# THE RETENTION GAP: WHY OPTIMAL FEATURES NEED NOT REMAIN STATIONARY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Implicit-bias analyses often explain selection by asking which features training captures. A separate question is whether captured optimal features remain dynamically stable as the residual evolves. This paper isolates that distinction in an explicit two-dimensional regression problem whose minimum representation norm and complete optimal contact geometry are known exactly. An occupied optimal dual contact can lose stationarity as training proceeds; this dynamical failure is termed the *retention gap*. For a compact family of states already resting on the optimal contacts, a computer-assisted proof certifies a uniform transversal first exit with positive charge, and an exact local calculation shows immediate motion into positive dual slack. An intrinsic canonical-measure identity then connects this local event to function selection: at interpolation, excess representation norm equals canonical mass weighted by signed dual slack. Local loss of retention and persistence to the endpoint are therefore distinct obligations. Complementary full-width experiments lie outside the certified entrance family and do not test its event ordering. Instead, inspected trajectories exhibit the same qualitative sequence of approach, margin loss, and departure at the other contact. Across the robustness sweep, training reaches near-machine-precision loss with a positive endpoint excess diagnostic and a consistent off-contact signature, while splitting the exact contact tie changes the observed scale dependence. Overall, these results show that reaching an optimal feature geometry does not by itself make that geometry dynamically stable, identifying retention as a distinct requirement in feature-capture accounts of implicit bias.

## 1 INTRODUCTION

When an overparameterised network interpolates its training data, which interpolating function does optimisation select? One approach identifies a minimum norm representation and studies how training acquires its features. Function space characterisations make the static target precise (Savarese et al., 2019; Ongie et al., 2019; Parhi & Nowak, 2021), while early alignment analyses describe how small initialisation directs neurons toward distinguished features (Maennel et al., 2018; Min et al., 2024; Kumar & Haupt, 2024). Connecting these perspectives raises a further question: once a feature reaches an optimal direction, what keeps it there?

The difficulty is that the field driving a feature continues to change. A network can occupy the contact geometry of a minimum-norm representation before its charges and affine component have fitted the training labels. The remaining residual then drives further motion and evolves as the rest of the network learns. We use *retention gap* specifically for the dynamical failure in which an occupied optimal contact loses stationarity before interpolation. *Persistence* is a separate endpoint question: whether the resulting displacement survives rather than returning or cancelling. Figure 1 separates static target, capture, retention and persistence, and locates our certified and empirical results within this progression.

We study this question in an explicit four-point regression problem whose minimum representation norm and complete optimal contact geometry are known exactly. The construction has two projective contacts, $A$ and $B$, and admits a reduced description while four signed cohorts rest on them. The main dynamical question is which contact loses stationarity first as the cohorts grow and the residual changes. A computer-assisted interval proof resolves this ordering uniformly on a compact



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

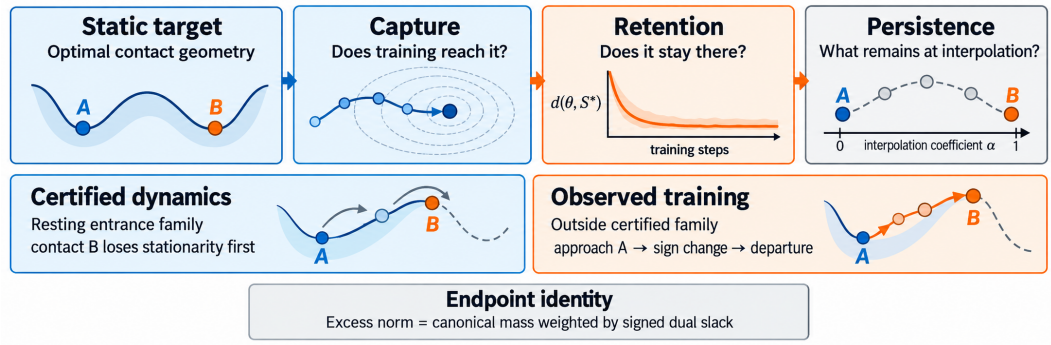



Figure 1: **The retention gap.** Top: static target, capture, retention, and persistence. Bottom left: certified dynamics from a resting entrance family, where contact $B$ loses stationarity first. Bottom right: observed training outside the certified family, where units approach contact $A$, its margin changes sign, and later depart. The endpoint identity links surviving positive dual slack to excess representation norm at interpolation.

entrance family, and an exact local calculation turns the certified stationarity loss into motion away from the optimal geometry. The theorem is conditional on entry into this family; establishing such entry from random initialisation is a separate question.

The endpoint consequence is governed by a canonical measure identity: at interpolation, excess representation norm equals canonical mass integrated against signed dual slack. Canonicalisation removes cancellation and representation ambiguity, making the identity intrinsic to the learned function. The two results play distinct roles: the exit theorem explains how displacement from the optimal geometry begins, while the identity determines when displacement that survives to interpolation yields a norm gap.

The experiments use the same four-point construction but occupy a different region of state space from the certified entrance family. The inspected trajectories therefore do not validate the theorem's event ordering: they exhibit approach, margin loss and departure at contact $A$, whereas the theorem certifies the first exit at $B$ under its stated entrance assumptions. Their role is complementary to test whether the same qualitative retention sequence occurs in full-width training and to examine what remains at interpolation. Matched-seed perturbations test sensitivity to the exact contact tie, while one-dimensional controls check the endpoint diagnostic against known minimum-norm solutions. Our contributions are:

1. **An exact link between optimal features and representation cost.** For the regression construction, we determine the minimum representation norm and the complete set of optimal contacts. We then derive an identity that explains how signed feature mass contributes to excess norm at interpolation, accounting for cancellation and equivalent representations (Section 2 and Appendix B).

2. **A proof that occupied optimal features can lose stationarity.** Starting from a specified compact family of states already resting on the optimal contacts, we prove that the same contact loses stationarity first throughout the family. The crossing is transversal and carries positive charge, which immediately moves into positive dual slack. This establishes a failure of retention in the exact reduced dynamics (Section 3).

3. **A geometric study of what training retains.** We measure approach to a contact, changes in its stationarity margin and subsequent departure, and examine the feature mass remaining at interpolation. Matched seed perturbations show how breaking the tie between optimal contacts changes the observed behaviour, while one dimensional controls test the diagnostic against known minimum norm solutions (Section 4).

# 2 THE RETENTION GAP AS AN INVARIANCE PROBLEM

## 2.1 CONTACT GEOMETRY AND THE REPRESENTATION NORM

Our question is conditional on interpolation: when gradient flow reaches an interpolating function, does it also attain the minimum representation norm among interpolants? The difficulty is that training and norm comparison naturally describe a hidden unit in different ways. Training follows individual balanced neurons, whereas the representation norm is a property of the function they collectively represent. We first connect these two descriptions and then identify exactly what creates an excess norm.

We study the two-layer ReLU network

$$h_\theta(x) = \sum_{j=1}^{m} a_j \left(\langle w_j, x\rangle + b_j\right)_+ + \langle v, x\rangle + c.$$

For training pairs $(x_i, y_i)$, the squared loss is $L(\theta) = \frac{1}{2n}\sum_i (h_\theta(x_i) - y_i)^2$, and training follows Clarke gradient flow $\dot\theta \in -\partial L(\theta)$ in time $\tau$. The affine term is trainable and unregularised. Hidden units are exactly balanced at initialisation, $|a_j| = \|(w_j, b_j)\|$, and balancedness is preserved by the flow.

For a hidden unit, write $u_j = (w_j, b_j)$ and $\rho_j = \|u_j\| = |a_j|$. Its direction is $\hat u_j = u_j/\rho_j \in \mathbb{S}^d$, while $s_j = \operatorname{sign}(a_j)$ records its conserved sign. We also write $p_j = a_j u_j$, so $\|p_j\| = \rho_j^2$. Thus $\hat u_j$ describes where the unit points, while $\|p_j\|$ measures how much dynamic charge it carries. These are the natural coordinates for the gradient-flow dynamics.

The representation norm sees the same unit differently. It charges $|a_j|\,\|w_j\|$ while leaving the hidden bias free. Its natural normalisation therefore fixes $\|w\|$, rather than $\|(w, b)\|$, and we work on the cylinder $\mathcal{Z} = \{(w, b) : \|w\|_2 = 1\}$. The map $\pi(w, b) = (w, b)/\|w\|$ connects the two descriptions. A direction $\hat u_j$ with $\hat w_j \neq 0$ contributes canonical coefficient $c_j = a_j\|w_j\| = s_j\|p_j\|\,\|\hat w_j\|$ at $\pi(\hat u_j)$. If $w_j = 0$, the unit is constant and is absorbed into the free affine term.

This second description is necessary because raw neurons are not intrinsic. Several neurons may represent the same ReLU kink, and reversing its orientation changes the function only by a linear term: $c\,(-t)_+ = c\,(t)_+ - ct$. Since linear terms are free here, what matters for the norm is the net signed coefficient carried by each hyperplane, not the particular collection of neurons used to represent it.

Let $\mathcal{F}$ be the class of finite-width networks of this form with $\|w_k\| = 1$ and a free affine part, and let $\mathcal{R}(f)$ be the infimum of $\sum_k |c_k|$ over all such representations. We collect coefficients on the same hyperplane into a canonical signed measure $\alpha_f$ on $\mathcal{Z}$ modulo $z \sim -z$. Antipodal representatives add with the same sign, with the resulting linear term absorbed by the affine component. After merging, the representation by pairwise distinct hyperplanes is unique, and merging can only decrease the total absolute coefficient. Hence $\mathcal{R}(f) = |\alpha_f|(\mathcal{Z})$; Appendix B gives the formal argument.

We can now express norm selection geometrically. Let $\hat g$ be an optimal dual feature satisfying $|\hat g(z)| \le 1$. Its *contacts* are the points where $|\hat g(z)| = 1$. Writing $s_f = d\alpha_f/d|\alpha_f|$ for the sign of the canonical mass, the quantity $1 - s_f(z)\hat g(z)$ is its dual slack. Correctly signed mass on a contact has zero slack and therefore contributes no excess over the dual optimum. Mass away from the contact set, or carrying the wrong sign, has positive slack.

This observation gives an exact decomposition of the norm gap.

**Lemma 1** (Intrinsic excess identity)**.** *If $f \in \mathcal{F}$ interpolates the data exactly, then for any optimal dual feature $\hat g$,*

$$\mathcal{R}(f) - \mathcal{R}_{\min} = \int_{\mathcal{Z}} \big(1 - s_f(z)\hat g(z)\big)\, d|\alpha_f|(z).$$

The identity says that the representation-norm gap is exactly canonical mass weighted by its dual slack. Correctly signed mass on a contact contributes zero; mass with positive slack contributes positively. Indeed, affine annihilation and dual attainment give $\int_{\mathcal{Z}} \hat g\, d\alpha_f = \mathcal{R}_{\min}$. Since $d\alpha_f = s_f\, d|\alpha_f|$ and $\mathcal{R}(f) = \int_{\mathcal{Z}} d|\alpha_f|$, subtracting the two expressions gives the identity. Appendix B gives the full proof, including the canonicalisation argument.

For a numerically trained network, a nonzero residual $r = f(X) - y$ adds the correction $\langle \hat{\lambda}, r \rangle$. Its absolute value is at most $8.6 \times 10^{-14}$ on all saved two-dimensional endpoints in Section 4, each with its own dual certificate, and at most $1.8 \times 10^{-13}$ on the four half-step reruns. The same identity gives the endpoint condition used later.

**Corollary 2** (Persistence criterion)**.** *If an interpolating $f$ has a measurable set $E \subset \mathcal{Z}$ carrying canonical mass at least $c > 0$, and its dual slack is at least $\sigma > 0$ for $|\alpha_f|$-almost every point in $E$, then $\mathcal{R}(f) - \mathcal{R}_{\min} \geq c\sigma$.*

In other words, a strict norm gap remains whenever a nonzero amount of canonical mass is still in positive dual slack at interpolation. This reduces the endpoint question to a concrete geometric one: whether mass that leaves the correctly signed contact set remains away from it until interpolation. The certified theorem below establishes the preceding dynamical event, a strict escape from the zero-slack contact geometry.

### 2.2 Why a recruited contact can become unstable

The endpoint criterion tells us what must persist. We next ask why a unit recruited at a contact need not remain there. The key is that the condition holding a direction in place depends on the current residual, and that residual changes during training.

Write $\phi_i = (x_i, 1)$, $r_i = h(x_i) - y_i$, and $\mu = -r/n$. For a direction $u$, define $g_\mu(u) = \sum_i \mu_i \left(\langle u, \phi_i \rangle\right)_+$. This quantity measures the correlation between the activation pattern generated by $u$ and the current negative residual field.

On each activation cell, balanced gradient flow separates into a change in charge and a change in direction:

$$\frac{d}{d\tau} \log \|p_j\| = 2 s_j g_\mu(\hat{u}_j), \qquad \frac{d\hat{u}_j}{d\tau} = s_j (I - \hat{u}_j \hat{u}_j^\top) \nabla g_\mu(\hat{u}_j).$$

The interpretation is direct. The first equation determines whether the unit grows: its charge increases when $s_j g_\mu(\hat{u}_j) > 0$. The second moves its direction by projected ascent of the same quantity on the sphere. The factor 2 comes from $\|p_j\| = \rho_j^2$.

The function $g_\mu$ is positively one-homogeneous and piecewise linear, with kinks where $\langle u, \phi_i \rangle = 0$. A unit can remain at a direction only while that direction is Clarke stationary for $s_j g_\mu$ on $\mathbb{S}^d$. At a kink, this means that zero lies in the convex hull of the one-sided tangent fields. The coefficients of this convex combination are the edge coordinates $\theta$ used in Section 3.

This is the source of the instability. A direction may be attractive and stationary for an early residual field. As other hidden units acquire charge, the residual changes, and the same direction must satisfy the stationarity condition again. If it no longer does, the unit can leave the contact. The early regime is particularly simple.

**Lemma 3** (Early aligned residual)**.** *Let $M = [\mathbf{1}, X] \in \mathbb{R}^{n \times (d+1)}$ have rank $d + 1$ and $n = d + 2$. Then $\ker M^\top$ is spanned by a single vector $\lambda^*$, and the residual of the affine least-squares fit lies on $\mathbb{R}\lambda^*$. If $\langle \lambda^*, y \rangle > 0$, the corresponding residual field $\mu$ is a positive multiple of $\lambda^*$.*

The reason is that the least-squares residual is orthogonal to the column space of $M$. While the hidden units are negligible and the affine part has relaxed, the residual field is therefore aligned with $\lambda^*$. Later, as the hidden units alter the fit, the residual leaves that ray. The holding condition at an early contact must therefore be checked again against the current residual rather than inherited from the early phase.

## 3 A certified failure of retention

### 3.1 The exact construction and its marginal contacts

Consider $X_{\text{old}} = \{(1, 0), (0, 1), (-1, \frac{1}{2}), (\frac{1}{2}, -1)\}$ with $y = (1, 0, 2, 1)$, where the points are ordered as $x_1, \ldots, x_4$ and indices are one based throughout. The inputs are in general position with $n = d + 2$. Both the affine null direction and the dual optimum are available in closed form, giving $\mathcal{R}_{\min} = 5\sqrt{17}/9$.

Modulo the canonical antipodal equivalence, the optimal dual geometry has exactly two projective contacts: $u_A = (1, 4, -1)/\sqrt{18}$, incident to $x_1, x_3$ with correct sign $-$, and $u_B = (4, 1, -1)/\sqrt{18}$, incident to $x_2, x_4$ with correct sign $+$; see Appendix B. They are tied in the exact dual landscape. This tie is deliberate: under the aligned residual field, both contacts are only marginally stationary.

That marginality can be seen without interval arithmetic. At $u_B$ the active point is $x_1$ and the incident points are $x_2, x_4$; the one sided tangent fields form a parallelogram, and zero lies in it with edge coordinates

$$\theta_2^B(\mu) = -\frac{\mu_1}{2\mu_2}, \qquad \theta_4^B(\mu) = -\frac{2\mu_1}{3\mu_4}.$$

The resting contact is interior to the stationary region when both lie in $(0, 1)$, lies on its boundary when either equals 0 or 1, and is inadmissible when either lies outside $[0, 1]$. For the aligned field $\hat{\lambda} \propto (3, -3, 2, -2)$, $(\theta_2^B, \theta_4^B) = (\frac{1}{2}, 1)$. Under the reflection that swaps the two input coordinates, which maps $\hat{\lambda}$ to $-\hat{\lambda}$, contact $A$ has $\theta_1^A = -\mu_2/(2\mu_1)$ and $\theta_3^A = -2\mu_2/(3\mu_3)$, with the same values $(\frac{1}{2}, 1)$. Both contacts therefore sit exactly on a stationarity face. Which of them crosses first depends on how the residual evolves; that is what the theorem resolves.

The theorem begins from a compact entrance family in an exact four-cohort resting system. A *cohort* $C \in \{A, -A, B, -B\}$ consists of units resting at $\pm u_A$ or $\pm u_B$ with the correct sign; its charge is $Q_C = \sum_{j \in C} \|p_j\|$, and $S = \sum_C Q_C$. While the four cohorts remain at these contacts, the balanced gradient flow closes exactly on the eight-dimensional state $(Q, \mu) \in \mathbb{R}^4_+ \times \mathbb{R}^4$:

$$\dot{Q}_C = 2 s_C g_\mu(\hat{u}_C) Q_C, \qquad \dot{\mu} = -\tfrac{1}{4}\big(\Phi\Phi^\top + D(Q)\big)\mu,$$

where $\Phi$ has rows $\phi_i$ and $D(Q) = \mathrm{diag}(Q_B, Q_A, \frac{9}{4}Q_{-B}, \frac{9}{4}Q_{-A})$. Thus *exact reduced dynamics* means the exact restriction to this stated resting four-cohort system, not an approximation to the full-width training runs; the complete coordinate form is in Appendix D.

On the entrance section $S = 10^{-4}$ we use the ratios $r_A = Q_A/Q_{-B}$, $r_{-A} = Q_{-A}/Q_{-B}$ and $r_B = Q_B/Q_{-B}$. The certified domain $\mathcal{B}_0 = [59/10, 473/80] \times [19/20, 21/20]^2$ is the first of 24 pre-specified $r_A$ slabs of $\mathcal{B} = [59/10, 31/5] \times [19/20, 21/20]^2$, with its boundaries fixed before the exit margins were inspected. It is an $A$-heavy certification domain ($r_A \approx 5.9$ while the other two ratios are near one), not an estimate of states reached from random initialisation. Floating-point integrations from 18 states per slab give the same first event in all 24 slabs; the uniform guarantee, however, is only on $\mathcal{B}_0$. The remaining entrance conditions are stated in Appendix D.

### 3.2 The first exit is uniform over the whole entrance family

Let $\ell_B = -(2\mu_1 + 3\mu_4)$ be the stationarity margin of contact $B$: $\ell_B > 0$ exactly when the resting contact is admissible, and $\ell_B = 0$ is its boundary face. In the exact reduced system, $\dot{\ell}_B = F_B - c_B \ell_B$ with $c_B > 0$; hence on the face the sign of $F_B = \dot{\ell}_B|_{\ell_B = 0}$ decides whether the trajectory crosses out of the resting regime. The explicit expressions are given in Appendix E.

The proof follows the event rather than a predetermined final time. Before the event, interval propagation controls all four charges and the margins that keep the other contacts admissible. Near $\ell_B = 0$, an event enclosure proves that this face is crossed with $F_B < 0$ while the competing quantities remain positive. The local calculation in Section 3.3 then converts this crossing into departure from the contact. Correlations between the residual and the charges are essential: replacing the reachable state by a single coordinatewise box can obscure the signs needed to distinguish competing events. Appendix D gives the full reduced system, entrance enclosure, monitored margins and coverage checks.

**Theorem 4** (Uniform failure of contact retention)**.** Assumptions. *Consider the exact four-cohort reduced dynamics above on $X_{\mathrm{old}}$, started from an entrance state: all four cohorts rest at their contacts with the correct signs; $S = 10^{-4}$; $(r_A, r_{-A}, r_B) \in \mathcal{B}_0$; and $\mu = \mu_\star + LQ + R$ with $\|R\|_2 \le 0.262\, S^2$ and $\lambda^{*\top} R = 0$, where $\mu_\star \in \mathbb{R}^4$ and $L \in \mathbb{R}^{4\times 4}$ are explicit (Appendix D). Equivalently, the conserved quantity $H = \lambda^{*\top}\mu + \frac{3}{4\sqrt{2}} S$ equals $5/4$, as it does for every realisable state with this resting configuration. On this family $\ell_A > 0$ and $\ell_B > 0$ are certified.* Conclusion. *For every such state: (i) the stationarity margin $\ell_B$ reaches zero in finite physical time; (ii) before that event none of the four charges and none of the four secondary Filippov margins vanishes, and*

| Certified quantity | Uniform consequence on $\mathcal{B}_0$ |
|---|---|
| Coverage | 4608 terminal leaves, each elementary cell covered once |
| $B$ charge at exit | $Q_B > 8.0 \times 10^{-2}$ |
| Transverse event margin | $-F_B > 4.8 \times 10^{-5}$ |
| Nearest monitored competing quantity at exit | $> 3.9 \times 10^{-3}$ |
| Physical event time | $T < 56.7$ |

Table 1: **Certified first exit bounds.** Rounded uniform bounds on $\mathcal{B}_0$, as recorded by the original theorem run; exact values, definitions, provenance and validation details are in Appendix D.

*$\ell_A$ can reach zero only on a face where the field points strictly into $\ell_A > 0$, so $B$ is the unique first cohort to lose stationarity; (iii) at the event $Q_B > 0$, the crossing is transversal ($\dot{\ell}_B = F_B < 0$), and all nine monitored competing quantities of Appendix D are strictly positive.*

The two remaining Filippov coefficients of the antipodal cohorts vanish identically, so they can never cross a face; Appendix D shows that resting is nevertheless the unique continuation for those cohorts. Entry into the family from initialisation (Appendix C) and the dynamics after the first event lie outside the statement. Table 1 gives the uniform bounds behind (i)–(iii).

Adaptive subdivision covers $\mathcal{B}_0$ by 4608 terminal leaves, all PASS, with exact rational verification that every elementary cell is covered once. The eight-dimensional validator uses outward-rounded interval arithmetic, correlated zonotopes and a validated fourth-order Taylor/Picard enclosure; an event-aware terminal step isolates the unique zero of $\ell_B$. The frozen source was separately audited for interval rounding, structural identities, face barriers and deliberately failing negative controls. Appendix D gives the monitored margins, event construction and exact certified constants.

The certificate combines three ingredients: an exactly identified target, uniform event ordering over the entrance family, and positive charge at a transversal stationarity loss. The local calculation below converts that event into motion away from the optimal geometry.

### 3.3 The exit has an immediate function space consequence

On the one sided post event chart, the $B$ cohort slides along the kink $\langle u, \phi_2 \rangle = 0$ toward the side where $x_4$ becomes active. With $e = (1, -2, 2)/3$ the unit tangent of that kink at $u_B$ pointing to this side, its tangential velocity is $v_B = \langle \mu_1 \phi_1 + \mu_4 \phi_4, e \rangle = \mu_1 + \frac{3}{2}\mu_4 = -\frac{1}{2}\ell_B$ (Appendix E), so at the event

$$v_B = 0, \qquad \dot{v}_B = -\tfrac{1}{2} F_B.$$

The certified sign in Theorem 4 therefore gives positive tangential acceleration into the positive slack side.

**Corollary 5** (Immediate escape from contact)**.** *For every trajectory in Theorem 4, the $B$ cohort moves into positive dual slack immediately after the certified loss of stationarity.*

Corollary 5 supplies the local escape from the optimal geometry. At an interpolating endpoint, surviving canonical mass at positive signed dual slack contributes a positive amount to the norm gap by Lemma 1. Corollary 2 specifies the surviving mass condition needed to pass from local displacement to an endpoint norm gap.

## 4 Retention along training runs

The theorem and the training runs concern different regions of state space. Theorem 4 conditions on $\mathcal{B}_0$ and certifies the first loss of stationarity at $B$; the full-width runs below do not satisfy those entrance conditions and are not a numerical validation of that event ordering. In particular, we do not use the static reflection between $A$ and $B$ to identify the observed full-width event at $A$ with the certified event at $B$. Instead, the experiments ask whether the same qualitative sequence of approach, margin loss, and departure occurs outside the certified family, and what representation remains at interpolation. Appendix A specifies the fixed training protocol.

### 4.1 EIGHT INSPECTED RUNS: ENTRANCE CONDITIONS AND EVENTS AT CONTACT $A$

We tracked eight canonical runs at $\lambda = 10^{-6}$ with $m \in \{50, 200, 400\}$ (configurations in Appendix A), using a numerical cohort assignment (units within $0.05$ rad of a contact with the correct sign; Appendix A). When the cohort charge first reaches $S = 10^{-4}$, more than $99.99\%$ of the charge sits in the four cohorts. The charge ratios, however, are $r_A \in [0.18, 0.65]$, $r_{-A} \in [0.007, 0.063]$ and $r_B \in [0.14, 0.56]$, far from $\mathcal{B}$. The residual remainder is $\|R\|_2/S^2 \in [0.59, 1.40]$, above the certified $0.262$, and $\ell_A < 0 < \ell_B$. The training runs therefore occupy a different region of state space from the entrance family of Theorem 4.

In this region the observed contact event occurs at $A$. Membership is recomputed at every log: the $A$ units are those of negative sign within $0.05$ rad of $u_A$, so the diagnostics describe the currently selected units, not a fixed set of neurons. We record three events (Appendix A). *Approach:* the $A$ units are non-empty, their charge weighted mean angle to $u_A$ is below $10^{-3}$ rad, and the residual field is aligned with $\lambda^*$ ($|\cos| > 0.999$); $\ell_A > 0$ at that log in every run. *Margin change:* $\ell_A$ then changes sign once in the logged record, at $S$ between $1.5 \times 10^{-6}$ and $5.6 \times 10^{-6}$, while the mean angle is still at most $7.2 \times 10^{-4}$ rad, and it stays negative until departure. *Departure:* the mean angle first exceeds $10^{-2}$ rad $4{,}570$ to $5{,}500$ steps after the margin change, after the runs have passed $S = 10^{-4}$. The negative sign is the one the relaxed residual predicts: at the charge ratios observed on the section $S = 10^{-4}$, the first order unstable graph of Appendix D gives $\ell_A/Q_{-B} \in [-0.163, -0.133]$ because $Q_{-B}$ dominates. The $B$ units meet the approach criterion only after $\ell_A$ has changed sign, and $\ell_B$ stays positive from then until the residual norm falls below $10^{-8}$. The observation is therefore an empirical sequence of approach, margin change and later departure at contact $A$; the certified reduced event at $B$ is a separate statement of Theorem 4.

### 4.2 WHAT REMAINS AT INTERPOLATION

The two local escape charts give distinct geometric signatures (Appendix E). On $A$'s chart, negative charge stays incident to $x_1$ as $x_3$ becomes active; on $B$'s chart, positive charge stays incident to $x_2$. At each trained endpoint we canonicalise and merge the finite-width atoms as in Section 2 and report $\widehat{\Delta}(f) = |\widehat{\alpha}_f|(\mathcal{Z}) - \mathcal{R}_{\min}$, the grouped canonical charge above the exact minimum. Because the grouping uses numerical tolerances, $\widehat{\Delta}$ is an endpoint diagnostic rather than the intrinsic norm itself. The residual correction from Lemma 1 is not folded into $\widehat{\Delta}$ and is below $8.6 \times 10^{-14}$ on these endpoints; Appendix A gives the extraction tolerances.

Across all 60 robustness runs ($m = 200$ with 20 seeds and $m = 50$ with 10 seeds, each at $\lambda \in \{10^{-4}, 10^{-6}\}$), training reaches squared loss below $1.9 \times 10^{-28}$ and $\widehat{\Delta} > 0$, with minimum $0.01597$. Wrong-sign contact charge is zero under the extraction thresholds. This does not conflict with the dominant negative charge: the latter is off contact, where its signed dual slack is positive. In every run this dominant off-contact family lies on a hyperplane through $x_1$ and no other training point; its distance to $x_1$ is below $10^{-16}$, while the next point is at least $0.17$ away. The same signature appears in all 20 additional $m = 400$ runs.

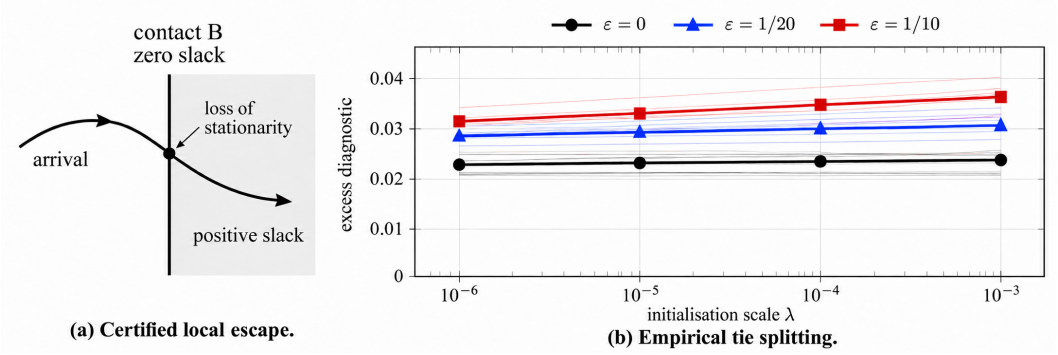
This conclusion is unchanged by tenfold variations of the grouping, contact, and incidence tolerances (diagnostic change below $10^{-15}$), and by four targeted half-step reruns (maximum change $1.02 \times 10^{-5}$); see Appendix A. Table 2 summarises the main evidence.

### 4.3 STRUCTURAL CONTROLS: BREAKING THE EXACT TIE

We replace the fourth point by $x_4 = (\frac{1}{2}, -1 + \varepsilon)$ and recompute the static dual geometry exactly for each $\varepsilon$, with an explicit primal interpolant attaining each dual value. At $\varepsilon = 0$ the two optimal contacts are tied; for $\varepsilon \in \{1/20, 1/10\}$ the tie splits and a single $B$ type contact remains. At width $m = 50$ we trained four seeds, fixed before inspecting the results, at $\lambda \in \{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$; the same seeds, and hence the same initial directions and signs, are used in every condition (Figure 2b). At $\varepsilon = 0$ the endpoint excess is flat in every seed, changing by at most $7.8 \times 10^{-4}$ in relative terms over $\lambda \in \{10^{-4}, 10^{-5}, 10^{-6}\}$. For both positive perturbations it decreases as $\lambda$ decreases in every seed: from $\lambda = 10^{-3}$ to $10^{-6}$ the median falls by $0.0015$ at $\varepsilon = 1/20$ and by $0.0043$ at $\varepsilon = 1/10$, against $0.0002$ at $\varepsilon = 0$. The endpoint signature also changes with the geometry: for $\varepsilon > 0$ the dominant off contact family carries positive charge through $x_2$ in all 32 runs. The paired

| Check | Result |
|---|---|
| Certified-family entrance | Not satisfied in any of the 8 tracked runs. |
| Contact $A$ | Approaches with $\ell_A > 0$, changes sign before $S = 10^{-4}$, and later departs. |
| Contact $B$ | $\ell_B > 0$ from approach until $\|r\| < 10^{-8}$ in the 8 tracked runs. |
| Endpoint excess | Positive in 60/60 runs; minimum 0.01597. Wrong-sign contact charge is zero. |
| Off-contact geometry | Negative charge through $x_1$ only in 60/60 runs and 20/20 additional $m = 400$ runs. |

Table 2: **Endpoint evidence under the tested training protocol.**




**(a) Certified local escape.** **(b) Empirical tie splitting.**

Figure 2: **Escape and tie splitting.** Left: certified local escape from contact $B$ into positive dual slack. Right: endpoint excess under tie perturbations across initialisation scales; thin lines show individual seeds and thick lines their medians.

changes in scale dependence and endpoint geometry show sensitivity to the exact tie across all four seeds. These finite scale measurements leave the asymptotic limit open.

The one dimensional controls check the diagnostic in a setting with a known norm optimum. In one dimension the representation norm reduces to the total variation of the derivative, and a minimum norm interpolant is given by linear spline interpolation (Savarese et al., 2019; Hanin, 2021). For the architecture with a free affine part used throughout, every interpolating control run on two one dimensional datasets ends within $3.1 \times 10^{-3}$ of that minimum, and the measured gap is consistent with $m\lambda^2$ scaling, the cost of dormant units at fixed width (Appendix F). These controls test the extraction pipeline and endpoint diagnostic against a known minimum: the extracted solutions remain close to the known optimum, rather than exhibiting a systematic positive excess from the diagnostic itself. They are not evidence for the two-dimensional retention dynamics. Appendix F also reports the output bias variant, whose different behaviour makes the role of the free affine part explicit.

## 5 Implications for implicit bias arguments

The construction distinguishes three ways a proof based on feature capture can fail: the trajectory may not reach the target geometry, an occupied contact may lose stationarity, or a local displacement may disappear before interpolation. The certified family isolates the second case. Its contacts are optimal and already occupied, yet the residual changes enough to force an exit. An argument that infers an endpoint support from early capture must therefore establish retention or account for subsequent departure and return.

The endpoint statement must also be intrinsic. Neurons can cancel or exchange antipodal representatives without changing the function. The canonical measure removes these ambiguities, and Lemma 1 then makes the remaining criterion exact: an interpolant has excess norm precisely when its canonical measure carries positive signed dual slack. This separates the dynamical question of where features move from the variational question of what function they ultimately represent.

## 6 Related work

**Static geometry.** Representation costs for two layer ReLU networks were characterised in one dimension by Savarese et al. (2019) and in multiple dimensions by Ongie et al. (2019). The representer results of Parhi & Nowak (2021) and the one dimensional description of Hanin (2021) further clarify the variational problem. We use this viewpoint to identify the exact optimal contact set and then ask whether occupied contacts remain stationary.

**Training dynamics.** Lazy and feature learning regimes are distinguished by Chizat et al. (2019) and Woodworth et al. (2020); Chizat & Bach (2018) study mean field convergence. Minimum variation norm selection holds in special settings such as orthogonal inputs (Boursier et al., 2022). Early alignment analyses (Maennel et al., 2018; Min et al., 2024; Kumar & Haupt, 2024; Kunin et al., 2026) explain movement toward distinguished directions. Our event occurs after the contacts are occupied and is driven by the subsequent residual evolution. Related analyses of nonsmooth gradient dynamics include Ji & Telgarsky (2020), Lyu & Li (2019) and Lee et al. (2022).

**Negative selection results.** Early alignment can lead to spurious stationary points (Boursier & Flammarion, 2025), and an optimisation threshold can separate interpolation from simpler non-interpolating solutions (Boursier & Flammarion, 2024). Gradient flow need not minimise rank in ReLU networks (Timor et al., 2023); in homogeneous classification, its limiting KKT point need not be a local optimum of the margin problem (Vardi et al., 2022). Our result isolates loss of stationarity at an already occupied optimal contact before interpolation. The uniform event certificate identifies when this loss occurs, and the canonical excess identity specifies its consequence when displaced mass survives to interpolation.

## 7 Conclusion

Reaching an optimal feature geometry does not guarantee that training will retain it. We make this distinction precise in an explicit regression construction whose minimum representation norm and complete optimal contact geometry are known exactly. For a compact family of states already resting on those contacts, we certify a uniform first loss of stationarity in the reduced gradient flow and show that the resulting motion enters positive dual slack. The canonical excess identity then connects this local event to function selection: at interpolation, excess representation norm is exactly canonical mass weighted by signed dual slack.

Training experiments on the same construction probe the corresponding geometry outside the certified entrance family. They exhibit the same qualitative sequence, approach, loss of the stationarity margin, and departure, at the other contact, while endpoint extraction reveals a consistent off-contact signature. Breaking the exact contact tie changes both this geometry and its dependence on initialisation scale. Overall, these results separate three questions that are often conflated in accounts of implicit bias: whether training finds an optimal feature geometry, whether that geometry remains dynamically stable, and whether any displacement persists to interpolation. Retention is therefore a distinct part of the selection problem, and stationarity margins and canonical dual slack provide concrete quantities for studying it.

### AI use statement

Generative AI tools were used primarily to assist manuscript writing and language polishing, with additional basic checks of references and consistency between the text and reported results. The authors reviewed the AI-assisted revisions and take responsibility for the final manuscript.

### Reproducibility statement

The accompanying anonymised supplementary material provides the code, experimental records and verification materials supporting our results. Further methodological details are given in the appendices. Upon acceptance, we will make the code, data and results publicly available on GitHub, together with instructions for reproduction.

## REFERENCES


Etienne Boursier and Nicolas Flammarion. Simplicity bias and optimization threshold in two-layer relu networks. *arXiv preprint arXiv:2410.02348*, 2024.

Etienne Boursier and Nicolas Flammarion. Early alignment in two-layer networks training is a two-edged sword. *Journal of Machine Learning Research*, 26(183):1–75, 2025.

Etienne Boursier, Loucas Pillaud-Vivien, and Nicolas Flammarion. Gradient flow dynamics of shallow relu networks for square loss and orthogonal inputs. *Advances in Neural Information Processing Systems*, 35:20105–20118, 2022.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in neural information processing systems*, 32, 2019.

Boris Hanin. Ridgeless interpolation with shallow relu networks in $1d$ is nearest neighbor curvature extrapolation and provably generalizes on lipschitz functions. *arXiv preprint arXiv:2109.12960*, 2021.

Ziwei Ji and Matus Telgarsky. Directional convergence and alignment in deep learning. *Advances in Neural Information Processing Systems*, 33:17176–17186, 2020.

Akshay Kumar and Jarvis Haupt. Directional convergence near small initializations and saddles in two-homogeneous neural networks. *arXiv preprint arXiv:2402.09226*, 2024.

Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James Simon, Michael Deweese, Surya Ganguli, and Nina Miolane. Alternating gradient flows: A theory of feature learning in two-layer neural networks. *Advances in Neural Information Processing Systems*, 38:4377–4424, 2026.

Sangmin Lee, Byeongsu Sim, and Jong Chul Ye. Support vectors and gradient dynamics of single-neuron relu networks. *arXiv preprint arXiv:2202.05510*, 2022.

Kaifeng Lyu and Jian Li. Gradient descent maximizes the margin of homogeneous neural networks. *arXiv preprint arXiv:1906.05890*, 2019.

Hartmut Maennel, Olivier Bousquet, and Sylvain Gelly. Gradient descent quantizes relu network features. *arXiv preprint arXiv:1803.08367*, 2018.

Hancheng Min, Enrique Mallada, and Rene Vidal. Early neuron alignment in two-layer relu networks with small initialization. In *International Conference on Learning Representations*, volume 2024, pp. 3431–3473, 2024.

Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.

Rahul Parhi and Robert D Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021.

Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pp. 2667–2690. PMLR, 2019.

Nadav Timor, Gal Vardi, and Ohad Shamir. Implicit regularization towards rank minimization in relu networks. In *International Conference on Algorithmic Learning Theory*, pp. 1429–1459. PMLR, 2023.

Gal Vardi, Ohad Shamir, and Nati Srebro. On margin maximization in linear and relu networks. *Advances in Neural Information Processing Systems*, 35:37024–37036, 2022.

Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

# A NUMERICAL PROTOCOL

All $X_{\text{old}}$ runs use the four point dataset of Section 3 and the model of Section 2. Initialisation is exactly balanced, $|a_j| = \|(w_j, b_j)\| = \lambda$, with directions i.i.d. uniform on $\mathbb{S}^2$, signs i.i.d. uniform, $v = 0$ and $c = 0$, and seed map `default_rng(1000003*seed+17)`, so initial directions and signs depend on the seed only. The integrator is explicit gradient descent on all parameters, including the affine part, with step $\eta_k = \min\{\eta_{\text{rel}}(\|\theta_k\| + 10^{-12})/(\|\nabla L(\theta_k)\| + 10^{-300}), \eta_{\max}\}$, $\eta_{\text{rel}} = 2\times10^{-3}$ and $\eta_{\max} = 10^{-2}$, for a fixed horizon of $9\times10^5$ steps. This explicit gradient descent with an adaptive step approximates gradient flow; finite steps introduce discretisation error, whose effect on the endpoint diagnostics is assessed by the step size check below. The ReLU derivative at zero is taken to be 0. Exact balancedness and the reduced flow identities belong to the continuous theory and hold only approximately along the discrete iterates.

**Extraction.** Units with $\|w_j\| \le 10^{-14}$ are discarded. Keys $(w, b)/\|w\|$ are canonicalised modulo $z \sim -z$ and grouped when all coordinates agree within $10^{-5}$, and groups with |net charge| $< 10^{-13}$ are dropped. A group is on a contact if its key is within $2 \times 10^{-3}$ of a contact key. An off contact family is significant if its charge exceeds $10^{-3}\mathcal{R}(f)$, and a training point is incident if $|\langle w, x_i\rangle + b| < 10^{-6}$. We call the resulting excess over $\mathcal{R}_{\min}$ the *dual slack excess diagnostic*. Because grouping tolerances enter, it is a diagnostic rather than the intrinsic norm itself.

**Scale sweep (30 runs).** Widths $m \in \{50, 200, 400\}$, $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ and seeds $0, 1$, with the protocol above. Every sweep, robustness, perturbation and step size run reported in this appendix was executed with its final parameters saved, and all reported values are recomputed from those endpoints. Ten sweep runs are shared with the robustness suite or the $m = 400$ cell below. All 30 runs finish with squared training loss below $1.4\times10^{-28}$, and the diagnostic is positive in all of them. At $\lambda = 10^{-6}$ the two seed means are 0.01686, 0.02621 and 0.02565 for $m = 50, 200, 400$. For each run, the value changes by at most $1.9 \times 10^{-3}$ in relative terms over $\lambda \in \{10^{-4}, 10^{-5}, 10^{-6}\}$. The dominant off contact family is the $A$ type family (negative charge through $x_1$) in all 24 runs with $\lambda \le 10^{-3}$. At $\lambda = 10^{-2}$, which is not a small initialisation, it is the $B$ type family (positive charge through $x_2$) in two of the six runs. The expanded seed results below provide the comparison across widths.

**Robustness suite (60 runs) and expanded seeds.** The suite uses $(m, \lambda) \in \{200\} \times \{10^{-4}, 10^{-6}\}$ with seeds 0–19 and $\{50\} \times \{10^{-4}, 10^{-6}\}$ with seeds 0–9. Seeds 0 and 1 of each cell coincide with the corresponding sweep runs. The suite was first run without saving endpoints; the rerun analysed here reproduces every originally recorded diagnostic to the six printed decimals. Table 3 reports expanded seed counts at $\lambda = 10^{-6}$ separately from the sweep. The $m = 400$ row is an additional cell under the identical protocol, whose seeds 0 and 1 are the sweep runs.

| $m$ | $\lambda$ | seeds | min | median | max |
|---|---|---|---|---|---|
| 50 | $10^{-6}$ | 10 | 0.01597 | 0.02499 | 0.03416 |
| 200 | $10^{-6}$ | 20 | 0.02163 | 0.02648 | 0.03434 |
| 400 | $10^{-6}$ | 20 | 0.02300 | 0.02670 | 0.02989 |

Table 3: Dual slack excess diagnostic with expanded seed counts. The ranges overlap across widths.

**Extraction and step size sensitivity.** On the saved endpoints of all 80 runs (the 60 run suite and the $m = 400$ cell), we varied the grouping ($10^{-5}$), contact ($2 \times 10^{-3}$) and incidence ($10^{-6}$) tolerances separately by factors 0.1, 1 and 10. The diagnostic changed by at most $10^{-15}$. Wrong sign contact charge stayed zero, and the dominant off contact family kept negative charge and incidence $\{x_1\}$ in all 80 runs and all nine settings. Only the number of significant off contact families depends on the grouping tolerance, with median 4, 2 and 1 over the 80 runs. The incident point lies within $6.3 \times 10^{-17}$ of the family's hyperplane, and the next point is at least 0.17 away. Four runs were repeated with $\eta_{\text{rel}} = 10^{-3}$ and $\eta_{\max} = 5 \times 10^{-3}$ at the same horizon: the smallest diagnostic ($m = 50$, $\lambda = 10^{-6}$, seed 0), the run whose second closest point is nearest the incidence threshold, the run whose dominant family has the smallest charge share, and the smallest $m = 400$ value. Final

losses were at most $7.1 \times 10^{-28}$. The diagnostic changed by $+6.6 \times 10^{-6}$ to $+1.02 \times 10^{-5}$. The dominant family's sign and incidence and the zero wrong sign contact charge were unchanged. The number of significant off contact families changed from six to five in one run.

**Perturbation control.** Figure 2b uses width $m = 50$, seeds 0–3, seed map `default_rng(4157*seed+11)` and $9 \times 10^5$ steps ($1.5 \times 10^6$ for $\lambda \le 10^{-5}$), as in an earlier single seed control, whose saved values the seed 0 runs reproduce to $8 \times 10^{-8}$. The seeds, scales, perturbation values and two tests were written down in a design note before these runs were started. Test (i): the median change of the excess from $\lambda = 10^{-3}$ to $10^{-6}$ at $\varepsilon = 0$ must be smaller than at $\varepsilon = 1/10$. Test (ii): at $\varepsilon = 1/10$, at most one of the four seeds may have the opposite sign of that change to the others. Both tests passed ($0.0002$ versus $0.0043$; all four seeds decrease). Squared loss is below $1.9 \times 10^{-28}$ and wrong sign contact charge is zero in all 48 runs. Each geometry's excess uses its own exact dual certificate (Appendix B).

**Trajectory diagnostic.** The eight runs of Section 4 are $m \in \{50, 200\}$ with seeds $0, 1, 2$ and $m = 400$ with seeds $0, 1$, all at $\lambda = 10^{-6}$ under the protocol above, logged every ten steps for the first $6 \times 10^4$ steps. At each log the $C$ units are those with sign $s_C$ whose direction is within $0.05$ rad of $\hat{u}_C$; membership is recomputed at every log, and $Q_C$ sums $|a_j| \, \|(w_j, b_j)\|$ over them. The entrance conditions are evaluated at the first log with $S \ge 10^{-4}$, using the exact $\mu_\star$ and $L$ of Appendix D. Approach is the first log at which the $C$ units are non-empty, their charge weighted mean angle to $\hat{u}_C$ is below $10^{-3}$ rad and $|\cos(\mu, \lambda^*)| > 0.999$; departure is the first later log at which that mean angle exceeds $10^{-2}$ rad. The mean angle is a distance to the contact, not a dispersion about the units' own mean direction. These definitions apply to the eight reported trajectories.

## B Exact certificates and datasets

$X_{\mathrm{old}} = \{(1,0), (0,1), (-1, \frac{1}{2}), (\frac{1}{2}, -1)\}$ with $y = (1, 0, 2, 1)$; $X_\varepsilon$ replaces $x_4$ by $(\frac{1}{2}, -1 + \varepsilon)$. With $n = d + 2 = 4$ the matrix $M = [\mathbf{1}, X]$ has rank 3, so $\ker M^\top$ is one dimensional and spanned by $\lambda^* = (3, -3, 2, -2)$; this is Lemma 3, and it is why the contact geometry depends on $X$ alone. Normalising, $\hat{\lambda} = \frac{\sqrt{17}}{9}(3, -3, 2, -2)$.

Certificates are computed in exact arithmetic. Dual feasibility requires $\sup_{\mathcal{Z}} |\hat{g}_{\hat{\lambda}}| \le 1$ on the cylinder. For fixed $w$ the map $b \mapsto \hat{g}$ is piecewise linear with breakpoints at $b = -\langle w, x_i \rangle$, so the supremum over $b$ is attained at a breakpoint. On each arc of $w$ between incidence changes $\hat{g}$ is linear in $w$, so the problem reduces to finitely many exact comparisons. This yields

$$\max_{\mathcal{Z}} |\hat{g}_{\hat{\lambda}}| = 1, \qquad \mathcal{R}_{\min}(X_{\mathrm{old}}) = \langle \hat{\lambda}, y \rangle = \tfrac{5\sqrt{17}}{9}.$$

The maximum equals one, as complementary slackness requires. Strong duality is also certified from the primal side: for every $c_0 \in [8/9, 13/9]$, the interpolant with affine offset $c_0$ and charges $c_A = \sqrt{17}(8/9 - c_0) \le 0$, $c_B = \sqrt{17}(13/9 - c_0) \ge 0$ on the two contact hyperplanes has total charge $5\sqrt{17}/9$. The minimiser is therefore not unique. For example, $c_0 = 1$, $v = (-4/3, -2/3)$, $c_A = -\sqrt{17}/9$ and $c_B = 4\sqrt{17}/9$ give an explicit interpolant of this cost, with the cylindrical contact representatives $z_A = (1, 4, -1)/\sqrt{17}$ and $z_B = (4, 1, -1)/\sqrt{17}$.

The contact set is $\{z \in \mathcal{Z} : |\hat{g}_{\hat{\lambda}}(z)| = 1\}$. Its $\mathbb{S}^2$ representatives are two antipodal pairs, $\pm u_A$ with $u_A = (1, 4, -1)/\sqrt{18}$ and $\pm u_B$ with $u_B = (4, 1, -1)/\sqrt{18}$, incident to $\{x_1, x_3\}$ and $\{x_2, x_4\}$ respectively. On the sphere the unnormalised feature takes the values $g_{\hat{\lambda}}(\pm u_A) = -G_{\max}$ and $g_{\hat{\lambda}}(\pm u_B) = +G_{\max}$ with $G_{\max} = \sqrt{34}/6$. Since $\|w_A\| = \|w_B\| = G_{\max}$ exactly, $\hat{g} = g/\|w\| = \mp 1$: the two numbers describe one contact in two normalisations. For the perturbed geometries the same routine gives $\mathcal{R}_{\min}(X_{1/20}) = 18\sqrt{1621}/319$ and $\mathcal{R}_{\min}(X_{1/10}) = 69\sqrt{386}/602$, each with a single contact carrying an explicit, correctly signed primal interpolant. For comparison, the diamond $\{(1,0), (0,1), (-1,0), (0,-1)\}$ with $y = (1, 0, 1, 0)$ and XOR $\{(1,1), (1,-1), (-1,1), (-1,-1)\}$ with $y = (1, -1, -1, 1)$ give 2 and $2\sqrt{2}$.

**Canonical representation and cost.** Identify $z \sim -z$ on $\mathcal{Z}$. Because $c\,(-t)_+ = c\,(t)_+ - c\,t$, an atom at $-z$ equals an atom at $z$ with the same charge, up to an affine function. After merging

equal keys, every $f \in \mathcal{F}$ therefore has a representation with pairwise distinct hyperplanes, and this representation is unique modulo affine functions: a finite combination of ReLU ridges on distinct hyperplanes is affine only if every coefficient vanishes, because each hyperplane carries a jump $c_k w_k$ of the gradient. Any representation of $f$ merges to this canonical one, and merging can only decrease $\sum_k |c_k|$, so $\mathcal{R}(f) = |\alpha_f|(\mathcal{Z})$.

**Weak duality.** For any signed measure $\alpha$ of finite total variation on $\mathcal{Z}$ that represents an exact interpolant together with an affine part, $\int \hat{g}\, d\alpha = \langle \hat{\lambda}, y \rangle$ by the computation below. Hence $|\alpha|(\mathcal{Z}) \geq \langle \hat{\lambda}, y \rangle$. With the explicit primal interpolants above, $\mathcal{R}_{\min} = \langle \hat{\lambda}, y \rangle$ is the infimum over finite and infinite width representations alike.

**Proof of Lemma 1.** Let $\hat{g}(z) = \sum_i \hat{\lambda}_i (\langle w, x_i \rangle + b)_+$ with $M^\top \hat{\lambda} = 0$, $\sup_{\mathcal{Z}} |\hat{g}| \leq 1$ and $\langle \hat{\lambda}, y \rangle = \mathcal{R}_{\min}$ (strong duality for the minimum norm problem with a free affine part). Since $M^\top \hat{\lambda} = 0$, $\hat{g}(-z) = \hat{g}(z) - \langle \hat{\lambda}, M(b, w) \rangle = \hat{g}(z)$, so $\hat{g}$ is well defined on $\mathcal{Z}/\sim$. If $f(x) = \int_{\mathcal{Z}} (\langle w, x \rangle + b)_+ \, d\alpha_f(z) + \langle v, x \rangle + c$ interpolates, then $\int_{\mathcal{Z}} \hat{g}\, d\alpha_f = \sum_i \hat{\lambda}_i \big(y_i - \langle v, x_i \rangle - c\big) = \langle \hat{\lambda}, y \rangle = \mathcal{R}_{\min}$, again by $M^\top \hat{\lambda} = 0$. Writing $d\alpha_f = s_f \, d|\alpha_f|$ and $\mathcal{R}(f) = \int_{\mathcal{Z}} d|\alpha_f|$ gives the identity. □

The same affine annihilation calculation applies to any finite representation $\{(c_j, z_j)\}$ of a network with residual $r = f(X) - y$, giving $\sum_j |c_j| - \mathcal{R}_{\min} = \sum_j (1 - s_j \hat{g}(z_j)) |c_j| + \langle \hat{\lambda}, r \rangle$. On all 80 saved endpoints of Appendix A this identity holds to $1.2 \times 10^{-15}$, $|\langle \hat{\lambda}, r \rangle| \leq 8.6 \times 10^{-14}$, and merging within the grouping tolerance changes $\sum_j |c_j|$ by at most $7.4 \times 10^{-13}$.

## C Initialisation and the entrance family

The entrance section in Theorem 4 isolates the retention stage. Lemma 3 identifies the affine least squares residual with the null direction $\lambda^*$, and the balanced dynamics couple directional motion to charge growth. This gives a motivation for studying occupied contacts at small total charge. The theorem itself starts from the resting section with the ratios, remainder bound and compatibility identity specified in Appendix D.

A result from random initialisation would require two additional estimates: an entrance probability for $\mathcal{B}_0$, and a bound controlling the effect of dormant units, directional deviations and the residual remainder over the certified time interval. The distinction between $\mathcal{B}$ and $\mathcal{B}_0$ matters here: the latter occupies $1/24$ of the former in $r_A$. These estimates remain open. Section 4 instead locates the eight recorded full width trajectories directly through their measured entrance coordinates and margins.

## D Computer assisted proof of Theorem 4

**Reduced dynamics.** While the four cohorts rest at $\pm u_A, \pm u_B$, gradient flow in physical time (the time $\tau$ of Section 2, written with a dot) reduces exactly to

$$\dot{Q}_C = 2 s_C \, g_\mu(\hat{u}_C) \, Q_C, \qquad \dot{\mu} = -\tfrac{1}{4}\big(\Phi\Phi^\top + D(Q)\big)\mu,$$

where $D(Q) = \operatorname{diag}\big(Q_B, Q_A, \frac{9}{4}Q_{-B}, \frac{9}{4}Q_{-A}\big)$, $g_\mu(\hat{u}_A) = \mu_2/\sqrt{2}$, $g_\mu(\hat{u}_{-A}) = 3\mu_4/(2\sqrt{2})$, $g_\mu(\hat{u}_B) = \mu_1/\sqrt{2}$ and $g_\mu(\hat{u}_{-B}) = 3\mu_3/(2\sqrt{2})$, where $\Phi$ is the matrix with rows $\phi_i$. The stationarity margins are $\ell_A = 2\mu_2 + 3\mu_3$ and $\ell_B = -(2\mu_1 + 3\mu_4)$, positive exactly when $\theta_3^A < 1$ and $\theta_4^B < 1$. The validator propagates the state $x = (\mu_1, \mu_2, \ell_A, \ell_B, Q_A, Q_{-A}, Q_B, Q_{-B})$, a linear change of coordinates of $(\mu, Q)$, under the quotient field $dx/dS = f(x)/\dot{S}$, where $S = \sum_C Q_C$ and $\dot{S} > 0$ is certified on every step. Since $\Phi^\top \lambda^* = 0$ and $g_{\lambda^*}(\hat{u}_C) = s_C \, 3/\sqrt{2}$, the quantity $H = \lambda^{*\top}\mu + \frac{3}{4\sqrt{2}} S$ is conserved and equals $\lambda^{*\top} y/n = 5/4$ from $h \equiv 0$. More directly, for any network whose hidden charge rests on the four cohorts, $\lambda^*$ annihilates the affine part and $\lambda^{*\top} h(X) = \sum_C s_C Q_C \, g_{\lambda^*}(\hat{u}_C) = \frac{3}{\sqrt{2}} S$, so $\lambda^{*\top}\mu = \frac{5}{4} - \frac{3}{4\sqrt{2}} S$ and $H = 5/4$ for every realisable resting state, not only along the flow. Together with $\sum_C dQ_C/dS = 1$, this is enforced throughout.

**Entrance family.** Let $\mu_\star = \frac{5}{104}\lambda^*$ be the affine least squares residual field and $\gamma = 2s_C g_{\mu_\star}(\hat{u}_C) = 15\sqrt{2}/104$, the common growth rate of all four cohorts at $\mu_\star$. The first order local unstable graph of the resting system is $\mu = \mu_\star + LQ$, with columns $L_{\cdot C} = -\frac{1}{4}(\gamma I + \frac{1}{4}\Phi\Phi^\top)^{-1} D_C \mu_\star$, where $D_C$ is the part of $D(Q)$ multiplying $Q_C$. An entrance state is a resting state with $S = 10^{-4}$, ratios in $\mathcal{B}_0$, $\mu = \mu_\star + LQ + R$, $\|R\|_2 \le 0.262\, S^2$ and $H = 5/4$. Since ${\lambda^*}^\top \mu_\star = 5/4$ and ${\lambda^*}^\top L_{\cdot C} = -\frac{3}{4\sqrt{2}}$ for every $C$, the last condition is equivalent to ${\lambda^*}^\top R = 0$. Every trajectory of the resting system on the local unstable manifold of $(\mu_\star, 0)$ satisfies this bound while $S \le 10^{-4}$, with no restriction on the charge ratios beyond $Q \ge 0$. *Proof.* Let $z = \mu - \mu_\star$, $A = -\frac{1}{4}\Phi\Phi^\top$ and $E(Q) = -\frac{1}{4}D(Q)$. Then $\dot{z} = Az - \frac{1}{4}D(Q)\mu_\star + E(Q)z$ and $\dot{Q} = \gamma Q + \operatorname{diag}(Q)Cz$, where the rows of $C$ have norm $\sqrt{2}$ or $3/\sqrt{2}$. Since $(\gamma I - A)L = -\frac{1}{4}D(\cdot)\mu_\star$, the remainder $w = z - LQ$ satisfies

$$\dot{w} = Aw + E(Q)LQ - L\operatorname{diag}(Q)CLQ + E(Q)w - L\operatorname{diag}(Q)Cw.$$

For $Q \ge 0$ with $S = \sum_C Q_C$ we have $\|E(Q)\| \le \frac{9}{16}S$, $\|LQ\| \le \ell S$ with $\ell = \max_C \|L_{\cdot C}\| < \frac{16}{125}$, and $\|L\operatorname{diag}(Q)v\| \le \ell S \max_C |v_C|$. Hence $\|\dot{w} - Aw\| \le h_2 S^2 + h_1 S\|w\|$ with $h_2 = \frac{9}{16}\ell + c\ell^2$, $h_1 = \frac{9}{16} + c\ell$ and $c = \frac{1061}{500} \ge 3/\sqrt{2}$. Since $\Phi\Phi^\top \succeq 0$, we have $\|e^{A\tau}\| \le 1$ for $\tau \ge 0$ and $\|(2\gamma I - A)^{-1}\| \le 1/(2\gamma)$. The linearisation at $(\mu_\star, 0)$ is block triangular with eigenvalue $\gamma > 0$ on the charges and spectrum in $(-\infty, 0]$ on $z$. The unstable manifold is therefore a smooth graph $w = H_2(Q,Q) + O(|Q|^3)$ with $(2\gamma I - A)H_2(Q,Q) = E(Q)LQ - L\operatorname{diag}(Q)CLQ$, so $\|H_2(Q,Q)\| \le h_2 S^2/(2\gamma) < 0.2619\, S^2$ and $\|w\| < 0.262\, S^2$ for small $S$. Suppose $\|w\| \le 0.262\, S^2$ up to a time $t$ with $S(t) \le S_{\max} = 10^{-4}$. Every charge then grows at rate at least $\alpha = \gamma - c(\ell S_{\max} + 0.262 S_{\max}^2) > 0$, so $S(s) \le S(t)e^{-\alpha(t-s)}$ for $s \le t$. Variation of constants from $s = -\infty$ gives $\|w(t)\| \le (h_2 + 0.262\, h_1 S_{\max})S(t)^2/(2\alpha)$. With the rational bounds $\ell < 16/125$ and $\gamma > 2039/10000$, the constant equals $1334858772990/5096820821009 < 0.26191 < 0.262$, so the bound can never be reached. □ All inequalities were checked in exact arithmetic, independently of the validator's own derivation, which produces the same rational constant. The validator's entrance set bounds each coordinate of $R$ by $0.262\, S^2$ and imposes $H = 5/4$ exactly, so it contains every such state with ratios in $\mathcal{B}_0$. Before every propagation step the validator checks $\ell_B > 0$ on the whole current enclosure and fails the leaf otherwise; since every leaf passes, $\ell_B > 0$ holds on the entire entrance family and is a certified consequence, not a filter. The exact entrance certificate gives $\ell_A \ge 1.6288 \times 10^{-6} > 0$ on all of $\mathcal{B}$.

**Monitored margins.** On every propagation step the validator requires strict positivity of the four charges and of the secondary Filippov margins $2\mu_1 + \mu_2$ ($\theta_1^A < 1$), $-\mu_1 - 2\mu_2$ ($\theta_2^B < 1$), $4\mu_1 + 3\mu_4$ ($\theta_1^{-A} < 1$) and $-4\mu_2 - 3\mu_3$ ($\theta_2^{-B} < 1$). These margins imply $\mu_1, \mu_3 > 0 > \mu_2, \mu_4$. If the interval for $\ell_A$ is inconclusive, the reachable face $\ell_A = 0$ is intersected exactly and the validator certifies $F_A > 0$, an inward pointing field. The remaining Filippov coefficients of the antipodal cohorts vanish identically, $\theta_3^{-A} \equiv \theta_4^{-B} \equiv 0$, because $\phi_4 \in \operatorname{span}(u_A, \phi_1)$ and $\phi_3 \in \operatorname{span}(u_B, \phi_2)$. This is a structural identity, not a monitored margin. Resting of the $\pm$ cohorts nevertheless follows from the one sided fields. Toward the side where the point with the vanishing coefficient becomes active, the velocity is $-\frac{3}{2}\mu_3 < 0$ for $-A$ and $\frac{3}{2}\mu_4 < 0$ for $-B$. On the other side it vanishes at the contact, with linearisation $\frac{3\sqrt{2}}{4}\mu_4 < 0$ for $-A$ and $-\frac{3\sqrt{2}}{4}\mu_3 < 0$ for $-B$. Both cohorts are therefore attracting under the certified sign conditions, and resting is their unique Filippov continuation. An identically vanishing coefficient cannot leave $[0, 1]$, so it can never produce a competing event. Every one of the eight Filippov coefficients of the four cohorts is therefore either monitored strictly, a stationarity margin ($\ell_A$, $\ell_B$), or identically zero. The certified event ordering does not depend on a strict inequality for the two vanishing coefficients.

**Event.** For each terminal leaf, an event aware terminal step isolates a unique zero of $\ell_B$, proves $\ell_B > 0$ immediately before and $\ell_B < 0$ immediately after, and certifies $d\ell_B/dS < 0$ and $\dot{S} > 0$. At the event the exact constraints $\ell_B = 0$ and $H = 5/4$ are imposed, and the correlated exit family is retained as a finite union of constrained zonotopes rather than a global axis aligned hull. The computation uses outward rounded interval arithmetic, correlated zonotopes, a validated fourth order Taylor jet with Picard enclosure, and certified rational enclosures of $\sqrt{2}$.

**Certified constants.** The constants are taken over the 4608 leaves covering $\mathcal{B}_0$. $q_B^*$ is the minimum of $Q_B$ at the event and $\eta_B^*$ the minimum of $-F_B$ at the event. $\delta_{\mathrm{exit}}^*$ is the minimum at the event of nine quantities: $2\mu_1+\mu_2$, $\ell_A$, $2\mu_1$ (equal to $4\mu_1+3\mu_4$ on $\ell_B=0$), $-\mu_1-2\mu_2$, $-4\mu_2-3\mu_3$, $Q_A$, $Q_{-A}$, $Q_B$ and $Q_{-B}$. $T_{\max}$ is the maximum over leaves of an exact rational upper bound on the physical event time, obtained from the certified steps as a sum of $h/\inf \dot{S}$. The values recorded by the aggregation of the theorem run are

$$q_B^* = 0.08084046934833176, \qquad \eta_B^* = 4.8922432494763354\times 10^{-5},$$

$$\delta_{\mathrm{exit}}^* = 0.0039848162226836835, \qquad T_{\max} = 56.61838307457595,$$

and every leaf carries a certified time bound.

**Validation.** The frozen source commit is `8656e10167b4`; all 46 source files match the hashes of the clean-room audit. The audit suite checks interval rounding, exact algebraic constants, the conserved identities, the face barrier logic and the correlated terminal event construction. A second validated step implementation is compared against the main step on early, middle and near event boxes, and deliberately unresolved and missing leaf inputs are rejected as negative controls. A separate audit cover of the sub box $[59/10,119/20]\times[19/20,24/25]^2$, which is not the theorem box, passes with 256 leaves and an exact no-gap verifier. For the theorem, the 4608 leaf boxes are checked by an exact rational aggregator: every one of the $4\times 64\times 58$ elementary cells of $\mathcal{B}_0$ is covered exactly once. The aggregate record of the theorem run reports, for the box $\mathcal{B}_0$, all 4608 terminal leaves PASS, no coverage errors, the frozen source commit and a single validator binary, together with the constants above. It was computed from the per leaf summaries and time records; the certificate files themselves were not re-hashed in that step, and their SHA-256 values are listed in the leaf manifest. For this paper we reconciled that record with the manuscript and the leaf manifest (domain, leaf count, commit, binary hash and all four constants agree exactly); this is a check of the original results, not a new certificate verification. Every leaf box coincides with the box encoded by its leaf identifier under the frozen round zero grid and midpoint splitting rule (256 round zero parents, terminal depths 3 to 6). These checks do not constitute a formal proof of the C++ implementation, but they exercise its main failure modes separately from the theorem run.

The post event identity used in Corollary 5 is symbolic (Appendix E), so the local conclusion requires no numerical propagation beyond the event. Long time persistence is a separate question, as specified in Appendix G.

## E EXACT LOCAL MARGINAL CONTACT CALCULATION

At $u_B$ the incident points are $x_2, x_4$ and $x_1$ is active. Writing zero in the edge coordinates of the one sided parallelogram gives $\theta_2^B(\mu) = -\mu_1/(2\mu_2)$ and $\theta_4^B(\mu) = -2\mu_1/(3\mu_4)$, and at the aligned field $(\theta_2^B,\theta_4^B) = (\frac{1}{2},1)$. With $\ell_B = -(2\mu_1+3\mu_4)$ the face is $\ell_B=0$. On the face the resting equations give

$$\dot{\ell}_B = F_B, \qquad F_B = \tfrac{1}{2}(\mu_1+\mu_2) + \tfrac{1}{2}\mu_1 Q_B - \tfrac{9}{8}\mu_1 Q_{-A},$$

and off the face $\dot{\ell}_B = F_B - c_B\ell_B$ with $c_B = \frac{13}{16} + \frac{9}{16}Q_{-A}$. The certified strict inequality $-F_B>0$ at the event makes the crossing transversal.

After the event $x_4$ becomes active and the $B$ cohort slides along the kink $\langle u,\phi_2\rangle = 0$. Its unit tangent at $u_B$ pointing to the side $\langle u,\phi_4\rangle > 0$ is $e=(1,-2,2)/3$, so the tangential velocity is

$$v_B = \langle \mu_1\phi_1 + \mu_4\phi_4, e\rangle = \mu_1 + \tfrac{3}{2}\mu_4 = -\tfrac{1}{2}\ell_B.$$

At the event $v_B=0$ and $\dot{v}_B = -\frac{1}{2}F_B>0$, which gives strictly positive initial acceleration into the positive dual slack side. This is the exact local statement used in Corollary 5.

For contact $A$ the mirror statements hold: $\ell_A = 2\mu_2+3\mu_3$, and on its escape chart $A$ keeps the incidence with $x_1$ while $x_3$ becomes active, with $v_A = -\frac{1}{2}\ell_A$. An escaped $B$ cohort therefore leaves positive charge on a hyperplane through $x_2$, and an escaped $A$ cohort leaves negative charge on a hyperplane through $x_1$.

## F ONE DIMENSIONAL CONTROL DETAILS

The datasets are $x = (-3, -1, 1, 3)$ with $y = (2, 0, 0, 2)$ and $x = (-2, -1, 1, 4)$ with $y = (1, 0, 0, 6)$. For the architecture with a free affine part, the minimum of $\mathcal{R}$ is the total variation of the slopes of the linear interpolant, 2 and 3 respectively. The suite uses widths $\{2, 8, 16, 32, 64\}$, $\lambda \in \{10^{-2}, 10^{-3}, 10^{-4}\}$, two direction laws (uniform on the circle and uniform on a half circle) and six seeds, with the relative step integrator for $3 \times 10^5$ steps. A run counts as interpolating if its loss is below $10^{-14}$.

With the free affine part, 174 of 180 and 177 of 180 runs interpolate. All interpolating runs end within $2.3 \times 10^{-3}$ and $3.1 \times 10^{-3}$ of the minimum, with medians $1.6 \times 10^{-6}$ and $3.5 \times 10^{-6}$. For fixed dataset and width the median gap divided by $m\lambda^2$ is constant across the three scales to within about $12\%$, with values between $0.011$ and $0.26$. This matches the cost of dormant units, which stay at size $O(\lambda)$ and cost $O(\lambda^2)$ each. The non interpolating runs occur only at $m = 2$ and end at the affine least squares fit. A variant with only a trainable output bias does not reach its own minimum: 9 of 153 and 130 of 152 interpolating runs miss it by more than $10^{-2}$. It is not used as a control. The one dimensional experiment checks the extraction pipeline and is not evidence for the two dimensional theorem.

## G RESULTS AND EVIDENCE

| Status | Claim |
|---|---|
| Exact | Variational minimum and primal attainment, optimal dual contact geometry, intrinsic dual slack excess identity, early aligned residual identity, reduced resting dynamics and conserved $H$, entrance remainder bound $\|R\|_2 \le 0.262S^2$ (analytic bootstrap, exact rational constants), structural vanishing of the two antipodal Filippov coefficients, local post event identity $\dot{v}_B = -F_B/2$. |
| Interval certified | Theorem 4 in the reduced dynamics: unique first $B$ stationarity loss on $\mathcal{B}_0$, cover of $\mathcal{B}_0$ by 4608 leaves, $\ell_B > 0$ at entrance, positive exit charge, transversality, positive monitored competing quantities, finite exit time (constants as recorded by the theorem run, Appendix D). |
| Numerical | Under the fixed four point dataset, architecture and balanced initialisation protocol: interpolation to near machine precision; a positive diagnostic in all reported runs; in eight tracked runs, tolerance defined approach to $A$, a later sign change of $\ell_A$, and subsequent crossing of the departure threshold with cohort membership recomputed at each log; the $A$ escape signature at every robustness endpoint; tie splitting changes the scale dependence across four seeds. |

The exact statements concern the specified representation cost and construction. The interval result is uniform over the reduced entrance family $\mathcal{B}_0$; the training observations concern the listed finite widths and protocols. Local exit and endpoint excess are connected by the surviving canonical mass condition of Corollary 2. Extending this connection to a theorem from random initialisation requires the entrance estimates described in Appendix C and control of the subsequent evolution up to interpolation.