (MNIST)	6.2×10^4	0.80	0.060	1.00	0.850
DeeperCNN (MNIST)	1.7×10^5	0.80	0.060	1.00	0.775
ConvLarge (MNIST)	1.97×10^6	0.97	0.060	1.00	0.925
ConvLarge (CIFAR-10)	1.97×10^6	0.70	0.035	0.58	0.42 (vacuous)

0.775–0.925 of each architecture's held-out samples. On CIFAR-10 the radius transfers (0.035 at audit 0.70) but the check proves only vacuous radius-0 certificates—theorem-level proofs currently hold on MNIST.

Beyond pixels: graded-cost text certificates (appendix). The same operator extends to weighted transport graphs, where the edge costs grade the edits—a step toward certificates for LLM alignment, not only vision. On a bag-of-clusters histogram of a document (Kusner et al., 2015) with a kNN graph over cluster centroids, Proposition E.1 certifies *graded* word-substitution budgets on any scorer affine in the document histogram. On a linear reward model over 300 UltraFeedback pairs (Cui et al., 2024), the certified radius corresponds to 1.1% of a response being synonym-rewritable, while the uniform-cost ℓ_∞ word-count box certifies nothing. Table 6 and the exactness audit are in Appendix E.2.

4 Related work

Bound propagation and concretization. Scalable verifiers—bound propagation, symbolic intervals, zonotopes, multi-neuron relaxations—share the paradigm "relaxation coefficients $\times$ the threat model's dual norm" (Weng et al., 2018); we extend it to Wasserstein constraints, where the dual is the maximal edge drop $\|B^\top A\|_\infty$, making any ℓ_p verifier a drop-in envelope generator on auto-LiRPA (Xu et al., 2020). Wang et al. (2021) also optimize bound coefficients at verification time, but *inside* the trust chain; our admission check consumes candidates from *outside* it—learned envelopes included—and abstains on failure.

Wasserstein threat models. Attacks came first (Wong et al., 2019; Wu et al., 2020). On the certification side, Levine & Feizi (2020) certify a smoothed classifier probabilistically and Kumar et al. (2023) give a distributional guarantee under W_2, while Wegel et al. (2021) take the deterministic route via flow-space lifting: the lifted ℓ_1 ball parametrizes the relaxed W_1 ball exactly, so their terminal certificate is isomorphic to ours (Section 3.3); our closed form obtains the same value reference-free, in $O(n)$, without a per-sample reference flow or an affine first layer. Distributionally robust optimization certifies *expected loss* over W_1 balls (Mohajerin Esfahani & Kuhn, 2018; Sinha et al., 2018); we certify per-input margins.

Semantic-edit certification. Beyond ℓ_p, work certifies named transformation classes—parametric geometric (Balunovic et al., 2019; Ruoss et al., 2021), semantic perturbations (Mohapatra et al., 2020), smoothing over composable families (Fischer et al., 2020; Korzh et al., 2024; Alfarra et al., 2022)—*lists* of families (the smoothing variants probabilistically). In NLP, set-based substitution certificates—IBP (Jia et al., 2019), SAFER (Ye et al., 2020), randomized-[MASK] smoothing (Zeng et al., 2023)—price every substitution in a synonym set at the same $0/1$ cost, treating "good→great" and "good→bad" identically; our closed form certifies *graded* costs on any scorer affine in the document representation.

5 Conclusion and discussion

We give bound propagation a closed-form Wasserstein concretization: the worst case of any valid affine bound under a 1-Wasserstein constraint is the budget times the maximal edge drop $\|B^\top A\|_\infty$—$O(n)$ on the pixel graph, exact on the relaxed ball, with an optional LP tier restoring exactness on the probability ball. Because the interface consumes validity, not provenance, any ℓ_p verifier is a ready-made envelope generator, and learned envelopes are admitted by a check built from the same closed form, which certifies or abstains but never issues a false certificate.

Limitations. The certified radius is bounded by envelope quality—a floor on the true W_1 radius, never a ceiling. The mass-preserving threat model does not cover insertions or deletions. The learned-envelope admission rate (0.775–0.925 proved on MNIST, vacuous on our CIFAR-10 setup) is the practical bottleneck; the learner itself is untuned.

AI use statement

Generative AI tools (OpenCode, Kimi K3, and Hermes Agent) assisted with implementing the methods, drafting and editing parts of the manuscript, and searching and formatting the literature. All

AI-generated code was reviewed, tested, and verified by the authors, and all AI-assisted text was checked for correctness. We take full responsibility for the final content of this work.

Reproducibility statement

Complete proofs and derivations are collected in the appendix; implementation details, all experimental tables, and per-sample data files are in Appendices C–G. Code, data, and experiment scripts are released, anonymized for review, at `https://anonymous.4open.science/r/wasserstein-concretization-6912`; paths such as `wasserstein/` elsewhere in the paper are relative to the repository root.

References

Motasem Alfarra, Adel Bibi, Naeemullah Khan, Philip H. S. Torr, and Bernard Ghanem. DeformRS: Certifying input deformations with randomized smoothing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6001–6009, 2022.

Mislav Balunovic, Maximilian Baader, Gagandeep Singh, Timon Gehr, and Martin Vechev. Certifying geometric robustness of neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Christopher Brix, Mark Niklas Müller, Stanley Bak, Taylor T. Johnson, and Changliu Liu. The first three years of the international verification of neural networks competition (VNN-COMP). *International Journal on Software Tools for Technology Transfer*, 25:329–339, 2023.

Hsi-Ming Chiu and Richard Y. Zhang. SDP-CROWN: Efficient bound propagation for neural network verification with tightness of semidefinite programming. In *International Conference on Machine Learning (ICML)*, 2025.

Ganqu Cui, Lifan Yuan, Ning Ding, Guangming Yao, Wei Zhu, Yuan Ni, Guotong Zheng, Zhiyuan Liu, and Maosong Sun. Ultrafeedback: Boosting language models with high-quality feedback. In *International Conference on Learning Representations (ICLR)*, 2024.

Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George J. Pappas. Efficient and accurate estimation of Lipschitz constants for deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

Marc Fischer, Maximilian Baader, and Martin Vechev. Certified defense to image transformations via randomized smoothing. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 8404–8417, 2020.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. Ai^2: Safety and robustness certification of neural networks with abstract interpretation. In *IEEE Symposium on Security and Privacy (S&P)*, 2018.

Sven Gowal, Krishnamurthy Dvijotham, Robert Stanforth, Rudy Bunel, Chongli Qin, Jonathan Uesato, Relja Arandjelovic, Timothy Mann, and Pushmeet Kohli. On the effectiveness of interval bound propagation for training verifiably robust models. In *International Conference on Learning Representations (ICLR)*, 2019.

Patrick Henriksen and Alessio Lomuscio. Efficient neural network verification via adaptive refinement and adversarial search. In *European Conference on Artificial Intelligence (ECAI)*, 2020.

Qi Huangfu and Julian A. J. Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018.

Robin Jia, Aditi R. Wang, Yash Khurana, and Anup Raghunathan. Certified robustness of neural networks against adversarial word substitutions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2249–2257, 2019.

Dmitrii Korzh, Mikhail Pautov, Olga Tsymboi, and Ivan Oseledets. General Lipschitz: Certified robustness against resolvable semantic transformations via transformation-dependent randomized smoothing. In *Proceedings of the 27th European Conference on Artificial Intelligence (ECAI)*, 2024.

Aounon Kumar, Alexander Levine, Tom Goldstein, and Soheil Feizi. Provable robustness against Wasserstein distribution shifts via input randomization. In *International Conference on Learning Representations (ICLR)*, 2023.

Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Q. Weinberger. From word embeddings to document distances. In *International Conference on Machine Learning (ICML)*, pp. 957–966. PMLR, 2015.

Alexander Levine and Soheil Feizi. Wasserstein smoothing: Certified robustness against Wasserstein adversarial attacks. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS)*, pp. 3938–3947, 2020.

Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1–2):115–166, 2018.

Jeet Mohapatra, Tsui-Wei Weng, Pin-Yu Chen, Sijia Liu, and Luca Daniel. Towards verifying robustness of neural networks against a family of semantic perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 244–252, 2020.

Gabriel Peyré and Marco Cuturi. Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5–6):355–607, 2019.

Anian Ruoss, Maximilian Baader, Mislav Balunovic, and Martin Vechev. Efficient certification of spatial robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 2504–2513, 2021.

Zhouxing Shi, Yihan Wang, Huan Zhang, J. Zico Kolter, and Cho-Jui Hsieh. Efficiently computing local Lipschitz constants of neural networks via bound propagation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Gagandeep Singh, Timon Gehr, Matthew Mirman, Markus Püschel, and Martin Vechev. Fast and effective robustness certification. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Gagandeep Singh, Rupanshu Ganvir, Markus Püschel, and Martin Vechev. Beyond the single neuron convex barrier for neural network certification. In *Advances in Neural Information Processing Systems*, volume 32, 2019a.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. In *Proceedings of the ACM on Programming Languages (POPL)*, volume 3, pp. 1–30, 2019b.

Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*, 2018.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272, 2020.

Shiqi Wang, Kexin Pei, Justin Whitehouse, Junfeng Yang, and Suman Jana. Efficient formal safety analysis of neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Shiqi Wang, Huan Zhang, Kaidi Xu, Xue Lin, Suman Jana, Cho-Jui Hsieh, and J Zico Kolter. Beta-CROWN: Efficient bound propagation with per-neuron split constraints for neural network robustness verification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 34, 2021.

Tobias Wegel, Felix Assion, David Mickisch, and Florens Greßner. A framework for verification of Wasserstein adversarial robustness. *arXiv preprint arXiv:2110.06816*, 2021.

Lily Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Luca Daniel, Duane Boning, and Inderjit Dhillon. Towards fast computation of certified robustness for ReLU networks. In *International Conference on Machine Learning (ICML)*, pp. 5276–5285. PMLR, 2018.

Eric Wong, Leslie Rice, and J Zico Kolter. Wasserstein adversarial examples via projected Sinkhorn iterations. In *International Conference on Machine Learning (ICML)*, pp. 6808–6817. PMLR, 2019.

Kaiwen Wu, Allen Houze Wang, and Yaoliang Yu. Stronger and faster Wasserstein adversarial attacks. In *International Conference on Machine Learning (ICML)*, pp. 10377–10387. PMLR, 2020.

Kaidi Xu, Zhouxing Shi, Huan Zhang, Yihan Wang, Kai-Wei Chang, Minlie Huang, Bhavya Kailkhura, Xue Lin, and Cho-Jui Hsieh. Automatic perturbation analysis for scalable certified robustness and beyond. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1129–1141, 2020.

Kaidi Xu, Zhouxing Li, Guannan Shi, Hao Du, Minlie Huang, Huan Sun, and Huan Zhang. Fast and complete: Enabling complete neural network verification with rapid and massively parallel incomplete verifiers. In *International Conference on Learning Representations (ICLR)*, 2021.

Mao Ye, Chengyue Gong, and Qiang Liu. SAFER: A structure-free approach for certified robustness to adversarial word substitutions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 3465–3475, 2020.

Jiehang Zeng, Jianhan Xu, Xiaoqing Zheng, and Xuanjing Huang. Certified robustness to text adversarial attacks by randomized [mask]. *Computational Linguistics*, 49(2):395–423, 2023.

Huan Zhang, Tsui-Wei Weng, Pin-Yu Chen, Cho-Jui Hsieh, and Luca Daniel. Efficient neural network robustness certification with general activation functions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

This appendix collects the proofs and supporting details for the paper's claims: the flow-feasibility analysis (pixel box) and its exact LP, the Lipschitz-constant proof, implementation details, the benchmark data, and the comparison protocols.

A Analysis of the flow-feasibility relaxation

We use the sets $\mathcal{S}_{\text{orig}} \subseteq \mathcal{S}_{\text{relax}}$ defined before Lemma 2.5 in Section 2.3: $\mathcal{S}_{\text{relax}}$ is the unconstrained flow ball (signed measures), and $\mathcal{S}_{\text{orig}}$ restricts to transports that stay in the legal pixel box $-m/s \leq x' \leq (1-m)/s$. The closed-form value is attained by the single-edge plan $\delta^* = \varepsilon(e_{u^*} - e_{v^*})$, and the only question is whether that plan is feasible on the box. This section proves the gap lemma and answers *(i) when is the flow-feasibility gap zero? (ii) when nonzero, how to compute it exactly?*

Proof of Lemma 2.5. Add and subtract $\min_{x' \in \mathcal{S}_{\text{orig}}} [A \cdot (x' - x_0) + b]$:

$$\min_{\mathcal{S}_{\text{orig}}} g_j - \big[b - \varepsilon \|B^\top A\|_\infty\big] = \Big(\min_{\mathcal{S}_{\text{orig}}} g_j - \min_{\mathcal{S}_{\text{orig}}} [A\cdot(x'-x_0)+b]\Big) + \Big(\min_{\mathcal{S}_{\text{orig}}} [A\cdot(x'-x_0)+b] - \big[b - \varepsilon\|B^\top A\|_\infty\big]\Big).$$

The first bracket is the *envelope gap*, nonnegative because $A \cdot (x' - x_0) + b \leq g_j$ on $\mathcal{S}_{\text{orig}}$. The constant b cancels in the second bracket, giving the *flow-feasibility gap* $\min_{\mathcal{S}_{\text{orig}}} A \cdot (x' - x_0) - \min_{\mathcal{S}_{\text{relax}}} A\cdot(x'-x_0)$, nonnegative because $\mathcal{S}_{\text{orig}} \subseteq \mathcal{S}_{\text{relax}}$. The closed form $\min_{\mathcal{S}_{\text{relax}}} A\cdot(x'-x_0) = -\varepsilon\|B^\top A\|_\infty$ is exact (Theorem 2.4), so no third term appears: the entire non-envelope loss is whether the single-edge extremal flow stays inside the pixel box. □

Proposition A.1 If both endpoints have box slack, the flow-feasibility gap is zero.
Let $\mathcal{G} = (V, E)$ be any transport graph, and let $e^* = (u^*, v^*)$ be the edge attaining the maximal

drop with $A_{v^*} \geq A_{u^*}$, so the extremal plan pushes ε mass out of the source v^* and into the sink u^*. If both endpoints have enough box slack,

$$x_{0,v^*} - \varepsilon \geq -\frac{m}{s} \qquad \text{and} \qquad x_{0,u^*} + \varepsilon \leq \frac{1-m}{s}, \tag{7}$$

then the single-edge plan is feasible on the box and the flow-feasibility gap is zero:

$$\min_{x' \in \mathcal{S}_{\text{orig}}} A \cdot (x' - x_0) = -\varepsilon \, \|B^\top A\|_\infty. \tag{8}$$

The first inequality (raw form $x^{\text{raw}}_{0,v^*} \geq \varepsilon s$) asks the source to supply the mass, the second asks the sink to receive it. The condition is sufficient, not necessary: equality holds iff the steepest-edge subgraph routes *any* budget-ε, sign-aligned flow that keeps every vertex in the box, and equation 7 is its cheapest-to-check special case. When the test fails the mass is rerouted along more edges—computed by the feasibility LP below.

Proof. "$\geq$" direction: follows from $\mathcal{S}_{\text{orig}} \subseteq \mathcal{S}_{\text{relax}}$ and the closed form (Theorem 2.4). "$\leq$" direction: take $\delta^* = \varepsilon(e_{u^*} - e_{v^*})$: only u^* and v^* change, and equation 7 keeps both inside the box, so $\delta^* \in \mathcal{S}_{\text{orig}}$ and $A \cdot \delta^* = -\varepsilon \|B^\top A\|_\infty$. Necessity and sufficiency: the equality-attaining flows of the relaxed problem are exactly those that are sign-aligned, supported on the steepest-edge set (edges with sub-maximal drop have strictly lower per-unit-budget return, so they do not attain equality), and satisfy $\|\phi\|_1 = \varepsilon$; if the boxed minimum equals the relaxed minimum, an equality-attaining perturbation must be one of these flows, so we need exactly one that is feasible against the box. For multi-channel inputs the drops are combined by $\max$ across channels, so checking the condition in the channel attaining the global maximum suffices. □

Remark Failure diagnosis.
The test fails when an endpoint lacks box slack. The common failure on MNIST is the source: the steepest edge starts at a background pixel (raw value 0) that cannot give mass; a saturated sink is rare on real margin rows. Q1 (b) confirms the diagnosis: the LP is strictly tighter exactly when the test fails.

Proposition A.2 A small LP computes the exact worst case on the pixel box.
When the single-edge plan fails equation 7, its exact value over $\mathcal{S}_{\text{orig}}$ is given by the following standard linear program, which restricts the transport to flows that keep every pixel in the box:

$$\begin{aligned} \text{val}_{\text{LP}}(A, x_0, \varepsilon) \;=\; \min_{\phi,\, t} \quad & (B^\top A) \cdot \phi \\ \text{s.t.} \quad & -\tfrac{m}{s} \;\leq\; x_0 + B\phi \;\leq\; \tfrac{1-m}{s} \quad \text{(component-wise)}, \\ & t_e \;\geq\; \phi_e, \quad t_e \;\geq\; -\phi_e, \quad \sum_{e \in E} t_e \;\leq\; \varepsilon, \qquad \forall\, e \in E. \end{aligned} \tag{9}$$

Then $\text{val}_{\text{LP}} = \min_{x' \in \mathcal{S}_{\text{orig}}} A \cdot (x' - x_0)$, and the flow-feasibility gap is

$$\Delta_{\text{box}}(A, x_0, \varepsilon) \;=\; \text{val}_{\text{LP}} + \varepsilon \, \|B^\top A\|_\infty \;\geq\; 0. \tag{10}$$

Proof. The t-variable linearization of the ℓ_1 budget satisfies $t_e = |\phi_e|$ at the optimum (otherwise reducing t_e does not change ϕ or the objective, a contradiction), so the feasible set of equation 9 coincides pointwise with $\mathcal{S}_{\text{orig}}$, the objective is unchanged, giving the first identity. The second follows by subtracting the closed form (Theorem 2.4) from the gap decomposition (Lemma 2.5). □

Complexity. The constraint matrix is the incidence matrix B plus the ℓ_1 linearization, all coefficients in $\{-1, 0, 1\}$—a minimum-cost-flow-type LP. Size: $2|E|$ variables and $2|E| + 2n + 1$ constraints (28×28 grid: $|E| = 1512$); solve times are in Section 3.2. The single-edge test is free, so the LP runs only when an endpoint lacks slack.

B Computation of the W_1-Lipschitz constant

Section 2.2 uses the Lipschitz constant formula for affine functions under the W_1 metric:

$$\text{Lip}_{W_1}(x \mapsto A \cdot x) \;=\; \max_{(u,v) \in E} |A_u - A_v| \;=\; \|B^\top A\|_\infty.$$

Upper bound ($\leq$). Let $g(x) = A \cdot x$. For any x, x', set $\delta = x - x'$. By the flow decomposition of Levine & Feizi (2020), $W_1(x', x) = \min_{x=x'+B\phi} \|\phi\|_1$; taking the optimal flow ϕ^* gives $\delta = B\phi^*$ and $\|\phi^*\|_1 = W_1(x', x)$. Hence

$$|g(x) - g(x')| \;=\; |A \cdot B\phi^*| \;=\; |(B^\top A) \cdot \phi^*| \;\leq\; \|B^\top A\|_\infty \cdot \|\phi^*\|_1 \;=\; \|B^\top A\|_\infty \cdot W_1(x', x),$$

by Hölder's inequality ($|\langle a, b\rangle| \leq \|a\|_\infty \cdot \|b\|_1$).

Equality ($\geq$). Take $e^* = (u^*, v^*) \in \arg\max_{(u,v)\in E} |A_u - A_v|$ and construct $\delta^* = \varepsilon(e_{u^*} - e_{v^*})$ as in the proof of Proposition A.1 (push ε mass across the steepest edge). Then $W_1(x_0, x^*) = \varepsilon$ and $|g(x^*) - g(x_0)| = \varepsilon \|B^\top A\|_\infty$. The two directions together give equality; the weighted-graph extension (edge costs c_e, maximal edge ratio $\max_e |A_u - A_v|/c_e$) is Proposition E.1.

C Implementation details

Algorithm 1 specifies the optional exactness tier layered on the closed form: it tests with zero extra cost whether the single-edge steepest-edge plan stays inside the pixel box (source can give and sink can receive ε mass) and, only if not, solves the flow-feasibility LP; by default the closed-form value alone is used.

Hardware. All benchmark radii are computed on a single 24 GB GPU using sparse forward intermediate bounds with final-specification α optimization.

Algorithm 1 Optional exactness tier: flow-feasible Wasserstein concretization

Require: coefficients $A \in \mathbb{R}^{H\times W}$, offset b, reference input $x_0 \in \mathbb{R}^{H\times W}$ (box $[-m/s,\ (1-m)/s]$), budget $\varepsilon \geq 0$

$L \leftarrow 0$ ▷ maximum edge drop
for $i = 1, \ldots, H$; $j = 1, \ldots, W-1$ **do** ▷ horizontal edges
  $L \leftarrow \max\big(L,\ |A_{i,j} - A_{i,j+1}|\big)$
end for
for $i = 1, \ldots, H-1$; $j = 1, \ldots, W$ **do** ▷ vertical edges
  $L \leftarrow \max\big(L,\ |A_{i,j} - A_{i+1,j}|\big)$
end for
$(u^*, v^*) \leftarrow \arg\max_{(u,v)\in E} |A_u - A_v|$ ▷ steepest edge, w.l.o.g. $A_{v^*} \geq A_{u^*}$
if $x_{0,v^*} - \varepsilon \;\geq\; -\frac{m}{s}$ **and** $x_{0,u^*} + \varepsilon \;\leq\; \frac{1-m}{s}$ **then** ▷ flow-feasibility: source supply and sink room
  return $b - \varepsilon \cdot L$ ▷ single-edge plan is feasible; closed form is exact
else
  solve the flow-feasibility LP of Proposition A.2
  $\mathrm{val}_{\mathrm{LP}} \leftarrow$ its optimal value
  return $b + \mathrm{val}_{\mathrm{LP}}$
end if

Remarks.

- *Multi-channel inputs.* The two edge loops run independently per channel and the channel-wise drops are combined following Levine & Feizi (2020); the rest is unchanged.
- *Numerical notes.* All quantities are computed in the normalized (probability) space and scaled back by s only through $\varepsilon_{\mathrm{eff}}$, so the computation is well-conditioned at radii down to 10^{-4}.

D VNN-COMP benchmark table

E Experiment details

Protocols and data tables supporting Section 3.

Table 3: Certified per-sample W_1 radii on the official VNN-COMP 2022 `mnist_fc` networks (grid2d-L1, α-CROWN, 100 correctly classified samples each; bracketed values are the 2.5% and 97.5% sample percentiles). Nineteen samples of the k=6 model fall back to plain CROWN (a deterministic auto-LiRPA issue; still sound). The official networks take raw $[0, 1]$ inputs while ours are standardized, so absolute radii are not comparable across the two blocks. All radii are relaxed-ball radii.

Network (ONNX)	Median radius (ε)	Percentiles [2.5%, 97.5%]	CROWN fallback
`mnist-net_256x2`	2.24	[0.75, 4.32]	0/100
`mnist-net_256x4`	2.55	[0.896, 4.28]	0/100
`mnist-net_256x6`	2.51	[0.53, 4.4]	19/100

Table 4: Clean test-set accuracy of every benchmark checkpoint (full 10k test images, evaluated once with the frozen weights) and the evaluation pools from which the certified samples in each table are drawn. "Correct in pool" explains the small n on CIFAR: e.g. CIFAR_MLP2 classifies 51.8% of all test images correctly, and 21 of the first 50 test images correctly—the latter is the pool the n=21 ablation rows use.

Model	Test acc.	Eval. pool	Correct	Used as
MNIST				
MLP2 baseline	97.5%	first 500	491	radius CDF and cert-training rows (n=100)
MLP2 cert-ε30	96.8%	first 500	490	radius CDF and cert-training rows (n=100)
MLP2 cert-ε60	94.2%	first 500	470	radius CDF and cert-training rows (n=100)
MLP4 / MLP6	97.6 / 98.0%	first 200	195 / 194	IBP ablation
LeNet5 / ConvNet3	98.8 / 98.9%	first 500	497 / 494	radius/ablation tables
DeeperCNN	97.9%	first 120	118	IBP ablation
PatchEmbedMLP	97.9%	first 500	495	floor-decision sweep
ConvLarge / *lip*	99.3 / 97.7%	first 100/500	100 / 490	radius tables and ℓ_p-reduction comparisons
CIFAR-10				
CIFAR_MLP2	51.8%	first 50	21	ablation rows (n=21)
CIFAR_MLP4	55.0%	first 50	28	ablation rows (n=28)
CIFAR_LeNet	75.9%	first 50	34	attack/ablation (n=34)
CIFAR_PatchEmb.	49.9%	first 50	24	ablation rows (n=24)
CIFAR_ConvLarge	86.7%	first 500	442	radius and learned-envelope transfer rows (n=100)
CIFAR_ConvLarge *lip*	36.7%	first 500	155	ℓ_p-reduction preview (n=38)

Proposition E.1 The closed form extends to any weighted graph.
Let $\mathcal{G} = (V, E)$ be connected with strictly positive edge costs c_e and ground metric given by the shortest-path distance $d_{\mathcal{G}}$. For any $A \in \mathbb{R}^{|V|}$, the closed form on the relaxed (signed-measure) ball is $\min_{W_1^{\mathcal{G}}(x_0,x')\le\varepsilon} A\cdot(x'-x_0) = -\varepsilon\cdot\max_{(u,v)\in E}|A_u - A_v|/c_e$, computed in $O(|E|)$ by the edge-collapse argument; restricting to box-feasible flows is a sound subset handled by the same feasibility LP.

The unit-cost grid is Theorem 2.4; this weighted form is what the text-rewrite experiment uses.

E.1 LP comparison protocol (Section 3.2)

Two tiers are distinguished.

- *Concretization exactness.* The primal LP $\min_\phi (B^\top A)\cdot\phi$ s.t. $\|\phi\|_1 \le \varepsilon$ (mass conservation, no box) is solved per margin row with SciPy HiGHS (Virtanen et al., 2020; Huangfu & Hall, 2018); the sweep uses the lower-face LP, and a separate two-sided run confirms the upper face adds a median gap $\le 3\times10^{-7}$ with no further changed decisions.
- *Flow-feasibility tier.* The zero-cost test of Proposition A.1 gates an LP solved by branch-and-bound on the LP relaxation.

A radius-level paired check makes the recovered headroom explicit: plain MLP2 (n=3 samples run in both tiers, zero-pixel-heavy inputs where the test fires often) certifies 1.30–1.42× the relaxed

radius under the floor-LP tier (median $1.41\times$), while the 13-model decision sweep above remains flip-free. Headline numbers are reported in Section 3.2; per-model-and-budget certified counts, the full decision-preservation records (`results/floor_decision/`), and the extended-sweep artifact (`wasserstein/experiments/floor/diag_floor_decision.py --eps 0.005 0.02 0.05 0.1`, records in `results/floor_decision_wide/`) are released under `wasserstein/`.

The Figure 2 lift under the floor-LP tier. The headline radii of Figure 2 are relaxed-ball radii; Table 5 re-certifies the same 100 paired samples per model on the exact probability ball (floor-LP tier for the flow zonotope; the box envelope clipped to the valid pixel range $[\max(x_0-\varepsilon, L), \min(x_0+\varepsilon, U)]$, a sound box envelope of the same probability ball). Two patterns emerge. On CIFAR—dense images where the zero-cost test rarely fires—both arms and the lift are unchanged ($17.8\times \to 17.3\times$ on CIFAR_ConvLarge). On MNIST the floor recovers 1.4–$1.9\times$ for the *box* arm (its raw box reaches far outside the valid pixel range on zero-background pixels) but only 1.00–$1.16\times$ for the flow zonotope, whose relaxed concretization already enforces mass conservation at the input; the lift is therefore discounted from $4.8\times$ ($2.0\times$) to $2.6\times$ ($1.6\times$) on ConvLarge (LeNet5) but survives on every model. Per-sample medians in normalized pixel units, paired across arms; records under `wasserstein/results/fig2_lp/`.

Table 5: Relaxed-ball vs. floor-LP-tier certified radii (median, normalized pixel units, n=100 paired samples) and the zonotope/box lift of Figure 2.

	box envelope		flow zonotope		lift (zono/box)	
Model	relaxed	floor-LP	relaxed	floor-LP	relaxed	floor-LP
MNIST ConvLarge	1.3×10^{-4}	2.3×10^{-4}	5.9×10^{-4}	5.9×10^{-4}	$4.8\times$	$2.6\times$
MNIST LeNet5	1.3×10^{-4}	1.8×10^{-4}	2.5×10^{-4}	2.9×10^{-4}	$2.0\times$	$1.6\times$
CIFAR ConvLarge	6.1×10^{-7}	6.4×10^{-7}	1.2×10^{-5}	1.2×10^{-5}	$17.8\times$	$17.3\times$
CIFAR LeNet	1.5×10^{-7}	1.5×10^{-7}	1.6×10^{-7}	1.6×10^{-7}	$1.1\times$	$1.1\times$

E.2 Graded-cost text certificates (Section 3.5)

Motivation: alignment. LLM outputs are increasingly scored, ranked, and filtered by reward models and LLM judges, so the certified object here is the aligner's verdict, not the policy: a W_1 certificate on the response bounds how much of it can be cheaply rewritten—synonyms and near-paraphrases priced by transport distance—before the preference ranking flips, a guarantee that $0/1$ synonym-set budgets cannot express.

Transport view of documents. A response is a histogram over 2000 clusters of a frozen embedding table (bge-small-en-v1.5; nomic-embed-text-v1.5 as a second geometry), the transport graph is a kNN graph over cluster centroids with inverse-cosine edge costs $c_e = 1/\max(\cos(u,v), 0.05)$, and a perturbation is a word substitution priced by shortest-path distance (Kusner et al., 2015). The edge-drop closed form extends from unit grid costs to these weighted costs unchanged (Proposition E.1).

Instantiation. On a linear reward model over bag-of-clusters histograms (pairwise AUC 0.62) trained on 300 UltraFeedback pairs (Cui et al., 2024), ε^* is the largest budget within which no substitution flips the ranking; any scoring head affine in the document histogram inherits the same certificate.

As Table 6 shows, transport budgets certify radii corresponding to 1.1% of a response being synonym-rewritable, while the uniform-cost ℓ_∞ box certifies nothing. The synonym-subgraph radius matches the full graph ($1.11\times$): the binding edge lies between two function-word clusters, not antonym edges.

Soundness of the text certificate. The 18-pair LP ground truth matches on every pair, margins stay positive at ε^*, and a matched-budget projected-Sinkhorn attack finds no counterexample (Clopper–Pearson 95% upper bound 14%). We certify the reward model, not the policy; representation limits (word order, insertions/deletions) are in Section E.3.

Table 6: Certified radius of a reward model's preference margin under semantic rewrites (median [p10, p90] over 300 pairs); "Equiv. subs." is the synonym-rewritable response fraction; the 18-pair exact floor-LP recomputation matches both W_1 rows. $p10 = 0$ certifies zero radius; nomic-embed-text-v1.5 results in Section E.3.

Threat model	ε^* median [p10, p90]	Equiv. subs.	LP check
W_1 full graph (bge-small-en-v1.5)	0.0130 [0.0000, 0.0732]	1.1%	match
W_1 cheapest-edge subgraph (bge-small-en-v1.5)	0.0145 [0.0000, 0.0974]	1.19%	match
ℓ_∞ box (word-count budget)	$< 10^{-3}$	$<10^{-3}$%	—

E.3 Semantic experiment: reward model and exactness (Section 3.5)

Construction. From the UltraFeedback preference corpus (train-prefs split, 62k pairs) we tokenize responses (lowercase alphabetic tokens of length ≥ 3), keep the top-2500 most frequent words, embed each with the frozen encoder (bge-small-en-v1.5: 384-d; nomic-embed-text-v1.5: 768-d; L2-normalized), and cluster with k-means into K=2000 clusters; a response becomes a probability histogram over clusters.

The transport graph is the symmetric 10-NN graph over cluster centroids with edge cost $c_e = 1/\max(\cos(u, v), 0.05)$ ("Equiv. subs." divides ε^* by the mean cost $\bar{c}_{\text{syn}}$ of the cheapest 20% of edges: $\bar{c}_{\text{syn}} = 1.22$ for bge-small-en-v1.5, 1.36 for nomic-embed-text-v1.5).

The reward model is a single linear scorer $r(h) = w \cdot h - b$ trained by logistic regression on the cluster histograms (chosen $= +1$, rejected $= -1$; held-out pairwise AUC 0.62); only the chosen-side histogram is perturbed, the rejected side is folded into the bias.

Exactness. Because the certified network is linear, the propagated affine envelope is exact; all semantic certificates use the LP floor (`floor=lp`), since the histogram's zero clusters make the nonnegativity floor active here (median floor gap 31%), unlike the pixel domain where it never binds.

On an 18-pair positive-margin subsample we recompute the certified radius from scratch by an independent exact floor-LP bisection: the two agree on every pair, for both the full graph and the synonym subgraph (Table 6, LP column). Attack budgets are the per-pair certified radii; the projected-Sinkhorn attack runs for 200 steps with row-normalized (mass-preserving) transport plans.

No negation or antonym rule is hand-curated: the binding edge of the certified dual $\max_e |w_u - w_v|/c_e$ connects two function-word clusters, and the two polarity-bearing edges among the 100 steepest have ratios below 15 (the binding edge is 22.3), so they never bind.

E.4 Envelope quality: supplementary data (Section 3.4)

Attack protocol. Projected Sinkhorn attack (Wong et al., 2019) with the same `grid2d` ground metric and exactly matched budgets; all 34 correctly classified CIFAR_LeNet samples at W_1 budgets $\varepsilon \in \{0.02, 0.05, 0.1, 0.2\}$ (zero successes; bound in main text).

E.5 Learned envelopes: supplementary data (Section 3.5)

Setup. The policy is a 63k-parameter CNN mapping a raw image to per-margin affine coefficients $(A_m^{(j)}, b_m^{(j)})$ through a full-resolution head (per-pixel on MNIST, per-(channel,pixel) on CIFAR), trained by reinforcement learning on a *teacher-free* corpus: the reward is the minimum-margin radius plus a validity penalty on fixed box points, with no per-sample teacher anchor. The certified radius is the closed form $\min_j b_m^{(j)}/\|B^\top A_m^{(j)}\|_\infty$. We did not tune the learner (capacity, reward shaping, curriculum), so the reported rates are a floor; stronger learners inherit the same guarantee. All validity figures below are held-out; the theorem-level check—not the sampled box-point audit—is what certifies.

Theorem-level check vs. box-point audit. The validity checker reduces a learned candidate to a propagated affine bound on the residual network $g_j - A \cdot (x - x_0)$ and certifies when its lower bound, concretized by the same closed form, covers b_m. A single bound-propagation pass is sound but incomplete: it abstains whenever its relaxation is too loose, and on the learned high-slope residual it abstains on up to 97% of large-budget samples.

We therefore run the *complete* form: branch-and-bound over ReLU phases (Appendix F) on the residual, resolving each abstention into a proof of validity or a witnessed real violation. Only a proof certifies; a witnessed violation is a genuine learner over-claim, and budget exhaustion remains an honest abstention. Algorithm 2 states the two-tier procedure.

Algorithm 2 Two-tier verification: learned envelope with complete admission check

Require: input x_0 (label y_0); network f; learned policy π; budget ε; node budget $N_{\max}$; time limit τ; teacher $\mathcal{T}$ (per-sample α-CROWN)

1: $(A_j, b_j)_{j \neq y_0} \leftarrow \pi(x_0)$ ▷ candidate affine minorant of each margin $g_j = f_{y_0} - f_j$
2: $r \leftarrow \min_j \; b_j \,/\, \|B^\top A_j\|_\infty$ ▷ candidate radius from the closed form
3: $\mu(x) \leftarrow \min_j \; g_j(x) - A_j \cdot (x - x_0) - b_j$ ▷ residual margin: candidate valid $\iff$ $\min_{\text{ball}} \mu \geq 0$; μ *is one network—the affine term edits only the last layer of* g_j
4: $d \leftarrow \textsc{Check}\big(\mu, \varepsilon, N_{\max}, \tau\big)$ ▷ decide $\min_{\text{ball}} \mu \geq 0$ *completely*
5: **if** $d =$ proved **then**
6: **return** $(r,$ learned$)$ ▷ every minorant valid on the ball
7: **else if** $d =$ violation **then**
8: **return** $\mathcal{T}(x_0, \varepsilon)$ ▷ genuine over-claim; discard
9: **else**
10: **return** $\mathcal{T}(x_0, \varepsilon)$ ▷ abstain
11: **end if**

12: **procedure** CHECK(μ, ε, $N_{\max}$, τ) ▷ is $\mu \geq 0$ on the whole W_1 ball?
13: $L, c \leftarrow$ an affine lower bound on μ from one bound-propagation pass, valid on any superset of the ball
14: $\ell \leftarrow c - \varepsilon \, \|B^\top L\|_\infty$ ▷ its exact ball worst case: the closed form (Theorem 2.4)
15: **if** $\ell \geq 0$ **then return** proved ▷ $\mu \geq \ell \geq 0$ on the ball
16: **else if** some ball point x' has $\mu(x') < 0$ **then return** violation ▷ counterexample witness
17: **else if** $N_{\max}$ or τ exhausted **then return** unknown ▷ honest abstention
18: **else**
19: fix one ReLU's phase in μ and recurse on both subnetworks ▷ splitting; converges to $\min_{\text{ball}} \mu$
20: **end if**
21: **end procedure**

Why the check is sound. Every step of CHECK is one-sided. A propagated affine bound is a valid lower bound on μ wherever it is computed (a superset of the ball suffices), and the closed form is the exact worst case of that bound on the relaxed ball (Theorem 2.4), so $\ell \leq \min_{\text{ball}} \mu$; $\ell \geq 0$ therefore proves $g_j(x) \geq A_j \cdot (x - x_0) + b_j$ for every j and every x in the ball, and the returned radius r is sound by Proposition 2.3.

A *violation* exhibits $\mu(x') < 0$—a real counterexample, not a proof failure—and *unknown* certifies nothing. Splitting is complete (Appendix F; at a fully-fixed leaf the exact floor LPs decide), so *unknown* reflects only the budget; in practice the root bound already decides (mean ≤ 1.4 nodes). The learner enters only as the candidate source in line 1 and is trusted for nothing.

Check budget and cost. The check runs with node budget $N_{\max}$=256 and a 60 s time limit (120 s on CIFAR-10), at a per-sample median of $\approx$11 s (ConvLarge, n=40 per budget), dominated by graph construction and the root propagation. Since each ε_{box} trains its own policy, the candidate changes with the budget, so each (sample, budget) is decided independently with no cross-budget reuse. The teacher bisection it replaces costs a median 7.7 s (mean 13.9 s); the fallback fires on 7.5–17.5% of samples (violations + abstentions, Table 8), adding the teacher's cost only there.

Table 7: Per-sample vs. learned envelope at a fixed budget $\varepsilon_{\text{box}} = 0.06$ (ConvLarge, $n = 200$ in-dist), both rows over the theorem-proved subset (n=187; Table 8). "Full-box" is the share of certified radii reaching the box ceiling; every proved learned sample reaches it.

Envelope	mean r	full-box
per-sample α-CROWN (teacher)	0.042	0.701
learned	0.060	1.00

Table 8: Complete (branch-and-bound) validity audit (ConvLarge, MNIST holdout, $n = 40$ per budget). "Provable" is the fraction proven valid over the full W_1 ball; a true violation is a witnessed counterexample (the learned minorant over-claims b_m); abstained is budget exhaustion. The single-pass audit it replaces abstained on up to 97% of samples at large budgets; the complete check resolves nearly all of these into a proof or a witness.

ε_{box}	provable (student)	true violations	abstained	student r
0.03	0.925	3	0	0.0292
0.04	0.900	4	0	0.0390
0.05	0.875	5	0	0.0500
0.06	0.925	3	0	0.0600
0.07	0.825	7	0	0.0700
0.08	0.875	4	1	0.0800
0.09	0.875	4	1	0.0900
0.10	0.825	5	2	0.1000

The complete check changes the picture at large budgets. The single-pass CROWN audit reported a provable rate that collapses to 0 by ε_{box}=0.08—but that collapse was a relaxation artifact, not unverifiability: re-expressing the residual network so CROWN propagates through a per-neuron-structured graph (the same graph the branch-and-bound evaluates) tightens the root lower bound from -266 to $+10.7$ on a representative ε_{box}=0.08 sample, and the branch-and-bound then *proves* validity on 0.825–0.925 of samples across all budgets. The residual failures are not abstentions but genuine violations, and they are concentrated, not diffuse: the same three samples (of 40) recur in the violation set at every budget, where the untuned learner's affine minorant genuinely over-claims b_m.

The complete checker is what exposes this—the box-point audit can sample a violation but can never *prove* validity, and the single-pass CROWN audit conflated these real violations with its own looseness. Operationally the two-tier design absorbs the violations: the per-sample α-CROWN teacher is constructively valid by construction ($40/40$ proved in the in-dist audit at ε_{box}=0.08), so a teacher fallback on the rejected subset restores full provable coverage while the learned envelope still sets the radius on the admitted majority. The same audit on the teacher at $\varepsilon_{\text{box}} = 0.06$ certifies a collapsed radius (mean 0.041), the collapse Figure 3 shows across the budget sweep.

Proof-level admission across architectures (Table 2). The same complete check gates the remaining rows of Table 2 at ε_{box}=0.06 on each architecture's own held-out set (n=40 MNIST, n=50 CIFAR-10, disjoint from its training corpus). On MNIST the admission rate tracks the empirical audit rate: LeNet5 0.85 ($34/40$ proved, 5 witnessed violations, 1 abstention) and DeeperCNN 0.775 ($31/40$, 8, 1), versus ConvLarge 0.925; every MNIST sample claims the full box, so a proof at ε_{box} is a proof at the claimed radius. CIFAR-10 is qualitatively different: the claimed radii are bimodal (21 samples at radius 0, 29 at the full box), and the check proves all 21 radius-0 certificates (vacuous) but *no* nonzero one—of the 29 full-box claims, 10 are witnessed over-claims and 19 exhaust the search budget (120 s per sample). The empirical audit rate told the same story softly (0.70); the complete check makes it exact: on CIFAR-10 the learned envelope's *nonzero* certificates are unproven, and the proved subset certifies only radius 0. Two-tier operation absorbs this: the CIFAR-10 row's reported radius is an audit-admitted estimate, not a theorem-level one. (Implementation note: the checker requires module-form nonlinearities for per-neuron splitting; LeNet5's functional `F.relu` calls were rewritten as `nn.ReLU` modules—numerically identical, no retraining.)

Admitted-subset radius. The admitted-subset radius is the mean certified radius over samples the validity check admits, always reported with its count n (radii are sound by the closed form). On the deeper DeeperCNN, sweeping ε_{box}: at $0.1/0.2/0.5$ the admitted-subset radius tracks the box $(0.10/0.20/0.50)$ over 21–29 admitted samples, but by $\varepsilon_{\text{box}} = 1.0$ it still reads 0.75 over only $n = 4$ and by 2.0 it is 0—so large-radius readings with tiny n are not convergent-certification claims; the usable region is $\varepsilon_{\text{box}} \leq 0.5$.

Future directions. Wasserstein-aware training that penalizes the maximal edge drop of the margin gradients, $\max_{(u,v)\in E} |\nabla_u g - \nabla_v g|$, would move the envelope geometry into the weights. Complete verification is demonstrated on MLPs and small convolutional networks (Appendix F); the threat model is mass-preserving, and unbalanced transport is left to future work.

F Completeness of the verification step

Two domains, two properties. On D_0 the envelope is *sound but loose* (looseness determined by the propagation algorithm); on P—the probability-measure ball $\mathcal{S}_{\text{orig}}$ (Section 2.3)—the concretization is *exact* (Theorem 2.4, Proposition A.2), and soundness is free since any lower bound valid on a superset $D_0 \supseteq P$ bounds P. The default box $[x_0 - \varepsilon, x_0 + \varepsilon]$ is the tightest uniform per-coordinate envelope (a pixel change of δ costs at least δ in transport budget, and ε is attained by moving mass from an adjacent pixel); where it blocks the single-edge plan, the flow-feasibility LP of Proposition A.2 recovers the exact value (Section 3.2).

Completeness in the splitting limit. Fixing every ReLU's phase makes the network affine on that region, so the envelope becomes exact (A is the input gradient) while the concretization stays exact; branch-and-bound over the finitely many phase assignments therefore converges to $\min_{x' \in P} f(x')$, the true Wasserstein worst case, *independently of the envelope domain—D_0* determines only the node count, never the limit value. A looseness placed on the concretization side (minimizing over a set larger than P) would survive the splitting limit.

Anti-contrast: ℓ_p-covering verifiers. An ℓ_p-covering verifier minimizes its affine bound over the *covering set*, a proper superset of the W_1 ball, so even exhaustive ReLU splitting leaves the residual gap $\min_{\text{cover}} - \min_P$: the split closes the envelope but cannot close the concretization domain.

Scope. The statement holds for both MLPs and convolutional networks (the splitting routine `complete_radius` reconstructs each ReLU, convolutional ones included, as a phase-fixable module). Two refinements: (i) the closed form allows any edge flow regardless of flow feasibility, leaving a feasibility residual that splitting cannot remove; the flow-feasibility LP tier (Proposition A.2) closes it, making the certificate exact for the propagated envelope on $P \cap \mathcal{S}_{\text{orig}}$; (ii) at a fully-fixed leaf with non-positive margin the decision is made by solving $K - 1$ exact LPs over the flow polytope (HiGHS (Huangfu & Hall, 2018), $\sim$1–5 ms each), so the B&B predicate itself is exact rather than heuristic.

G W_1 vs. lifted-ℓ_1 (Wegel) radius comparison

Reference-invariance: formal statement and proof.

Proposition G.1 The closed form is the same for every reference distribution.
In the flow coordinates of any reference distribution R (flow δ^R, $R = x_0 + B\delta^R$), the relaxed W_1 ball is the radius-ε ball $\{x_0 + B(\delta - \delta^R) : \|\delta - \delta^R\|_1 \leq \varepsilon\}$; under the bijection $\delta \mapsto x_0 + B\delta$ it is the same set for every δ^R. Hence the extremal value of Theorem 2.4 is independent of R, and reference fine-tuning $\max_R \text{cert}_R$ (Wegel et al., 2021, Sec. 3.2.2) cannot improve an exact concretization.

The close agreement below is an instance of Proposition G.1. Wegel et al. anchor flow coordinates at a reference distribution R: per sample, the map $\Delta(R, \delta) = \mu$ is affine ($x = R + A\delta$, their Theorem 2), the first layer folds as $\tilde{W}^{(1)} = W^{(1)}A$, $\tilde{b}_1 = W^{(1)}R + b_1$, and $W_1(\mu, \nu) = \|\delta^\mu - \delta^\nu_*\|_1$ holds for every R (their Theorem 3).

Our incidence-matrix parametrization is the special case R = a Dirac anchor: both describe the same relaxed ball $\{x_0 + B\phi : \|\phi\|_1 \le \varepsilon\}$, differing only by a fixed translation of flow coordinates. The terminal problem $\min A \cdot (x' - x_0)$ is solved exactly over this set, so its value is coordinate-independent—equal to $-\varepsilon\|B^\top A\|_\infty$.

Corollary G.2 With a complete certifier, Wegel's certificate equals the closed form for every reference.
Vanilla flow-space certification with a *complete* ℓ_1-ball certifier is independent of the reference distribution R and equals the closed form of Theorem 2.4 for every R. The closed form is therefore the exact, reference-free limit of the vanilla method; reference fine-tuning ($\max_R \mathrm{cert}_R$) is a search over envelope generators—an instance of the enclosing-domain freedom of Proposition 2.2, not an improvement over the exact concretization.

Detailed ablations. The covering-gap ablations of Section 3.4 (intermediate-bound and iteration-budget variations move the margin by $< 0.03\%$) leave the main-text conclusions unchanged. On the extended n=100 sample (unpaired, identically selected correct samples) the lift arm keeps median radius 0.177 (bootstrap 95% CI $[0.174, 0.187]$, i.e. an unpaired $1.02\times$ the Wegel arm's n=20 median), and its first 20 indices reproduce the paired run to within the $\sim 10^{-7}$ solver tolerance above. Paired over all 100 samples the lift/Wegel ratio is a median $1.004\times$ (bootstrap 95% CI $[1.003, 1.006]$), $97/100$ at or above parity ($67/100$ within 1%) with a single outlier at $1.47\times$—the two pipelines reproduce each other's certificates to within a fraction of a percent on the median.

Protocol and per-sample data. α-CROWN in the same backend, identical models, and identically selected correctly classified MNIST test samples ($n = 20$ per model for the ratio arms; the Wegel arm was extended to the same $n = 100$ samples as Q2); certified radius in normalized units (ε^*); identical edge set (4-neighbor grid, 1,512 edges) and iteration budget (15 iterations, warm-started). Two refinements of ours are included: (i) the native W1 support function ("Ours"); (ii) a hybrid ("Ours+ℓ_1-box") that propagates the ℓ_1-2ε box through CROWN and takes the deviation from the exact W_1 support—sound because $\mathrm{Ball}_{W_1}(\varepsilon) \subseteq \{\Delta : \|\Delta\|_1 \le 2\varepsilon\}$ (Section 3.4). No arm applies the nonnegativity floor LP; all report the relaxed-ball certificate. Per-sample data for each arm is under `wasserstein/results/`. Headline per-sample ratios on ConvLarge: lift/native median $2.67\times$ [2.60–3.05] (bootstrap 95% CI on the median), wins $20/20$; lift/ℓ_1-box median $2.09\times$ [2.03–3.09], $20/20$; lift/SDP median $1.88\times$ [1.80–1.97], $20/20$ (all $n = 20$).

Implementation note. The lift arm is our own framework instantiated with the flow zonotope: a `LiftLinear` entry layer computes $x = x_0 + Bh$ (folded into the first linear map implicitly), the perturbation object is the stock auto-LiRPA `PerturbationLpNorm(norm=1)` on h, and the original network weights are untouched. The fold is a linear input reparametrization: it applies verbatim when the first layer is a convolution (any stride/padding) and leaves downstream residual, concatenation, and pooling structure untouched. The Wegel arm itself uses explicit per-sample Ling–Okada decomposition followed by first-layer folding ($\tilde{W}^{(1)} = W^{(1)}B$).

Margin decomposition. The flow generators carry the mass-conservation coupling ($|\{e : v \in e\}|$ generators per pixel, divergence-free fields) while the native arm's IBP patch covers intermediate neurons by an $\ell_\infty(\varepsilon)$ box admitting $\Theta(n)$ times more total mass; the ℓ_1-2ε hybrid recovers part of the gap ($1.32\times$ median) and the flow-zonotope lift closes it to the isomorphism limit—the binding factor is the *envelope*, not the terminal support function.

Unit conventions. The flow-ℓ_1 radius ε and the W_1 radius ε coincide (both inclusions hold: any $W_1 \le \varepsilon$ perturbation admits a flow with $\|h\|_1 = W_1$, and conversely $W_1(x_0, x_0 + Bh) \le \|h\|_1$), so no conversion is applied; the MLP$_2$ radii agree to median $\sim 10^{-7}$ relative (max 1.13×10^{-4}), confirming this empirically.