# NORMALIZATION COUPLES WHAT THE OBJECTIVE SEPARATES: REPRESENTATION SELECTION IN GROUP COMPOSITION NETWORKS

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

## ABSTRACT

We study when the representation selected by a normalized network can be predicted by scoring candidate representations independently. The model is a two-layer quadratic network on a finite group's multiplication table, with real Fourier sectors as the candidates. At small logit scale the leading objective has no cross-sector terms, but the sphere-projection multiplier depends on the sum of all sector terms. A sector-local rule may score each candidate from its own Fourier coefficients and from every sector's per-layer magnitudes, but not from another sector's signs, phases or matrix directions. Let the joint input sphere have metric weight $\omega$ and the output sphere weight one; ordinary normalization has $\omega = 1$. For groups with at least three nontrivial real Fourier sectors, sector-local rules recover the leading selector almost everywhere if and only if $\omega = 2$, the degree-matched weight. At every other fixed weight, the errors of measurable sector-local rules under uniform initialization are bounded away from zero. The proof reflects the output block of a third sector that wins at neither start and thereby switches the winner between two rivals whose permitted inputs are unchanged.

Under ordinary normalization, for each fixed group and finite width and almost every initialization, exact cross entropy inherits the leading labels at its infinite-time endpoints as the scale vanishes. At width one, with three distinct nontrivial real linear characters, the optimal sector-local error of continuous exact cross entropy is positive at every sufficiently small positive scale, even at $\omega = 2$, where it converges to zero with the scale, and uniformly over any fixed compact set of weights. On two-bit XOR under ordinary normalization, a simple sector-local rule errs on $1.7\%$ of sampled initializations, while the leading selectors of jointly and separately normalized networks have identical label frequencies but disagree on an estimated $18\%$ of paired starts. These examples separate independent representation competition from both approximate prediction and agreement of label frequencies.

## 1 INTRODUCTION

Fix an architecture, a task and an initialization. Which representation does training select? The object of interest is the map from an initialization to the representation its trajectory reaches; final loss and population frequencies answer different questions. When the leading objective is a sum of one term per candidate representation, the natural answer is an independent contest: score each candidate from its own initial data, and the best score wins. Two recent selector formulas in restricted Abelian settings have this form: He et al. (2026b, v2, Theorem 5.3(ii)) and He et al. (2026a, Corollary 6.1). But an additive objective separates only the gradients. Normalized training follows their projections onto spheres, and each projection subtracts a radial term proportional to the sum of all candidates' terms. We ask when this shared term can change the winner, in a model where both the candidates and the contest are explicit.

**Two-bit XOR: the objective separates.** The task is composition in a finite group $G$: the data is the complete multiplication table $\{((g, h), gh)\}$. The smallest example with three nontrivial real



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

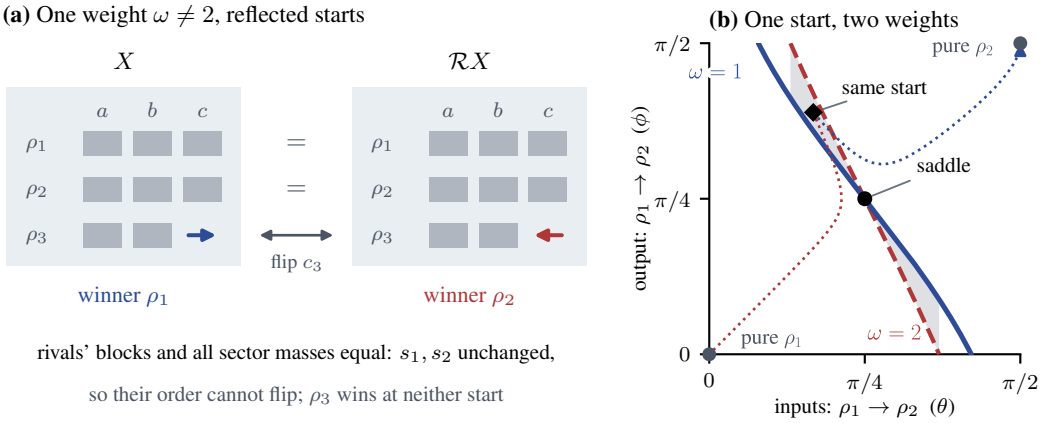


Figure 1: **A third sector that wins at neither start can decide between two unchanged rivals. (a)** At a fixed mismatched weight, $X$ and $\mathcal{R}X$ differ only in the sign of the third sector's output block. The rivals' blocks and all sector masses are unchanged, yet the winner switches; Theorem 3.1 lifts such pairs to paired open sets. At $\omega = 2$ each sector's clock depends only on its own blocks, so almost surely this reflection cannot exchange the two rivals. **(b)** On an invariant surface of balanced inputs and aligned blocks, the ordinary and degree-matched metrics have different stable boundaries through the same mixed saddle, so the marked start reaches different pure sectors. The curves are computed from the reduced system of Appendix B.7; Proposition 3.2 gives their slopes at the saddle analytically. The two sectors have equal ceilings (maxima of $f$ on each).

sectors is coordinatewise XOR on two-bit strings, $G = C_2 \times C_2$, whose three nonconstant features are the parities $\chi_1(g) = (-1)^{g_1}$, $\chi_2(g) = (-1)^{g_2}$ and $\chi_3(g) = (-1)^{g_1+g_2}$. Each turns XOR into multiplication, $\chi_j(g+h) = \chi_j(g)\chi_j(h)$. A neuron carries two input vectors $a, b$ and an output vector $c$; with unit vectors $e_j = \chi_j/2$ and coefficients $a_j = \langle a, e_j \rangle$, and likewise $b_j, c_j$, the first nonconstant term of cross entropy at one neuron is a *negative* multiple of

$$f(a, b, c) = \tfrac{1}{8}\big(a_1 b_1 c_1 + a_2 b_2 c_2 + a_3 b_3 c_3\big), \tag{1}$$

so leading cross-entropy descent ascends $f$. One character coordinate across $a, b, c$ is one *sector*: one candidate representation. Almost every leading trajectory ends on a single sector, and the map from an initialization to that sector is the *selector*. Equation (1) has no term coupling two sectors, which makes three independent competitors the natural hypothesis.

**Normalization couples them anyway.** Each neuron's first layer is one vector, the concatenation $(a, b)$ that reads the one-hot pair, so normalizing it puts both inputs on one sphere, $\|a\|^2 + \|b\|^2 = 1$, with $\|c\| = 1$. Ascent constrained to a sphere replaces $\nabla f$ by $\nabla f - \langle x, \nabla f \rangle\, x$, and the subtracted multiplier reads the whole state. As $f$ is trilinear, Euler's identity gives $\langle a, \nabla_a f \rangle = \langle b, \nabla_b f \rangle = \langle c, \nabla_c f \rangle = f$, and the input sphere adds the first two:

$$\dot{a} = \nabla_a f - 2fa, \qquad \dot{b} = \nabla_b f - 2fb, \qquad \dot{c} = \nabla_c f - fc. \tag{2}$$

The gradients keep the sectors apart; the multipliers do not. Every block feels the total $f$, the sum of all three sectors' terms, and the input blocks feel it twice as strongly as the output block. This shared radial feedback is the only channel between the candidates.

**A losing third sector can decide between two others.** Could a better score, built for this coupled geometry, still rank the candidates one at a time? Let a candidate's score read its own coefficients and every sector's magnitude in each layer, but no rival's signs, phases or matrix directions. Now change exactly what such scores cannot see: keep sectors 1 and 2 entirely fixed, keep every magnitude fixed, and reflect the sign of sector 3's output coefficient, giving states $X$ and $\mathcal{R}X$. If the winner switches from sector 1 to sector 2, no such rule is right at both states. The scores of sectors 1 and 2 are the same at $X$ and $\mathcal{R}X$, so their order cannot reverse, and choosing sector 3, which wins at neither, is wrong at both (Figure 1(a)). The question is when such switches happen on sets of positive probability.

The main result characterizes when such an independent ranking exists. Give the joint input sphere metric weight $\omega$ and the output weight one; ordinary normalization is $\omega = 1$. For groups with at least three nontrivial real Fourier sectors, such *sector-local* scores reproduce the leading selector almost everywhere if and only if $\omega = 2$. At every other fixed weight the reflection above switches the winner on open sets, so every measurable rule in the class has a positive error floor (Theorem 3.1). The distinguished weight is determined by degree: the input sphere carries two of the cubic's three factors, and weight two matches its radial rate to the output's.

Under ordinary normalization, exact cross entropy (CE), as flow and as retracted descent, converges at small scale to endpoints carrying the leading labels (Theorem 4.2), so the floor persists under the actual loss (Corollary F.2). At width one and $\omega = 2$, the optimal sector-local error of continuous exact CE tends to zero with the scale (Proposition D.4), but with three distinct nontrivial real linear characters it remains positive at every sufficiently small positive scale; the same obstruction persists when the weight is retuned within a fixed compact interval (Theorem 5.1). On two-bit XOR under ordinary normalization, a simple sector-local rule has empirical error $1.7\%$, and the optimal error, positive by Theorem 3.1, is estimated at $0.08\%$ (Section 6). Metric-dependent basins, radial removal, earliest-blow-up selection, layer balance and the compensating factor (He et al., 2026b) are known ingredients. Theorem 3.1 supplies the converse over the full sector-local class and extends both directions to the full joint sphere; Sections 4–5 determine what remains under exact CE.

## 2 The model and the class of independent scores

Let $G$ be nontrivial with $N = |G| \geq 2$ and let the data contain every $((g, h), gh)$. At common input/output scale $s$ the width-$M$ network, $1 \leq M < \infty$, has

$$z_k^{(g,h)}(X; s) = \frac{s^3}{M} \sum_{m=1}^{M} (c_m)_k (a_{m,g} + b_{m,h})^2, \qquad \|a_m\|^2 + \|b_m\|^2 = 1, \quad \|c_m\| = 1, \tag{3}$$

so $\mathcal{M}_M = (S^{2N-1} \times S^{N-1})^M$. Up to the sphere constraints and the scale this is the unnormalized two-layer quadratic network of Gromov (2023); we impose the constraints and compare gradient metrics on them. At width one every logit vector is a nonnegative multiple of one output vector, so the model isolates sector selection and cannot represent the full classifier.

Softmax ignores a logit vector's mean; with the centering projection $P = I - N^{-1}\mathbf{1}\mathbf{1}^\top$ define

$$\Omega(a, b, c) = \frac{1}{N^2} \sum_{g,h} a_g b_h c_{gh}, \qquad f(a, b, c) = \Omega(a, b, Pc) = \Omega(Pa, Pb, Pc), \tag{4}$$

which is (1) on $C_2 \times C_2$. The summed full-batch loss splits exactly as $L_s = N^2 \log N - \lambda_s \mathcal{F} + R_s$, with $\mathcal{F} = \sum_m f(a_m, b_m, c_m)$, $\lambda_s = 2s^3 N^2/M$, and a remainder $R_s$ bounded in (39). Dividing by $\lambda_s$, exact-CE descent becomes ascent on $\Phi_s = \mathcal{F} - R_s/\lambda_s$, with $\Phi_s - \mathcal{F} = O(s^3)$ in every finite derivative order used below (Appendix B). The leading potential evolves each neuron separately; the remainder couples them. Dropping $R_s$ gives the first-stage spectral objective of He et al. (2026b) up to an additive constant and a positive factor.

**Real sectors.** In general the parities become matrix blocks. The real group algebra splits as

$$\mathbb{R}[G] \cong \bigoplus_{\rho} M_{n_\rho}(D_\rho), \qquad D_\rho \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}, \tag{5}$$

and a *real sector* is one summand across $a, b, c$, with orthogonal projector $\Pi_\rho$. It may be of real, complex or quaternionic type, a conjugate pair of complex irreducibles counts as one sector, and centering removes the trivial one. The potential $f$ is a sum of one term per nontrivial sector and its gradient is block diagonal, so sectors interact only through the constraints. Rank means intrinsic rank over $D_\rho$; a quaternionic block of intrinsic rank one has complex rank two (Wolf, 1936; Zhang, 1997).

**The metric family and its comparator.** Let $g_{ab}$ and $g_c$ be the round metrics of the input and output spheres, and give the input sphere weight $\omega > 0$: $g_\omega^J = \omega g_{ab} + g_c$ on $\mathcal{M}_1$. Ordinary normalization is $\omega = 1$, with ascent flow (2); the weight divides the whole input gradient, radial part included, by

$\omega$. The separately normalized comparator puts $a, b, c$ on their own spheres, $\widetilde{\mathcal{M}} = (S^{N-1})^3$ with $g_\lambda = \lambda g_a + \lambda g_b + g_c$, and ascends the same centered potential. The map $\Psi(\widetilde{a}, \widetilde{b}, c) = (\widetilde{a}/\sqrt{2}, \widetilde{b}/\sqrt{2}, c)$ identifies $\widetilde{\mathcal{M}}$ with the balanced joint states $\|a\| = \|b\| = 1/\sqrt{2}$, which the joint flow preserves; it halves the potential, which only rescales time, and halves the input weight,

$$\Psi^* g_\omega^J = g_{\omega/2}: \qquad \text{ordinary } \omega = 1 \leftrightarrow \lambda = \tfrac{1}{2}, \qquad \text{degree-matched } \omega = 2 \leftrightarrow \lambda = 1 \text{ (separate).} \quad (6)$$

Results are stated in $\omega$; $\lambda = \omega/2$ appears only in proof coordinates and for the separate comparator, and the loss factor $\lambda_s$ is unrelated to it. Under every fixed weight, almost every leading trajectory converges to a maximum supported on a single nontrivial sector (Proposition 4.1 and Appendix C). Write $S_\omega^J$ for the resulting leading selector on $\mathcal{M}_1$, $S_J = S_1^J$ for ordinary normalization, and $S_\lambda$ for the separate selector under $g_\lambda$ (Theorem B.1).

**Definition 2.1** (Sector-local rules)**.** For $X = (a, b, c)$ let $\mathfrak{m}(X) = \left(\|\Pi_k a\|^2, \|\Pi_k b\|^2, \|\Pi_k c\|^2\right)_k$ be the per-layer sector masses over all real sectors $k$, the trivial one included. A rule is *sector-local* if

$$\sigma(X) = \arg\max_\rho \; s_\rho\big(\Pi_\rho a, \Pi_\rho b, \Pi_\rho c; \mathfrak{m}(X)\big)$$

over the nontrivial sectors $\rho$, with ties broken by one fixed state-independent priority, where each $s_\rho$ may depend on $\rho$, $G$ and the fixed metric weight and is otherwise unrestricted. A score thus reads its own sector's complete blocks and every sector's magnitudes, never another sector's orientation: its signs, phases or matrix directions. Against a selector $S$ the class has optimal error $\mathrm{err}^*(S) = \inf_\sigma \mathbb{P}[\sigma \neq S]$, the infimum over measurable sector-local rules.

Both selector formulas rank a sector from its own coefficients and aggregate magnitudes, so they lie in this class (Section 7). The restriction is substantive: on $C_2 \times C_2$, the masses fix the magnitudes of the three summands of (1), and the exact total $f$ then fixes their signs almost surely, so one additional scalar suffices for almost-sure prediction (Remark F.4).

## 3 When independent ranking is exact

**Why weight two.** Drop the constraints and follow the weighted gradient in an unconstrained lift, $a' = \omega^{-1}\nabla_a f$, $b' = \omega^{-1}\nabla_b f$, $c' = \nabla_c f$. The input sphere carries two of the cubic's three factors and the output one, so Euler's identity gives $\mathrm{d}(\|a\|^2 + \|b\|^2)/\,\mathrm{d}\xi = 4f/\omega$ against $\mathrm{d}\,\|c\|^2\,/\,\mathrm{d}\xi = 2f$: these *radial rates* agree, and equal radii stay equal, exactly at $\omega = 2$, one unit of metric weight per factor carried. At $\omega = 1$ the squared input radius changes twice as fast, the $2f$ against $f$ of (2). The network's rescaling symmetry fixes the matched weight. The logits (3) are unchanged by the per-neuron rescaling $(a_m, b_m, c_m) \mapsto (e^\tau a_m, e^\tau b_m, e^{-2\tau} c_m)$, so any loss $L$ of the logits, exact CE included, has $\langle a_m, \nabla_{a_m} L\rangle + \langle b_m, \nabla_{b_m} L\rangle = 2\langle c_m, \nabla_{c_m} L\rangle$, and unconstrained gradient flow in $\omega g_{ab} + g_c$ conserves $\frac{\omega}{2}(\|a_m\|^2 + \|b_m\|^2) - \|c_m\|^2$ for every neuron: the layer balance of homogeneous networks (Du et al., 2018; Kunin et al., 2021). Only at $\omega = 2$ does it keep equal input and output radii equal as they grow. The cubic adds homogeneity, which turns that common radius into a change of time; as $f$ is block diagonal over the real sectors, the matched lift is a disjoint family of sector systems, and the constraints only reparametrize its time.

**Theorem 3.1** (Sector-local ranking is exact only at the degree-matched weight)**.** *Let $G$ have at least three nontrivial real sectors, and for fixed $\omega > 0$ let $S_\omega^J$ be the leading selector on the joint manifold $\mathcal{M}_1 = S^{2N-1} \times S^{N-1}$ under $g_\omega^J = \omega g_{ab} + g_c$. Then*

$$S_\omega^J \text{ has a sector-local representation almost everywhere} \iff \omega = 2.$$

*For every fixed $\omega \neq 2$, the optimal error $\mathrm{err}^*(S_\omega^J)$ is positive under normalized volume; in particular $\mathrm{err}_J^* := \mathrm{err}^*(S_J) > 0$ for ordinary normalization. The positive direction at $\omega = 2$ holds for every nontrivial finite group. On the separately normalized comparator $\widetilde{\mathcal{M}}$ the same holds for $S_\lambda$, with $\lambda = 1$ in place of $\omega = 2$ and $\mathrm{err}_\lambda^* := \mathrm{err}^*(S_\lambda) > 0$ for $\lambda \neq 1$.*

Appendix C proves the separate form and carries it to the full joint manifold, unbalanced inputs included (Corollary C.4). The two directions have different proofs.

**At $\omega = 2$: independent clocks.** With matched rates and no interaction between the lifted sectors, each sector $j$ has a maximal forward lifetime $T_j \in (0, \infty]$, finite if its lifted norm blows up: a *clock*, computed from its own blocks alone. On XOR, sector $j$'s lifted system is $a_j' = b_j c_j/16$, $b_j' = a_j c_j/16$, $c_j' = a_j b_j/8$, started at its initial coordinates. A common time change cannot reorder the clocks, the earliest finite one belongs to the winner almost surely, and $s_j = (1 + T_j)^{-1}$ is the required rule; selecting the feature that blows up first is the rule of Zhao et al. (2026, Section 5.1) for the scalar system. On XOR the clock is an elliptic integral, $T_j = 8\sqrt{2}\, R_F(a_j^2, b_j^2, c_j^2/2)$ with $R_F$ Carlson's symmetric form when $a_j b_j c_j > 0$ (Appendix H.3 gives the other sign); for a larger block it is the blow-up time of the sector's own equations.

**At $\omega \neq 2$: the reflection.** Take two rivals and the mixed saddle of their competition on the invariant surface where only they are active. Its stable manifold separates their basins, and its slope already depends on the weight (Proposition 3.2 below). Now add a third sector with input and output amplitudes $u$ and $v$. The boundary between the rivals becomes a graph whose first term odd in $v$, the only kind the reflection $v \mapsto -v$ can see, has coefficient $(\lambda - 1)\widetilde{C}(\lambda)$ with $\widetilde{C}(\lambda) \neq 0$ and $\lambda = \omega/2$: nonzero exactly at a mismatched weight. There are therefore reflected pairs on opposite sides of it. A local exit argument sends the two sides into different pure-sector traps, continuous dependence extends the pair to paired open sets of the ambient manifold, and pairing each set with its reflection gives a positive error lower bound for every allowed rule (Appendix C). At $\omega = 2$ the clocks force the rival boundary to be reflection invariant.

**Proposition 3.2** (The weight alone moves the winner)**.** *Let $|G| \geq 4$, which is exactly when $G$ has at least two nontrivial real sectors; they may be of any types and ceilings. For $\lambda_1 \neq \lambda_2$, ascent of the same centered potential on $\widetilde{\mathcal{M}}$ under $g_{\lambda_1}$ and under $g_{\lambda_2}$ selects different sectors from every point of a nonempty open set, which has positive probability under uniform initialization. In the aligned chart of the two sectors, the stable slope of their mixed saddle is $m(\lambda) = -(\lambda + \sqrt{\lambda^2 + 8\lambda})/2$, strictly decreasing in $\lambda = \omega/2$.*

Metric-dependent basins are classical (Fiedler et al., 1999; Belabbas, 2023); this is the explicit instance here (Figure 1(b), Appendix B.7). Proposition 3.2 compares the selectors induced by two metrics. Theorem 3.1 fixes a single mismatched metric and bounds the error of every measurable sector-local rule away from zero, including rules built for that metric.

**Scope.** The theorem covers every group with at least three nontrivial real sectors, of any types, among them $C_2 \times C_2$, $D_4$, $Q_8$, $C_6$ and $C_7$. Groups with exactly two, such as $C_4$, $C_5$, $S_3$ and $A_4$, keep the clocks at $\omega = 2$ and Proposition 3.2, but offer no third block to reflect; whether their selector is sector-local at $\omega \neq 2$ is open (Remark F.3).

## 4 From the leading selector to exact cross-entropy endpoints

At small scale, exact CE keeps the leading labels at its infinite-time endpoints (Theorem 4.2). Finite-horizon closeness of the two fields is insufficient: settling near the flat stratum is not uniform, and the Gronwall error grows with time. The bridge is a capture argument starting where leading trajectories end; the next statement's separate-sphere analogue is He et al. (2026b, v2, Theorem 4.3).

**Proposition 4.1** (Leading maxima and almost-sure convergence)**.** *The positive local maxima of one-neuron $f$ on $S^{2N-1} \times S^{N-1}$ are exactly the states supported on one nontrivial real sector $\rho$ with balanced inputs and aligned blocks of intrinsic rank one, $\Pi_\rho a = \alpha p q^*$, $\Pi_\rho b = \beta r p^*$, $\Pi_\rho c = \gamma r q^*$ for unit $p, q, r$ over $D_\rho$ and $\alpha, \beta, \gamma > 0$. Each maximum component is Morse–Bott: its Hessian is uniformly negative in every normal direction, and its kernel is exactly the tangent space of the aligned orbit. Every other critical point is a strict saddle for ascent except the fully trivial flat stratum $\{(a, b) = (\cos t, \sin t)\mathbf{1}/\sqrt{N},\ c = \pm\mathbf{1}/\sqrt{N}\}$, on which the Hessian vanishes identically. Every trajectory of (2) converges to one critical point, and under any initialization law absolutely continuous with respect to Riemannian volume almost every trajectory converges to one of those maxima, simultaneously for all neurons at finite width.*

Analyticity gives convergence to a point, the strict-saddle-subset theorem of Bah et al. (2022, arXiv v5, Theorem 28) excludes the other saddles, and a conserved ratio of the trivial coordinates puts the basin of the flat stratum, which strict-saddle avoidance does not reach, in a null set (Appendix B). A

leading trajectory therefore enters any neighborhood of its limiting maximum component in finite time, a horizon over which exact CE stays close. The Morse–Bott normal gap gives that neighborhood a collar of strictly lower potential, which exact CE inherits at small scale and cannot cross, because $\Phi_s$ increases along its own flow; analytic convergence then supplies an exact-CE endpoint inside the collar. For the retracted iteration $X_{k+1} = \mathrm{Ret}_{X_k}(h_s \operatorname{grad} \Phi_s(X_k))$, with $h_s = \eta\lambda_s$ and $\mathrm{Ret}_x(v) = (x+v)/\|x+v\|$ on each sphere block, sufficient ascent with a step-length bound also rules out a single step jumping the collar. A fixed physical learning rate $\eta$ drives the effective step to zero with $s$.

An endpoint's *label* is the unique maximizer of the canonical scores

$$S_\rho(a,b,c) = \|\Pi_\rho a\|^2 \|\Pi_\rho b\|^2 \|\Pi_\rho c\|^2 \tag{7}$$

read at that endpoint; a trajectory that does not converge, or ends at a tie, is unresolved, which counts as an error for every rule (Appendix D.8). Here $S_\rho$ reads the evolved state, not the initialization, and the label it records does not require exact Fourier support. For an initialization whose leading endpoint is a maximum, let $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_M)$ be its neuronwise labels and $K_{\boldsymbol{\rho}}$ the connected leading-maximum component containing that endpoint.

**Theorem 4.2** (Exact-CE sector endpoints)**.** *Fix a nontrivial finite group and a finite width, use ordinary normalization ($\omega = 1$), and let the initialization law be absolutely continuous with respect to Riemannian volume on $\mathcal{M}_M$.*

*(i) For almost every $X_0$, there is $s_0(X_0) > 0$ such that continuous exact CE converges to an endpoint whose neuronwise labels equal those of the leading endpoint, for all $0 < s < s_0$.*

*(ii) For every $0 < \eta_{\max} < \infty$ and almost every $X_0$, there is $s_0(X_0, \eta_{\max}) > 0$ such that the retracted CE iteration converges to an endpoint with the same neuronwise label vector, for every $0 < s < s_0$ and $0 < \eta \le \eta_{\max}$.*

*(iii) The continuous endpoint in (i) and every retracted endpoint in (ii) satisfy $\mathrm{dist}(X_\infty^{\mathrm{CE}}, K_{\boldsymbol{\rho}}) = O(s^3)$, with a constant independent of $s$ and $\eta$; the position along $K_{\boldsymbol{\rho}}$ is unconstrained.*

**Sector-local prediction under exact CE.** Changing the target label changes the optimal error of a fixed class of rules by at most the probability that the two targets differ. Under ordinary normalization at width one that probability vanishes as $s \downarrow 0$. So for groups with at least three nontrivial real sectors, under normalized volume, the best exact-CE endpoint-label error of sector-local rules, even rules chosen after $s$ and $\eta$, converges to $\mathrm{err}_J^* > 0$, uniformly over $0 < \eta \le \eta_{\max}$ (Corollary F.2). For continuous flow this holds at every fixed weight, with limit $\mathrm{err}^*(S_\omega^J)$, zero at $\omega = 2$ (Proposition D.4). Under i.i.d. uniform initialization the expected fraction of changed labels vanishes uniformly in the width, so the expected fraction of mispredicted neurons has asymptotic lower bound $\mathrm{err}_J^*$ as $s \downarrow 0$, even for rules that also read the other neurons' complete initial states (Corollary F.1).

## 5 Does exact sector-local ranking survive the actual loss?

At width one and $\omega = 2$, the optimal sector-local error of continuous exact CE vanishes as $s \downarrow 0$ (Proposition D.4). For groups with three distinct nontrivial real linear characters, however, it is positive at every sufficiently small positive scale.

**Theorem 5.1** (Exact CE is not sector-local at small positive scale)**.** *Let $G$ have three distinct nontrivial real linear characters, equivalently a $C_2 \times C_2$ quotient, as $C_2 \times C_2$, $D_4$ and $Q_8$ do, and let $M = 1$. For continuous exact CE on $\mathcal{M}_1$ with the degree-matched metric $g_2^J = 2g_{ab} + g_c$, there is $s_* > 0$ such that for each $0 < s < s_*$ the endpoint-label error of measurable sector-local rules has a positive infimum under normalized volume. The same holds simultaneously for every weight in a fixed compact subset of $(0, \infty)$, hence for any weight $\omega(s)$ retuned with the scale inside such a set (Corollary E.1). The rule may depend on $s$; no bound uniform as $s \downarrow 0$ is asserted.*

**Why exactness fails at positive scale.** The clocks needed matched radial rates and a lifted system that is homogeneous and splits into sectors. The layer balance of Section 3 holds for every loss of the logits, so at $\omega = 2$ equal radii persist under exact CE. What is lost is the homogeneous, sector-separable lift. A common radius $R$ rescales the effective logit scale from $s$ to $sR$, and the cubic and its first correction, quartic in the joint input and quadratic in the output, carry different

Table 1: The error quantities of Section 6 at width one; ranges run over $C_2 \times C_2$, $D_4$ and $Q_8$. The disagreement rates are point estimates from 20,000 paired draws per group (Figure 2(a)). The $\mathrm{err}_J^*$ entry is a Monte Carlo estimate, not a bound, computed after the fact like the raw-block rate (Appendix H.3). The witness entries are each group's best one-sided 95% lower confidence bound, simultaneous over its reflections and conditional on the numerical labels (Appendix H.2).

| quantity | compares | relation | measured |
|---|---|---|---|
| $\mathrm{err}_J^*$ | best sector-local rule vs. $S_J$ | target | $8.1 \times 10^{-4}$ on $C_2 \times C_2$ |
| raw-block clock | one sector-local rule vs. $S_J$ | $\geq \mathrm{err}_J^*$ | 0.0170 on $C_2 \times C_2$ |
| $\frac{1}{2}\mu_J(W_j)$ | reflected pairs with two rival labels | $\leq \mathrm{err}_J^*$ | $4.40 \times 10^{-5}$ to $6.09 \times 10^{-4}$ |
| $q_{\mathrm{op}}$ | joint vs. separate selector, one draw | $\geq \mathrm{err}_J^*$ | 0.145 to 0.184 |
| $q_{\mathrm{met}}$ | metric $\lambda = 1$ vs. $\frac{1}{2}$, one state | $\geq \mathrm{err}_{1/2}^*$ | 0.109 to 0.120 |

powers of $R$; no radial time change therefore reduces their sum at a fixed scale to the family the clocks need. To turn this loss of factorization into a statement about labels, the proof tracks the rival stable boundary to first order in the finite-scale correction. At zero scale the clocks make the rival boundary at $\omega = 2$ invariant under reflecting the third character; its first-order correction in the logit scale $s^3$ has a nonzero reflection-odd part. Reflected states therefore fall on opposite sides of the exact-CE stable boundary and enter different pure-character traps, although the reflection preserves every sector mass and both rivals' blocks (Appendix E). The obstruction persists under retuning within a fixed compact set: the weight and the scale move two reflection-odd boundary coefficients in independent directions, so no weight near $\omega = 2$ cancels both, and away from $\omega = 2$ the leading coefficient is already nonzero.

## 6 What the measurements show

Table 1 separates the optimal sector-local error $\mathrm{err}_J^*$ from four empirical quantities that bound it or compare different selectors. The paired *joint-versus-separate disagreement* $q_{\mathrm{op}}$ draws one Gaussian triple, normalizes it both ways, $X_J = \left((A,B)/\left\|(A,B)\right\|, C/\left\|C\right\|\right)$ and $X_S = T(X_J)$ with $T(a,b,c) = (a/\left\|a\right\|, b/\left\|b\right\|, c)$, and asks whether the ordinary joint and separate leading selectors disagree; this changes the normalized start as well as the metric. The *metric-only disagreement* $q_{\mathrm{met}}$ instead holds a separately normalized state fixed and changes only the metric, from $\lambda = 1$ to $\lambda = \frac{1}{2}$, the pullbacks of $\omega = 2$ and $\omega = 1$. Both are selector-disagreement probabilities rather than estimates of sector-local prediction error. Because the separate clocks applied after $T$ form a sector-local rule, $\mathrm{err}_J^* \leq q_{\mathrm{op}}$ and $\mathrm{err}_{1/2}^* \leq q_{\mathrm{met}}$ (Remark C.2).

**How close independent ranking comes.** Lifting the ordinary joint flow as at $\omega = 2$ leaves its sectors coupled only through an output-to-input rate ratio that reads the total input radius. Freezing it at its initial value gives independent races on each sector's raw initial blocks, a sector-local rule (Appendix H.3). On the $C_2 \times C_2$ draws behind $q_{\mathrm{op}}$ it disagrees with $S_J$ on 340 of 20,000 (1.70%, Wilson 95% interval 1.52 to 1.89%), against 3,686 for the comparator. On $C_2 \times C_2$ the optimum on each mass fiber is a finite optimization: given the masses, the label depends only on the three product signs, so the best rule is the best of 720 orderings of six sign-dependent scores. Averaging this fiberwise optimum over 40,000 sampled mass fibers gives $\widehat{\mathrm{err}}_J^* \approx 8.1 \times 10^{-4}$, with Monte Carlo standard error $0.5 \times 10^{-4}$ (Appendix H.3). The *reflected-witness* set $W_j$, of states whose label changes when sector $j$'s output block is reflected, with neither label equal to $j$, gives the complementary lower bound $\frac{1}{2}\mu_J(W_j) \leq \mathrm{err}_J^*$ for every reflected sector $j$ (Remark C.1), and Appendix H.2 estimates these volumes. On this distribution every sector-local rule errs and the simple raw-block rule is already accurate. Thus small empirical error is compatible with failure of exact sector-locality.

**Paired effects, equal frequencies.** On $C_2 \times C_2$, $GL(2, \mathbb{F}_2) \cong S_3$ permutes the three characters while preserving both metrics, the potential and the coupling, so each character has probability $1/3$ under both leading selectors: the label distributions agree exactly while a fraction $q_{\mathrm{op}} > 0$ of paired draws disagree (Corollary D.2). On $D_4$ and $Q_8$ the largest empirical total variation between the two label histograms was 0.0046, which does not establish equality. Paired-start agreement and agreement of the marginal label distributions are therefore distinct statistics.

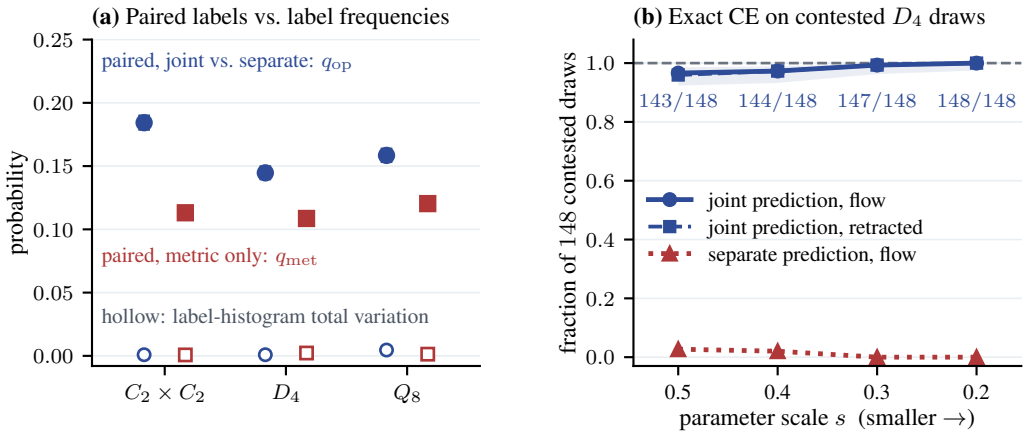


Figure 2: **Paired labels differ far more often than label frequencies do, and small-scale exact CE follows the joint prediction. (a)** Width-one disagreement over 20,000 paired draws per group for the joint-versus-separate rate $q_{\mathrm{op}}$ and the metric-only rate $q_{\mathrm{met}}$; Wilson 95% intervals (Table 2) are narrower than the markers. Hollow markers show total variation between the corresponding label histograms. These selector disagreements bound the optimal sector-local errors from above (Table 1). **(b)** Among the 148 of 1000 $D_4$ draws whose leading labels disagree, the fraction whose finite-horizon exact-CE label follows each leading prediction as $s$ decreases. Counts shown are for the flow. One draw at each of the three larger scales follows neither prediction. The band is the Wilson 95% interval of the flow's joint agreement.

**Exact CE follows the joint prediction.** Among 1000 further $D_4$ draws, the 148 whose two leading labels differ are the only ones that can show which selector exact CE follows; this conditional summary was chosen after the run. At the protocol's fixed horizon $\tau \approx 60$, continuous and retracted CE carried the joint label on $143/148$ and $142/148$ of them at $s = 0.50$ and on $148/148$ at $s = 0.20$ (Figure 2(b)). Over all 1000 draws at $s = 0.20$ the counts were $999/1000$ and $1000/1000$; the one continuous exception persists under refinement at that scale and carries the joint label at the smaller scales tested (Appendix H.1). These are finite-horizon measurements; Theorem 4.2 and Corollary D.1 concern infinite-time endpoints in the small-scale limit.

## 7 Related Work

**Selector formulas and the compensating factor.** He et al. (2026b, v2, Theorem 5.3(ii)) identify the initially dominant representation for their separately normalized projected-sphere Stage-I flow, on Abelian groups with tied inputs, matched scales, aligned phases and no nontrivial self-conjugate characters; He et al. (2026a, Corollary 6.1) identify, with high probability, the initially least-misaligned frequency under equal initial magnitudes, for an unconstrained leading flow on a cyclic quadratic model. Both rank a representation from its own coefficients and aggregate magnitudes, the shape of our class, and each works under its hypotheses; Theorem 3.1 asks whether any rule of that form works at each fixed weight of a metric family. Their tied-input flow already halves the input gradient to compensate for the shared embedding (He et al., 2026b, v2, Section 5), which is the matched rate for one sphere carrying two factors; we use the same rate on the untied joint sphere. Theorem 3.1 supplies the converse, a positive error floor for every measurable rule of this form at every other fixed weight, and extends both directions to untied inputs on the full joint sphere; Sections 4–5 treat exact CE. Their Stage-I flow provides the separate-normalization comparator used here (Appendix A.7).

**Radial removal and metric-dependent basins.** Common radial removal, with the survivor selected by the minimum of subsystem-local quantities, is classical in homogeneous competition (Hofbauer, 2002); for real scalar sectors our clocks are, after fixed rescalings, the blow-up times of the three-factor system of Zhao et al. (2026, Section 5.1), whose distance-speed argument already selects the feature that blows up first, and the conservation law behind the matched weight is the layer balance

of homogeneous networks (Du et al., 2018; Kunin et al., 2021). Metric-dependent basins for a fixed potential are classical (Fiedler et al., 1999; Belabbas, 2023), as are dynamics-dependent separatrices (Sakos et al., 2024) and the Riemannian account of parameterization dependence (Kristiadi et al., 2023; Lawrence et al., 2022; Marcotte et al., 2023; Li et al., 2022; Achour et al., 2026; Morwani & Ramaswamy, 2022; Cao et al., 2023). The distinction here is the quantifier: a moved boundary compares two flows, whereas Theorem 3.1 quantifies over every measurable rule in a stated class. The failure is informational: the basins are open and a full-state predictor exists.

**Representations and endpoints.** Which representations trained group networks contain, and which are favored, is well documented (Chughtai et al., 2023; Stander et al., 2024; Morwani et al., 2024; Wu et al., 2025; Marchetti et al., 2024; Tian, 2025); under stated assumptions an encoding-dependent acquisition order is proved (Marchetti et al., 2026), and alternating gradient flows give the order in which small-initialization quadratic networks learn Fourier features for modular addition (Kunin et al., 2025). These characterize circuits, optimization targets and orders. Ding et al. (2024) find empirically that frequencies with high initial signal and gradients are more likely to survive in modular addition, and model their competition with linear dynamics; we ask when independent scores determine the survivor exactly. Shutman et al. (2025) also decouple aligned sectors over all three real algebras, but in a squared-loss Hadamard model whose tensor rank is not the Fourier-block rank, with neither normalization metric. Finite-time approximation followed by local analytic capture is an established route to endpoint statements (Li et al., 2021, Theorems 5.6 and 5.8), with standard persistence tools (Shub, 1987; Li, 1999); the model must supply its geometry, the Morse–Bott collars of the real-block maxima, the null basin of the flat stratum and a step-length bound for retracted descent.

## 8 Limitations

*Model and regime.* The analysis fixes a finite group and its full multiplication table, quadratic activation, full-batch training and exact real arithmetic; stochastic and momentum dynamics, other normalizations and finite precision lie outside this setting. At fixed parameter norms, $s \downarrow 0$ sends the logits to zero and the loss to $N^2 \log N$. The theorems concern representation selection in this small-logit regime and give no low-loss, accuracy, generalization or training-time guarantee (Appendix D.3); a sector mismatch may leave the learned function unchanged.

*The characterization.* Theorem 3.1 is restricted to sector-local predictors. Its floors extend to initialization laws equivalent to volume; absolute continuity alone is insufficient, since a law supported in one basin admits a perfect constant rule. The floors depend on the law and on the weight, with no bound uniform over such laws or as $\omega \to 2$ (Remarks C.3 and C.5).

*Exact CE.* Whole-vector persistence fixes the width; only the expected fraction of changed labels is uniform over widths. Outside the hypotheses of Theorem 5.1, for other sector triples, larger widths, retracted descent or weights tending to 0 or $\infty$, exact sector-locality at positive scale is open, and no value or rate is proved for its floor. The measurements are finite-horizon reads on three groups, and $\mathrm{err}_J^*$ and the raw-block rate are post hoc estimates from numerical labels on $C_2 \times C_2$ only.

## 9 Conclusion

Normalization can couple representation selection even when the leading objective is additive across sectors. The sphere constraints feed the total potential back to every candidate, so a sector that wins at neither of two starts can decide between two others. Matching metric weights to the network's degrees reduces this coupling to a common time change in the leading flow; with three or more nontrivial real sectors, every other fixed weight leaves a positive sector-local error floor. Under ordinary normalization, exact cross entropy inherits the leading winners at its infinite-time endpoints as the scale vanishes. At width one and $\omega = 2$, sector-locality returns in that limit, but with three distinct nontrivial real linear characters the optimal sector-local error of continuous exact CE is positive at every sufficiently small positive scale, including under compactly retuned weights. On XOR at $\omega = 1$, a simple sector-local rule errs on $1.7\%$ of starts, while the jointly and separately normalized leading selectors have identical label frequencies and still disagree on paired starts. These results separate approximate prediction and agreement of label frequencies from the stronger claim that representations compete independently.

## REPRODUCIBILITY STATEMENT

Appendices A–E contain the complete proof chain for every numbered result in the main text. Appendices A and B establish the real Fourier landscape and almost-sure convergence of the leading flow; Appendix C proves the sector-local characterization and its joint-manifold form; Appendix D proves exact-CE endpoint persistence; and Appendix E proves the positive-scale failure of exact sector-locality, including compactly retuned weights. Appendix F gives extensions cited after the main theorems, Appendix G records comparisons and corrections, and Appendix H gives the complete experimental protocols and numerical results, distinguishing prespecified analyses from later exploratory ones.

Code, machine-readable outputs, and reproduction commands accompany the submission as supplementary material. The manuscript itself states all assumptions, experimental settings, resolution rules, numerical checks, and reported summaries used in the paper, so it is self-contained; the supplement serves reproduction. Group laws, representation homomorphisms, character orthogonality, and Frobenius–Schur types were checked exhaustively, and every Fourier transform received a real spatial-domain input.

## AI USE STATEMENT

In this work, we used generative AI tools to formulate and refine the mathematical claims, draft proofs, refine hypotheses, implement the analysis and experiment code, and interpret results. We also used them to search for and summarize related literature, draft sections of the paper, and edit the writing and references. We verified every claim, proof and computation before including it in the paper. The authors take full responsibility for all content, including every proof.

## ETHICS STATEMENT

This work uses synthetic finite-group tasks and no human-subject, personal, or sensitive data.

## REFERENCES

Alberto Abbondandolo and Pietro Majer. On the global stable manifold. *Studia Mathematica*, 177: 113–131, 2006. URL https://arxiv.org/abs/math/0506292v2.

Pierre-Antoine Absil, Robert Mahony, and Ben Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005. doi: 10.1137/040605266. URL https://sites.uclouvain.be/absil/2005.901. See the authors' online erratum to Proposition 3.3.

El Mehdi Achour, Kathlén Kohn, and Holger Rauhut. The Riemannian geometry associated to gradient flows of linear convolutional networks. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 946–954. PMLR, 2026. URL https://proceedings.mlr.press/v300/achour26a.html.

Bubacarr Bah, Holger Rauhut, Ulrich Terstiege, and Michael Westdickenberg. Learning deep linear neural networks: Riemannian gradient flows and convergence to global minimizers. *Information and Inference: A Journal of the IMA*, 11(1):307–353, 2022. doi: 10.1093/imaiai/iaaa039. URL https://arxiv.org/abs/1910.05505v5. Theorem numbering follows arXiv:1910.05505v5.

Mohamed-Ali Belabbas. On gradient flows initialized near maxima. *SIAM Journal on Control and Optimization*, 61(5):2826–2848, 2023. doi: 10.1137/21M1450756.

Yuan Cao, Difan Zou, Yuanzhi Li, and Quanquan Gu. The implicit bias of batch normalization in linear models and two-layer linear convolutional neural networks. In *Proceedings of the 36th Conference on Learning Theory*, volume 195 of *Proceedings of Machine Learning Research*, pp. 5699–5753. PMLR, 2023. URL https://proceedings.mlr.press/v195/cao23a.html.

Bilal Chughtai, Lawrence Chan, and Neel Nanda. A toy model of universality: Reverse engineering how networks learn group operations. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6243–6267. PMLR, 2023. URL https://proceedings.mlr.press/v202/chughtai23a.html.

Tobias H. Colding and William P. Minicozzi, II. Łojasiewicz inequalities and applications. *Surveys in Differential Geometry*, 19:63–82, 2014. doi: 10.4310/SDG.2014.v19.n1.a3.

Xiaoman Delores Ding, Zifan Carl Guo, Eric J. Michaud, Ziming Liu, and Max Tegmark. Survival of the fittest representation: A case study with modular addition. *arXiv preprint arXiv:2405.17420*, 2024. URL https://arxiv.org/abs/2405.17420v1.

Simon S. Du, Wei Hu, and Jason D. Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. In *Advances in Neural Information Processing Systems*, volume 31, 2018. URL https://papers.nips.cc/paper/7321-algorithmic-regularization-in-learning-deep-homogeneous-models-layers-are-automatically-balanced.

Bernold Fiedler, Mohamed Belhaq, and Mohamed Houssni. Basins of attraction in strongly damped coupled mechanical oscillators: A global example. *Zeitschrift für angewandte Mathematik und Physik*, 50:282–300, 1999. doi: 10.1007/s000330050151.

Andrey Gromov. Grokking modular arithmetic. *arXiv preprint arXiv:2301.02679*, 2023. URL https://arxiv.org/abs/2301.02679v1.

Jianliang He, Leda Wang, Siyu Chen, and Zhuoran Yang. On the mechanism and dynamics of modular addition: Fourier features, lottery ticket, and grokking. *arXiv preprint arXiv:2602.16849*, 2026a. URL https://arxiv.org/abs/2602.16849v1.

Jianliang He, Leda Wang, Fengzhuo Zhang, Siyu Chen, and Zhuoran Yang. Neural networks provably learn spectral representations for group composition. *arXiv preprint arXiv:2606.02993*, 2026b. URL https://arxiv.org/abs/2606.02993v2.

Josef Hofbauer. Competitive exclusion of disjoint hypercycles. *Zeitschrift für Physikalische Chemie*, 216:35–39, 2002. URL https://homepage.univie.ac.at/josef.hofbauer/02zpc.pdf.

Agustinus Kristiadi, Felix Dangel, and Philipp Hennig. The geometry of neural nets' parameter spaces under reparametrization. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2302.07384.

Daniel Kunin, Javier Sagastuy-Brena, Surya Ganguli, Daniel L. K. Yamins, and Hidenori Tanaka. Neural mechanics: Symmetry and broken conservation laws in deep learning dynamics. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=q8qLAbQBupm.

Daniel Kunin, Giovanni Luca Marchetti, Feng Chen, Dhruva Karkada, James B. Simon, Michael R. DeWeese, Surya Ganguli, and Nina Miolane. Alternating gradient flows: A theory of feature learning in two-layer neural networks. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2506.06489v4.

Hannah Lawrence, Kristian Georgiev, Andrew Dienes, and Bobak T. Kiani. Implicit bias of linear equivariant networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12096–12125. PMLR, 2022. URL https://proceedings.mlr.press/v162/lawrence22a.html.

Ming-Chia Li. Structural stability for the Euler method. *SIAM Journal on Mathematical Analysis*, 30 (4):747–755, 1999. doi: 10.1137/S0036141097322558.

Zhiyuan Li, Yuping Luo, and Kaifeng Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2012.09839.

Zhiyuan Li, Tianhao Wang, Jason Lee, and Sanjeev Arora. Implicit bias of gradient descent on reparametrized models: On equivalence to mirror descent. In *Advances in Neural Information Processing Systems*, volume 35, pp. 34626–34640, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/dfa1106ea7065899b13f2be9da04efb4-Abstract-Conference.html.

Giovanni Luca Marchetti, Christopher J. Hillar, Danica Kragic, and Sophia Sanborn. Harmonics of learning: Universal Fourier features emerge in invariant networks. In *Proceedings of the Thirty Seventh Conference on Learning Theory*, volume 247 of *Proceedings of Machine Learning Research*, pp. 3775–3797. PMLR, 2024. URL https://proceedings.mlr.press/v247/marchetti24a.html.

Giovanni Luca Marchetti, Daniel Kunin, Adele Myers, Francisco Acosta, and Nina Miolane. Sequential group composition: A window into the mechanics of deep learning. *arXiv preprint arXiv:2602.03655*, 2026. URL https://arxiv.org/abs/2602.03655v2. Accepted at ICML 2026; cited version arXiv:2602.03655v2.

Sibylle Marcotte, Rémi Gribonval, and Gabriel Peyré. Abide by the law and follow the flow: Conservation laws for gradient flows. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/c7bee9b76be21146fd592fc2b46614d5-Abstract-Conference.html.

Depen Morwani and Harish G. Ramaswamy. Inductive bias of gradient descent for weight normalized smooth homogeneous neural nets. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pp. 827–880. PMLR, 2022. URL https://proceedings.mlr.press/v167/morwani22a.html.

Depen Morwani, Benjamin Edelman, Costin-Andrei Oncescu, Rosie Zhao, and Sham Kakade. Feature emergence via margin maximization: Case studies in algebraic tasks. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=i9wDX850jR.

Iosif Sakos, Stefanos Leonardos, Stelios Andrew Stavroulakis, William Overman, Ioannis Panageas, and Georgios Piliouras. Beating price of anarchy and gradient descent without regret in potential games. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/921ac785fa9edc73cacaf2664f43d234-Paper-Conference.pdf.

Jean-Pierre Serre. *Linear Representations of Finite Groups*, volume 42 of *Graduate Texts in Mathematics*. Springer, New York, 1977. doi: 10.1007/978-1-4684-9458-7.

Michael Shub. *Global Stability of Dynamical Systems*. Springer-Verlag, New York, 1987. doi: 10.1007/978-1-4757-1947-5.

Maor Shutman, Oren Louidor, and Ran Tessler. Learning words in groups: Fusion algebras, tensor ranks and grokking. *arXiv preprint arXiv:2509.06931*, 2025. URL https://arxiv.org/abs/2509.06931v1.

Dashiell Stander, Qinan Yu, Honglu Fan, and Stella Biderman. Grokking group multiplication with cosets. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46441–46467. PMLR, 2024. URL https://proceedings.mlr.press/v235/stander24a.html.

Yuandong Tian. Composing global solutions to reasoning tasks via algebraic objects in neural nets. In *Advances in Neural Information Processing Systems*, 2025. URL https://arxiv.org/abs/2410.01779v4.

Louise A. Wolf. Similarity of matrices in which the elements are real quaternions. *Bulletin of the American Mathematical Society*, 42:737–743, 1936. doi: 10.1090/S0002-9904-1936-06417-7.

Wilson Wu, Louis Jaburi, Jacob Drori, and Jason Gross. Towards a unified and verified understanding of group-operation networks. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.07476v3.

Fuzhen Zhang. Quaternions and matrices of quaternions. *Linear Algebra and its Applications*, 251: 21–57, 1997. doi: 10.1016/0024-3795(95)00543-9.

Jiajie Zhao, Jianxing Wang, Junjie Yang, Zhiwei Bai, and Yaoyu Zhang. Gradient flow dynamics and implicit bias of diagonal linear networks under infinitesimal initialization. *arXiv preprint arXiv:2607.12332*, 2026. URL https://arxiv.org/abs/2607.12332v1.

# PART I CORE PROOF CHAIN

## A PROOFS FOR THE REAL FOURIER REDUCTION AND LANDSCAPE

This appendix develops the real Fourier reduction and proves Theorem A.1 using the spatial convention (8). The trace below is the ordinary complex trace after embedding each real block in $M_{d_\rho}(\mathbb{C})$.

The route is indirect by design. We first solve the uncentered real-form problem $\mathcal{E} = \Omega$, whose block structure is algebraically clean: its critical points, sector ceilings and Hessian inertia can be read off the Wedderburn decomposition directly. The centered network potential $f = \Omega(Pa, Pb, Pc)$ then factors through $\mathcal{E}$ on every nontrivial sector, by (37), so the maxima, equality cases and open trapping neighborhoods established for $\mathcal{E}$ transfer to $f$ in Appendix A.7. Working with $f$ from the start would carry the trivial-sector projector through every block computation without changing any of these conclusions.

### A.1 FOURIER CONVENTIONS

For $d_\rho^\sharp = d_\rho |\operatorname{Orb}(\rho)|$, the convention is

$$\widehat{a}_\rho = N^{-1} \sum_g a_g \rho(g)^*, \tag{8}$$

$$a_g = \sum_\rho d_\rho^\sharp \operatorname{Re}\operatorname{tr}(\widehat{a}_\rho \rho(g)), \tag{9}$$

$$\langle a, b \rangle = \sum_\rho N d_\rho^\sharp \operatorname{Re}\operatorname{tr}(\widehat{a}_\rho^* \widehat{b}_\rho), \tag{10}$$

$$\Omega(a, b, c) = \sum_\rho d_\rho^\sharp \operatorname{Re}\operatorname{tr}(\widehat{c}_\rho^* \widehat{b}_\rho \widehat{a}_\rho). \tag{11}$$

For $\kappa_\rho = 1, 1, 2$ in real, complex, and quaternionic type,

$$\operatorname{Re}\operatorname{tr}(C^* B A) \le \frac{\|A\|_F \|B\|_F \|C\|_F}{\sqrt{\kappa_\rho}}. \tag{12}$$

For nonzero $A, B, C$, equality holds exactly at cross-layer-aligned, intrinsic-$D_\rho$-rank-one triples; if a factor is zero, equality is automatic. The corresponding sector ceiling is

$$\gamma_\rho = \frac{1}{N^{3/2} \sqrt{d_\rho^\sharp \kappa_\rho}}. \tag{13}$$

Choose one unitary irrep $\rho$, of complex degree $d_\rho$, from each complex-conjugation orbit, in a Frobenius–Schur-adapted basis when $\rho$ is self-dual. The real Fourier transform $\widehat{a}_\rho =$

$N^{-1}\sum_g a_g\rho(g)^*$ identifies $\mathbb{R}^G$ isometrically with $H = \bigoplus_\rho \mathcal{A}_\rho$, the blocks $\mathcal{A}_\rho = M_{n_\rho}(D_\rho)$ of (5), for the weighted inner product $\langle A, B\rangle_H = \sum_\rho N d_\rho^\sharp \operatorname{Re}\operatorname{tr}(A_\rho^* B_\rho)$ whose weights are forced by spatial normalization.

Intrinsic rank means rank over $D_\rho$, so a nonzero quaternionic block of intrinsic rank one has ordinary complex rank two. With $d_\rho^\sharp = d_\rho|\operatorname{Orb}(\rho)|$, the uncentered auxiliary form is

$$\mathcal{E}(A, B, C) = \sum_\rho d_\rho^\sharp \operatorname{Re}\operatorname{tr}(C_\rho^* B_\rho A_\rho) \quad \text{on } S(H)^3. \tag{14}$$

The form $\mathcal{E}$ includes the trivial sector. The centered joint landscape is treated in Appendix B.

**Theorem A.1** (Fixed-real-form landscape)**.** *On $S(H)^3$, the positive local maxima of (14) are exactly single-sector, cross-layer-aligned, intrinsic-$D_\rho$-rank-one states. Every other critical point, at every energy sign, is a strict saddle for ascent. The sector ceiling is $\gamma_\rho = (N^{3/2}\sqrt{d_\rho^\sharp \kappa_\rho})^{-1}$, where $\kappa_\rho = 1, 1, 2$ for real, complex, and quaternionic type.*

### A.2 The transform and its forced weights

This calculation fixes the metric weights the Fourier transform induces; changing them would change the normalization being compared.

For all complex irreducibles, the regular-character identity is

$$\sum_{\chi \in \operatorname{Irr}_{\mathbb{C}}(G)} d_\chi \chi(x) = N \mathbf{1}_{\{e\}}(x). \tag{15}$$

Expanding the Fourier coefficients and using unitarity gives

$$\sum_\chi d_\chi \operatorname{tr}(\widehat{a}_\chi^* \widehat{b}_\chi) = N^{-1} \langle a, b\rangle, \tag{16}$$

$$\sum_\chi d_\chi \operatorname{tr}(\widehat{a}_\chi \rho_\chi(g)) = a_g. \tag{17}$$

For a real signal, the contributions of two non-self-conjugate irreducibles are conjugate scalars. Collapsing their orbit changes $d_\rho$ into $d_\rho^\sharp = 2d_\rho$ and takes a real part. A self-conjugate block has real complex trace. This proves (9) and (10).

The Frobenius–Schur trichotomy realizes a self-dual representation in $M_d(\mathbb{R})$ or $M_{d/2}(\mathbb{H})$, while a non-self-dual pair contributes $M_d(\mathbb{C})$ (Serre, 1977). Their real dimensions are, uniformly,

$$\dim_{\mathbb{R}} \mathcal{A}_\rho = |\operatorname{Orb}(\rho)|\, d_\rho^2. \tag{18}$$

These dimensions sum to $N$. Since the Fourier map is injective by (17), it is a real-linear bijection from $\mathbb{R}^G$ onto the displayed direct sum of real forms.

For the multiplication tensor, expansion over all complex irreducibles gives

$$\operatorname{tr}(\widehat{c}_\chi^* \widehat{b}_\chi \widehat{a}_\chi) = N^{-3} \sum_{k,h,g} c_k b_h a_g \chi(kh^{-1}g^{-1}). \tag{19}$$

Weighting by $d_\chi$ and applying (15) leaves $k = gh$ and a factor $N$. Orbit collapse proves (11).

There is no remaining choice in either weight. Surjectivity supplies a real signal supported on any chosen block. Applying a proposed metric identity to that signal forces $m_\rho = N d_\rho^\sharp$; inserting three such signals in the trilinear identity forces $w_\rho = d_\rho^\sharp$.

### A.3 The sector bound and the spatial spectral norm

Two things are needed: the largest value attainable in each sector, and the configurations that attain it. The first fixes the ceilings; the second is what makes the maxima aligned and of intrinsic rank one.

**Lemma A.2** (Sharp block inequality over $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{H}$)**.** *Let $D \in \{\mathbb{R}, \mathbb{C}, \mathbb{H}\}$, let $\kappa_D = 1, 1, 2$, respectively, and use the standard complex embedding of $M_n(D)$. For nonzero $A, B, C \in M_n(D)$,*

$$\operatorname{Re}\operatorname{tr}(C^* BA) \le \frac{\|A\|_F \|B\|_F \|C\|_F}{\sqrt{\kappa_D}}. \tag{20}$$

*Equality holds exactly when there are unit $D$-columns $p, q, r$ and positive scalars $\alpha, \beta, \gamma$ such that*

$$A = \alpha q r^*, \qquad B = \beta p q^*, \qquad C = \gamma p r^*.$$

*In particular, all three blocks have intrinsic $D$-rank one and are cross-layer aligned.*

*Proof.* All traces here are ordinary complex traces, and the singular-value multiplicity is derived directly; Wolf (1936) and Zhang (1997) provide classical quaternionic matrix background. For a quaternionic Hermitian matrix $H$, the real Rayleigh quotient attains its maximum on the unit sphere. Its real first variation gives $Hv = v\lambda$ with $\lambda \in \mathbb{R}$. The quaternionic orthogonal complement of $v$ is $H$-invariant, since $v^* Hw = (Hv)^* w = \lambda v^* w$. Induction gives a quaternionic orthonormal eigenbasis. Apply this to $B^* B$: its eigenvalues are nonnegative, and the vectors $Bv_i/\sigma_i$ for $\sigma_i > 0$ are orthonormal. Complete them to an orthonormal basis; the zero eigenvectors belong to $\ker B$. Thus $B = U\Sigma V^*$ with quaternionic unitary factors and real nonnegative diagonal $\Sigma$. The standard complex star-embedding gives

$$\iota(B) = \iota(U)\,\mathrm{diag}(\Sigma, \Sigma)\iota(V)^*,$$

whose outer factors are complex unitary. Each intrinsic singular value therefore occurs twice, and complex rank is twice quaternionic rank. The real and complex cases have multiplicity one. In all cases the multiplicity is $\kappa_D$, so

$$\|B\|_{\mathrm{op}} \le \|B\|_F / \sqrt{\kappa_D}, \tag{21}$$

with equality exactly when $B$ has intrinsic $D$-rank one. Real Frobenius Cauchy–Schwarz and the operator–Frobenius inequality give

$$\operatorname{Re}\operatorname{tr}(C^* BA) \le \|C\|_F \|BA\|_F \le \|C\|_F \|B\|_{\mathrm{op}} \|A\|_F,$$

which proves (20). At equality, write $B = \beta p q^*$ using the $D$-singular-value decomposition. Equality in the middle inequality forces the range of $A$ into the right singular $D$-line of $B$, hence $A = \alpha q r^*$ after absorbing a unit $D$-scalar into $r$. Equality in real Frobenius Cauchy–Schwarz forces $C$ to be a positive real multiple of $BA$, hence $C = \gamma p r^*$. The converse follows by substitution. For $D = \mathbb{H}$, the multiplicity two above is already the entire quaternionic factor; no additional trace weight appears. □

After the Fourier metric is imposed, Lemma A.2 becomes (12) and gives (13).

There is also a direct spatial proof of the global ceiling. For unit $a, c$, put $r(h) = \sum_g a_g c_{gh}$. Cauchy–Schwarz gives $|r(h)| \le 1$, hence

$$\sum_h b_h r(h) \le \|r\|_2 \le \sqrt{N}. \tag{22}$$

Equality requires every right translate of $c$ to equal $a$ up to sign. Comparing translations makes these signs a homomorphism $G \to \{+1, -1\}$. Thus the unnormalized multiplication tensor has real spectral norm $\sqrt{N}$, and the $N^{-2}$ normalization gives $\max_{(S^{N-1})^3} \Omega = N^{-3/2}$, attained exactly by normalized real linear characters with layer signs of positive product.

### A.4 Critical rigidity

It is enough to work in the general block algebra of (14). Sphere criticality and Euler's identity give

$$\frac{w_\rho}{m_\rho} B_\rho^* C_\rho = E A_\rho, \qquad \frac{w_\rho}{m_\rho} C_\rho A_\rho^* = E B_\rho, \qquad \frac{w_\rho}{m_\rho} B_\rho A_\rho = E C_\rho. \tag{23}$$

Suppose $E > 0$, fix an active sector, and write $\phi = mE/w$. Right-multiplying the first critical equation by $A^*$ and using the second yields $B^* B = AA^*$. Adjointing the third equation, multiplying by $B$, and using the first gives

$$A^* AA^* = \phi^2 A^*, \qquad AA^* A = \phi^2 A. \tag{24}$$

Thus $A/\phi$ is a partial isometry. The same argument makes $B/\phi$ a partial isometry, and $C = BA/\phi$ aligns the corresponding initial and final spaces. All three layers share an intrinsic rank $r$. Their metric norm in one layer is $m\kappa r\phi^2$. Summing over active sectors and using the unit sphere constraint yields

$$E = \left( \sum_{\rho \in S} \frac{\kappa_\rho r_\rho m_\rho^3}{w_\rho^2} \right)^{-1/2}. \tag{25}$$

The compact unitary groups of $M_n(D)$ act by

$$(A_\rho, B_\rho, C_\rho) \longmapsto (P_\rho A_\rho Q_\rho^*, R_\rho B_\rho P_\rho^*, R_\rho C_\rho Q_\rho^*). \tag{26}$$

This action and singular-value decompositions over $\mathbb{R}$, $\mathbb{C}$, and $\mathbb{H}$ simultaneously put the three slots of a positive critical point in diagonal form. In the metric-orthonormal diagonal basis, they coincide with a vector $z$ satisfying

$$\sum_i z_i^2 = 1, \qquad \gamma_i z_i = E. \tag{27}$$

If at least two directions are active, choose a unit $h \perp z$ on their span and set every layer equal to $z\cos t + h \sin t$. Expanding the cubic along this curve gives

$$\left. \frac{\mathrm{d}^2 \mathcal{E}}{\mathrm{d}t^2} \right|_{t=0} = 6 \sum_i \gamma_i z_i h_i^2 - 3E = 3E > 0. \tag{28}$$

The squared speed is three, so the Hessian has a positive unit-direction eigenvalue. This handles both higher rank inside one sector and support on several sectors.

### A.5 Why the remaining positive points are maxima

For an arbitrary point, let $a_j = \sqrt{m_j}\,\|A_j\|_F$ and define $b_j, c_j$ cyclically. Equation (20) gives the global sector budget

$$\mathcal{E}(A, B, C) \le \sum_j \gamma_j a_j b_j c_j. \tag{29}$$

Fix a one-sector rank-one critical point in sector $j_0$. Let $\epsilon, \eta, \zeta$ be the three off-sector layer norms and put $\sigma^2 = \epsilon^2 + \eta^2 + \zeta^2$. The active mass product obeys

$$\sqrt{1-\epsilon^2}\sqrt{1-\eta^2}\sqrt{1-\zeta^2} \le 1 - \frac{\sigma^2}{2} + \frac{\sigma^4}{12}, \tag{30}$$

while Hölder and arithmetic–geometric mean give

$$\sum_{j \neq j_0} a_j b_j c_j \le \epsilon\eta\zeta \le \frac{\sigma^3}{3^{3/2}}. \tag{31}$$

With $\Gamma = \max_j \gamma_j$, subtraction of the critical value gives

$$\mathcal{E} - \gamma_{j_0} \le -\sigma^2 \left[ \frac{\gamma_{j_0}}{2} - \sigma \left( \frac{\Gamma}{3^{3/2}} + \frac{\gamma_{j_0}}{12} \right) \right]. \tag{32}$$

This is negative for sufficiently small positive $\sigma$. On the single-sector locus, equality in (20) holds exactly on the aligned rank-one gauge orbit. The orbit is therefore locally maximal despite the Hessian's gauge kernel.

### A.6 Zero and negative critical values

At any critical point, differentiating the constrained Lagrangian gives

$$\operatorname{Hess} \mathcal{E}(v, v) = 2\{\mathcal{E}(\dot{A}, \dot{B}, C) + \mathcal{E}(\dot{A}, B, \dot{C}) + \mathcal{E}(A, \dot{B}, \dot{C})\} - E\,\|v\|^2. \tag{33}$$

If $E < 0$, every nonzero tangent direction that moves one layer has curvature $-E\,\|v\|^2 > 0$.

At $E = 0$, the critical equations say $B_j^* C_j = C_j A_j^* = B_j A_j = 0$ in every block. Let $U$ be the unit of the direct-sum algebra, with its metric normalization understood, and set

$$\nu = \sqrt{\|U\|^2 - \langle U, A\rangle^2}, \qquad \dot{A} = \frac{U - \langle U, A\rangle A}{\nu}, \quad \dot{B} = 0, \quad \dot{C} = B. \tag{34}$$

The denominator cannot vanish: otherwise $A$ would be a normalized unit, contradicting $BA = 0$ and $\|B\| = 1$. The vector is tangent, and substitution in (33) gives

$$\frac{\operatorname{Hess} \mathcal{E}(v, v)}{\|v\|^2} = \nu^{-1} \sum_j w_j \|B_j\|_F^2 > 0. \tag{35}$$

Thus every nonpositive critical point is a strict saddle. Together with the positive analysis, this proves Theorem A.1.

### A.7 The Stage-I objective, centering, and the transfer of the traps

Three separate things happen here. First the published separate Stage-I objective is identified with the leading potential of the main text. Then the traps of the auxiliary uncentered problem are transferred to the centered separate and joint constraints. Finally the weighted-table remark records which data weightings admit the exact identity, which is a scope statement rather than a step in either argument.

The Stage-I objective of He et al. (2026b) is the first-order expansion of cross entropy at zero logits over the same quadratic architecture: the correct logit is subtracted and the mean logit added back, at a fixed output scale $A > 0$, with all $3M$ parameter vectors on unit spheres. Writing $F_k^{(g,h)} = \frac{A}{M} \sum_m (c_m)_k (a_{m,g} + b_{m,h})^2$ and

$$\mathcal{R}(X) = -\sum_{g,h} F_{gh}^{(g,h)} + \frac{1}{N} \sum_{g,h} \sum_k F_k^{(g,h)} + N^2 \log N,$$

centering the output turns the pair sum into $-\frac{A}{M} \sum_m \sum_{g,h} (Pc_m)_{gh} (a_{m,g} + b_{m,h})^2$. Expanding the square, the $a_{m,g}^2$ term carries $\sum_h (Pc_m)_{gh} = 0$ for each fixed $g$ and the $b_{m,h}^2$ term carries $\sum_g (Pc_m)_{gh} = 0$ for each fixed $h$, because left and right translation are bijections of $G$ and $Pc_m$ sums to zero. Only the cross term survives, giving $2N^2 \Omega(a_m, b_m, Pc_m)$, so

$$\mathcal{R}(X) = N^2 \log N - \frac{2AN^2}{M} \mathcal{F}(X) \tag{36}$$

identically in $X$, with no constraint on any layer and without using associativity. The two sides differ by an additive constant and a positive factor and $\mathcal{F}$ is a sum over neurons, so the sphere-projected gradient flow of $\mathcal{R}$ decouples across neurons and is, in each neuron, the projected gradient ascent flow of $f$ on $(S^{N-1})^3$ under the time change $t \mapsto (2AN^2/M)t$, with the same orbits, equilibria and $\omega$-limit sets. Both initialization laws are uniform on the same sphere product, so the objective and initialization law match directly.

Theorem A.1 and the local budget (32) concern the uncentered $\mathcal{E}$, while both the network objective and the Stage-I objective of He et al. (2026b) produce the same centered scalar potential $f(a, b, c) = \Omega(a, b, Pc)$. This subsection works on $(S^{N-1})^3$. To transfer the sector basins, fix a nontrivial sector $j_0$ of ceiling $\gamma_{j_0} > 0$, let $K_{j_0}$ be its compact set of cross-layer-aligned intrinsic-$D_{j_0}$-rank-one states, and let $W$ be a neighborhood of $K_{j_0}$ on which (32) gives $\mathcal{E} \le \gamma_{j_0}$ with equality exactly on $K_{j_0}$. Then there is a nonempty open $A \subset (S^{N-1})^3$, one and the same set for every Riemannian metric of class $C^1$, from which every gradient-ascent trajectory of $f$ has its whole $\omega$-limit set contained in $K_{j_0}$; any full-support initialization law gives that event positive probability.

Nontriviality gives $K_{j_0} \subset (\mathbf{1}^\perp)^3$, so the normalization map below is defined and equal to the identity on $K_{j_0}$. Because $Pc$ sums to zero and left and right translation are bijections, $\Omega(\mathbf{1}, b, Pc) = \Omega(a, \mathbf{1}, Pc) = 0$. Expanding the first two slots along $\mathbb{R}\mathbf{1} \oplus \mathbf{1}^\perp$ therefore leaves $f = \Omega(Pa, Pb, Pc)$, and trilinearity gives the exact factorization

$$f(x) = \tau\, \mathcal{E}(\pi(x)), \qquad \pi(x) = \Big(\frac{Pa}{\|Pa\|}, \frac{Pb}{\|Pb\|}, \frac{Pc}{\|Pc\|}\Big), \qquad \tau = \|Pa\|\, \|Pb\|\, \|Pc\| \in (0, 1] \tag{37}$$

wherever $\tau > 0$, with $f = 0$ where any factor vanishes. Put $W' = \pi^{-1}(W)$, which is open because $\{\tau > 0\}$ is open and $\pi$ is continuous on it, and which contains $K_{j_0}$ because $\pi$ is the identity there; carving $W'$ as a preimage rather than as a tube keeps it clear of the locus where $\pi$ is undefined. For $x \in W'$, if $\mathcal{E}(\pi(x)) \geq 0$, then $f(x) \leq \mathcal{E}(\pi(x)) \leq \gamma_{j_0}$. If $\mathcal{E}(\pi(x)) < 0$, then $f(x) < 0 < \gamma_{j_0}$. The condition $\gamma_{j_0} > 0$ is needed here: multiplying a negative value by $\tau \leq 1$ raises it, so the naive bound $f \leq \mathcal{E} \circ \pi$ fails in that case. Equality at $\gamma_{j_0}$ forces the nonnegative case and $\mathcal{E}(\pi(x)) > 0$. The chain $\gamma_{j_0} = \tau\mathcal{E}(\pi(x)) \leq \mathcal{E}(\pi(x)) \leq \gamma_{j_0}$ then gives $\mathcal{E}(\pi(x)) = \gamma_{j_0}$, hence $\pi(x) \in K_{j_0}$, and also $\tau = 1$. Thus $a, b, c \perp \mathbf{1}$ and $x = \pi(x)$. So $f \leq \gamma_{j_0}$ on $W'$ with equality exactly on $K_{j_0}$, and every point of $K_{j_0}$ is an interior local maximum of $f$ on $(S^{N-1})^3$, hence a critical point.

The collar argument runs on $(f, W')$ as follows. Since $K_{j_0}$ is compact, nonempty and a proper subset of the connected $(S^{N-1})^3$, small tubes have nonempty boundary spheres, and $r$ may be taken with the closed tube inside $W'$; the boundary sphere is compact and misses $K_{j_0}$, so $b_r = \max_{\mathrm{dist}=r} f < \gamma_{j_0}$. The critical set $\{x : \mathrm{d}f|_{T_x(S^{N-1})^3} = 0\}$ is real algebraic, so by Tarski–Seidenberg its image under $f$ is semialgebraic and by Sard it is Lebesgue null, hence finite; let $\delta$ be smaller than the gap down to the largest critical value below $\gamma_{j_0}$, and arbitrary if there is none. On $A = \{\mathrm{dist} < r\} \cap \{f > \gamma_{j_0} - \min(\delta, \gamma_{j_0} - b_r)\}$, ascent keeps $f > b_r$, so by the intermediate value theorem applied to $t \mapsto \mathrm{dist}(x(t), K_{j_0})$ the trajectory never reaches the boundary sphere; $(S^{N-1})^3$ is compact, so the field is complete and the forward orbit is precompact. Its $\omega$-limit set is nonempty, invariant, carries a single critical value in $(\gamma_{j_0} - \delta, \gamma_{j_0}]$, and therefore sits at $\gamma_{j_0}$, whose level set inside the tube is $K_{j_0}$. Criticality is the metric-independent condition $\mathrm{d}f|_{T(S^{N-1})^3} = 0$, and $A$ is built from $f$ and the round distance alone, so one $A$ serves every metric of class $C^1$, that regularity being what makes $\mathrm{grad}\, f$ locally Lipschitz and the $\omega$-limit set invariant. Finally $A$ is open and nonempty, so a full-support law charges it. For a general $C^1$ metric the conclusion is containment of the $\omega$-limit set. For the product round metric, Theorem B.1 below gives convergence to one point.

For $Q_8$ the quaternionic sector has $\gamma_\rho > 0$, so a set of initializations of positive measure has its $\omega$-limit set inside the aligned intrinsic-$\mathbb{H}$-rank-one component, and this separately normalized Stage-I flow reaches that component with positive probability. Every point of the component has two equal, strictly positive complex singular values, so the smaller one cannot vanish in the limit and the surviving coefficient has ordinary complex rank two. For quaternionic type, this qualifies only the ordinary-rank-one clause of He et al. (2026b, v2, Theorem 4.3(ii)). Their single-representation statement and cross-layer proportionality relations still hold here, and the argument does not affect their finite-horizon comparison, Abelian population law, or groups with only real- or complex-type irreducibles. The corresponding real-form statement is intrinsic rank one over $D_\rho$, as in Theorem A.1.

## B Centered joint geometry and the leading selector

This appendix proves Propositions 4.1 and 3.2, Corollary B.2, and Theorem B.1. The uncentered real-form landscape of Theorem A.1 is used only after the joint input-sphere critical equations force a reduction to it.

### B.1 The projection identity and exact centering

Write $L_s$ for the summed full-batch cross entropy of the network (3), $Z$ for its logits without the factor $s^3/M$, and $\psi(w) = \log(N^{-1}\sum_k e^{w_k})$. The exact split referred to in the main text is

$$L_s = N^2 \log N - \lambda_s \mathcal{F} + R_s, \qquad \mathcal{F} = \sum_m f(a_m, b_m, c_m), \qquad \lambda_s = \frac{2s^3N^2}{M}, \tag{38}$$

$$R_s = \sum_{g,h} \psi\left(\frac{s^3}{M} P Z^{(g,h)}\right), \qquad |R_s| \leq N s^6, \qquad \|\mathrm{grad}\, R_s\| \leq 2Ns^6\sqrt{\frac{N+4}{M}}, \tag{39}$$

so that, after dividing by the positive leading coefficient, exact-CE descent is ascent on

$$\Phi_s = \mathcal{F} - R_s/\lambda_s, \qquad \Phi_s - \mathcal{F} = O(s^3). \tag{40}$$

Write $a = Pa + \alpha\mathbf{1}$ and $b = Pb + \beta\mathbf{1}$. For group convolution,

$$\mathbf{1} * x = \left(\sum_g x_g\right)\mathbf{1}, \qquad x * \mathbf{1} = \left(\sum_g x_g\right)\mathbf{1}.$$

Every term in $a * b$ containing a constant component is therefore constant and orthogonal to $Pc$. This proves

$$\Omega(a, b, Pc) = \Omega(Pa, Pb, Pc)$$

without using the later Fourier decomposition.

To obtain (38), split one sample's logit vector as $z = \bar{z}\mathbf{1} + Pz$. Then

$$\log \sum_k e^{z_k} - z_y = \log N + \psi(Pz) - (Pz)_y.$$

Summing the linear term over all samples and expanding $(a_g + b_h)^2$, the two diagonal terms vanish because $h \mapsto gh$ and $g \mapsto gh$ are bijections:

$$\sum_{g,h} (Pc)_{gh} a_g^2 = \left( \sum_k (Pc)_k \right) \|a\|^2 = 0,$$

and similarly for $b$. The cross term is $2N^2\Omega(a, b, Pc)$, proving the exact split.

The softmax Hessian is $\operatorname{Hess} \psi(w) = \operatorname{diag}(p) - pp^\top$, where $p = \operatorname{softmax}(w)$. For every unit vector $v$,

$$v^\top \operatorname{Hess} \psi(w) v = \operatorname{Var}_p(v) \le \frac{(\max v - \min v)^2}{4} \le \frac{1}{2}.$$

The value and gradient bounds in (39) then follow from the estimates

$$\sum_{g,h} \left\| PZ^{(g,h)} \right\|^2 \le 4NM^2, \qquad \sum_{g,h} \left\| DZ^{(g,h)} \right\|_F^2 \le 4NM(N+4),$$

and Cauchy–Schwarz over the $N^2$ samples.

### B.2 Nonzero critical points force balance

Let $(a, b, c)$ be critical for $f$ on $S^{2N-1} \times S^{N-1}$, with $E = f(a, b, c) \neq 0$. There is one input multiplier $\lambda$ and one output multiplier $\mu$:

$$\operatorname{grad}_a f = \lambda a, \qquad \operatorname{grad}_b f = \lambda b, \qquad \operatorname{grad}_c f = \mu c.$$

Euler's identities give

$$\lambda \|a\|^2 = \lambda \|b\|^2 = \mu = E.$$

Since $\|a\|^2 + \|b\|^2 = 1$, this forces $\lambda = 2E$, $\mu = E$, and $\|a\|^2 = \|b\|^2 = 1/2$. Every Euclidean partial gradient lies in $\mathbf{1}^\perp$; the nonzero multipliers then give $a, b, c \perp \mathbf{1}$. Putting $u = \sqrt{2}a$ and $v = \sqrt{2}b$ turns the three equations into the separate-sphere critical equations for $\Omega$ on the nontrivial ideal $\mathbf{1}^\perp$, at value $2E$.

The tangent space splits orthogonally into the input-balance direction, one trivial direction in each of the three slots, and the reduced tangent space. Direct substitution into the constrained Hessian gives curvatures

$$-4E, \qquad -2E, \qquad -2E, \qquad -E$$

on the first four summands. Every cross term between these summands and the reduced tangent vanishes. On the reduced tangent, the map

$$(\dot{u}, \dot{v}, \dot{c}) \longmapsto (\dot{u}/\sqrt{2}, \dot{v}/\sqrt{2}, \dot{c})$$

identifies the reduced joint Hessian, by congruence, with one half of the separate-sphere Hessian. Congruence preserves inertia, so every nonmaximum positive critical point is a strict saddle by Theorem A.1.

Local maximality requires a separate argument. For nontrivial sectors $j$, put

$$\alpha_j = \|\Pi_j a\|, \quad \beta_j = \|\Pi_j b\|, \quad \chi_j = \|\Pi_j c\|, \quad x = \|a\|^2.$$

The real-form sector bound gives

$$f \le \sum_{j \neq \mathrm{triv}} \gamma_j \alpha_j \beta_j \chi_j, \qquad \sum_j \alpha_j \beta_j \chi_j \le \sqrt{x(1-x)} \le \frac{1}{2}.$$

Thus every leading sector ceiling is halved and the trivial sector is absent. Near a maximum in sector $j_0$, define

$$t = x - \frac{1}{2}, \quad e^2 = x - \alpha_{j_0}^2, \quad r^2 = 1 - x - \beta_{j_0}^2, \quad h^2 = 1 - \chi_{j_0}^2, \quad \sigma^2 = t^2 + e^2 + r^2 + h^2.$$

For $\sigma < 1/4$,

$$\alpha_{j_0}\beta_{j_0}\chi_{j_0} \le \frac{1}{2}\left(1 - \frac{\sigma^2}{2} + \frac{3}{2}\sigma^4\right),$$

while all rival sectors together contribute at most $\sigma^3/3^{3/2}$. With $\Gamma' = \max_{j \neq \mathrm{triv}} \gamma_j$,

$$f - \frac{\gamma_{j_0}}{2} \le -\sigma^2 \left[\frac{\gamma_{j_0}}{4} - \sigma\left(\frac{\Gamma'}{3^{3/2}} + \frac{3\gamma_{j_0}}{4}\right)\right] < 0$$

for sufficiently small positive $\sigma$. On $\sigma = 0$, equality in the single-sector bound holds exactly on the aligned intrinsic-rank-one orbit. This proves the positive half of Proposition 4.1.

### B.3 The complete nonpositive branch

At a zero-energy critical point, both sphere multipliers vanish. The ambient Euclidean Hessian $H$ has zero diagonal layer blocks and therefore trace zero. It also annihilates the two sphere normals by trilinearity and criticality. Suppose its tangent restriction were zero. Symmetry would put $H(T\mathcal{M}_1)$ in the normal span, while annihilation of the normals would make $H(T\mathcal{M}_1)$ orthogonal to that span. Hence $H = 0$ on the full ambient space.

Let $x = Pa$, $y = Pb$, and $z = Pc$. The nontrivial ideal $\mathbf{1}^\perp$ is a unital Frobenius algebra. The vanished $a$–$b$ Hessian block, evaluated with the unit in one slot, gives $\langle \eta, z\rangle = 0$ for every $\eta$, hence $z = 0$. The other two mixed blocks similarly give $x = y = 0$. Conversely, $x = y = z = 0$ makes every Hessian block vanish. Thus every zero critical point off the fully trivial stratum has a nonzero symmetric trace-free tangent Hessian and therefore a positive eigenvalue. On the stratum the Hessian is identically zero.

Each mixed Hessian block is linear in one projected base layer. The multilinear norm bound therefore gives

$$\lambda_{\max}(\mathrm{Hess}\, f) \le 2N^{-3/2}\big(\|Pa\| + \|Pb\| + \|Pc\|\big)$$

at zero energy. The escape curvature can vanish continuously near the flat stratum. At negative energy, direct trace computation gives

$$\mathrm{tr}\,\mathrm{Hess}\, f = -(5N-1)f > 0,$$

so every negative critical point is a strict saddle. This completes Proposition 4.1.

### B.4 Analytic convergence and saddle avoidance

Set $q = (a, b)$ and use the scale-free extension

$$f_0(q, c) = \frac{f(q, c)}{\|q\|^2 \, \|c\|}.$$

The polynomial ambient extension of the joint field obeys

$$\frac{\mathrm{d}}{\mathrm{d}t}(\|q\|^2 - 1) = -4f(\|q\|^2 - 1), \qquad \frac{\mathrm{d}}{\mathrm{d}t}(\|c\|^2 - 1) = -2f(\|c\|^2 - 1).$$

The manifold is invariant. Along the flow,

$$\frac{\mathrm{d}}{\mathrm{d}t} f(X(t)) = \left\|\mathrm{grad}_{\mathcal{M}_1} f(X(t))\right\|^2.$$

Compactness gives accumulation points. The Łojasiewicz gradient inequality for $f_0$ near any one of them (Colding & Minicozzi, 2014) gives

$$|f(X) - f(p)|^{1-\theta} \le C \left\|\mathrm{grad}_{\mathcal{M}_1} f(X)\right\|$$

for some $\theta \in (0, 1/2]$. The usual energy-length calculation bounds the remaining arc length after a sufficiently late return, so the trajectory converges to one critical point.

For $L = -f$, use the strict-saddle-subset argument of Bah et al. (2022, arXiv v5, Theorems 28–29), which requires no global strict-saddle property. The manifold is smooth, finite dimensional and second countable, the metric is smooth, and $L$ is $C^2$; compactness makes its flow complete. For any fixed $t > 0$, its time map $H = \mathrm{Fl}^t_{-\mathrm{grad}\, L}$ is a $C^1$ diffeomorphism, hence $\det DH \neq 0$. At a strict saddle $x$, equilibrium linearization gives $DH(x) = \exp(-t \,\mathrm{Hess}\, L(x))$, which has an eigenvalue greater than one. Thus the whole strict-saddle set lies in the unstable fixed-point set of $H$, to which their Theorem 29 applies. The basin of this set is therefore Riemannian null, without any countability assumption on the saddles. The flat stratum prevents a global strict-saddle hypothesis and is handled separately below.

It remains to exclude the fully trivial stratum. With $e = N^{-1/2}\mathbf{1}$ and trivial coordinates $\alpha, \beta, \gamma$, the partial gradients are centered and hence

$$\dot\alpha = -2f\alpha, \qquad \dot\beta = -2f\beta, \qquad \dot\gamma = -f\gamma.$$

If $\gamma = 0$, the hyperplane $\gamma = 0$ is invariant and the orbit cannot reach a trivial output. Otherwise

$$J = \frac{\alpha^2 + \beta^2}{\gamma^4} \tag{41}$$

is conserved. This conserved ratio equals one on the flat stratum, so the basin is contained in the proper analytic zero set $\alpha^2 + \beta^2 - \gamma^4 = 0$. This set has volume zero and proves Proposition 4.1.

### B.5 Almost-sure resolution of the separate centered flow

**Theorem B.1** (Almost-sure convergence of the separate flow)**.** *Let $G$ be a finite group of order $N \ge 2$, and give $\widetilde{\mathcal{M}} = (S^{N-1})^3$ the product round metric. Every gradient-ascent trajectory of $f(a, b, c) = \Omega(Pa, Pb, Pc)$ converges to one critical point. Under any initialization law absolutely continuous with respect to volume on $\widetilde{\mathcal{M}}$, almost every trajectory converges to a positive, single-nontrivial-sector, cross-layer-aligned local maximum of intrinsic $D_\rho$-rank one, at value $\gamma_\rho$. In particular, if $R$ is the event of convergence to such a maximum, the separate-sphere Gaussian initialization law satisfies $\mathbb{P}[R] = 1$.*

*Proof.* Write $G_a, G_b, G_c$ for the Euclidean partial gradients of $f$. They are centered, and trilinearity gives $\langle a, G_a\rangle = \langle b, G_b\rangle = \langle c, G_c\rangle = f$. The separate field is therefore

$$\dot a = G_a - fa, \qquad \dot b = G_b - fb, \qquad \dot c = G_c - fc. \tag{42}$$

It is analytic and tangent to the compact sphere product, so its flow is complete. The extension $f/(\|a\|\,\|b\|\,\|c\|)$ is analytic near that product and has Euclidean gradient equal there to the Riemannian gradient. Along a trajectory, $\dot f = \|\mathrm{grad}\, f\|^2$, so $f$ increases to a limit $E_\infty$ and the squared gradient is integrable. Its derivative is bounded by compactness, hence the gradient tends to zero and every accumulation point $p$ is critical. The Łojasiewicz inequality near $p$ gives $(E_\infty - f)^{1-\theta} \le C\,\|\mathrm{grad}\, f\|$ for some $\theta > 0$. Until the trajectory leaves that neighborhood, its remaining arc length is at most $C(E_\infty - f)^\theta/\theta$. A sufficiently late return close to $p$ makes this smaller than the distance to the neighborhood's boundary, preventing exit and giving finite remaining length. If the energy gap is zero, the trajectory is already stationary. Thus every trajectory converges to one critical point, without an isolation assumption.

*Nonzero critical points.* At a critical value $E \neq 0$, the equations $G_a = Ea$, $G_b = Eb$, $G_c = Ec$ force $Pa = a$, $Pb = b$, $Pc = c$. In the nontrivial Fourier blocks, (23) becomes

$$B_\rho^* C_\rho = NEA_\rho, \qquad C_\rho A_\rho^* = NEB_\rho, \qquad B_\rho A_\rho = NEC_\rho.$$

For $E > 0$, the critical-rigidity calculation above gives $B^*B = AA^*$ and $AA^*A = (NE)^2 A$. Thus, with $\phi = N|E|$, every active block has the form

$$A_\rho = \phi U_\rho V_\rho^*, \qquad B_\rho = \phi W_\rho U_\rho^*, \qquad C_\rho = \operatorname{sgn}(E)\phi W_\rho V_\rho^*,$$

where $U_\rho, V_\rho, W_\rho$ have $r_\rho$ orthonormal $D_\rho$-columns; inactive blocks vanish in all three slots. The $E < 0$ case follows by replacing $c$ by $-c$, and the converse follows by substitution. Plancherel and the complex singular-value multiplicities give the complete list of nonzero critical values:

$$E = \pm \left( N^3 \sum_{\rho \neq \mathrm{triv}} d_\rho^\sharp \kappa_\rho r_\rho \right)^{-1/2}, \qquad \kappa_\rho = 1, 1, 2 \quad (D_\rho = \mathbb{R}, \mathbb{C}, \mathbb{H}). \tag{43}$$

Here each active $r_\rho$ is intrinsic rank, not ordinary complex rank.

If $E > 0$ and at least two intrinsic singular directions are active, put the blocks in the canonical coordinates (27). The common great-circle variation $z \cos t + h \sin t$, with $h \perp z$ in their active span and $\|h\| = 1$, has curvature $3E > 0$. This is feasible over each $D_\rho$ and covers both higher rank and several active sectors. The remaining positive points have one active sector of intrinsic rank one. The centered local budget proved above makes them local maxima, including against perturbations in the trivial directions. At $E < 0$, moving any one layer along a nonzero sphere tangent gives curvature $-E\|v\|^2 > 0$. Such a direction exists even when $N = 2$.

*Zero critical points and the exceptional set.* At $E = 0$, all three Euclidean partial gradients vanish. The ambient Hessian $H$ has zero diagonal layer blocks by trilinearity, hence $\operatorname{tr} H = 0$. By criticality, it also annihilates each sphere normal $(a, 0, 0)$, $(0, b, 0)$, $(0, 0, c)$. Consequently, the tangent restriction is the constrained Hessian and also has trace zero. If that restriction vanishes, symmetry and annihilation of the normal span imply $H = 0$ on the whole ambient space.

Put $x = Pa$, $y = Pb$, $z = Pc$, and let $u = \delta_{1_G} - N^{-1}\mathbf{1}$ be the two-sided convolution unit of $\mathbf{1}^\perp$. If $H = 0$, its $a$–$b$ block evaluated at $(u, z)$ gives $\Omega(u, z, z) = N^{-2}\|z\|^2 = 0$. Its $a$–$c$ and $b$–$c$ blocks give, respectively, $\Omega(u, y, y) = N^{-2}\|y\|^2 = 0$ and $\Omega(x, u, x) = N^{-2}\|x\|^2 = 0$. Thus $x = y = z = 0$; conversely this condition makes the gradient and Hessian vanish. On three unit spheres it describes exactly the eight points

$$\mathcal{T} = \{(\epsilon_a e, \epsilon_b e, \epsilon_c e) : \epsilon_a, \epsilon_b, \epsilon_c \in \{\pm 1\}\}, \qquad e = N^{-1/2}\mathbf{1}.$$

Every other zero critical point has a nonzero symmetric trace-free tangent Hessian, hence a positive eigenvalue. Points of $\mathcal{T}$ are neither maxima nor minima: for any unit aligned nontrivial triple $(a_*, b_*, c_*)$, the feasible curve $(\epsilon_a\sqrt{1-t^2}e + ta_*, \epsilon_b\sqrt{1-t^2}e + tb_*, \epsilon_c\sqrt{1-t^2}e + tc_*)$ has value $t^3\gamma_\rho$ of either sign near $t = 0$. Thus every nonmaximum critical point outside $\mathcal{T}$ is a strict saddle for ascent, meaning it has a positive Hessian eigenvalue.

*The flat basin is null.* Let $\alpha = \langle a, e\rangle$, $\beta = \langle b, e\rangle$, and $\gamma = \langle c, e\rangle$. Centered partial gradients in (42) give

$$(\alpha, \beta, \gamma)(t) = \exp\left( -\int_0^t f(X(v))\, \mathrm{d}v \right) (\alpha, \beta, \gamma)(0). \tag{44}$$

A zero coordinate stays zero and cannot approach a point of $\mathcal{T}$. Otherwise the coordinate ratios are constant, whereas every point of $\mathcal{T}$ has $|\alpha| = |\beta| = |\gamma| = 1$. Convergence to $\mathcal{T}$ therefore requires $\alpha(0)^2 = \beta(0)^2 = \gamma(0)^2$. In particular, the basin of $\mathcal{T}$ is contained in the zero set of $\alpha^2 - \beta^2$, a nonzero real-analytic function on the connected manifold $(S^{N-1})^3$ for $N \geq 2$: choosing $a = e$ and $b \perp e$ makes that function equal to one. This is a volume-null necessary condition, not a characterization of that basin.

Finally, apply the strict-saddle-subset argument of Bah et al. (2022, arXiv v5, Theorems 28–29) to $L = -f$. The manifold is finite dimensional, smooth, and second countable; the metric is smooth, $L \in C^2$, and compactness makes the flow complete. Its time-$t$ map is a $C^1$ diffeomorphism. At each strict saddle, equilibrium linearization gives an eigenvalue greater than one for its derivative $\exp(-t \operatorname{Hess} L)$. Their Theorem 29 therefore makes the basin of the strict-saddle set Riemannian null without requiring a countable critical set or a global strict-saddle property, the latter being violated by the flat stratum. Together with analytic convergence and the null basin of $\mathcal{T}$, this leaves exactly the stated positive maxima almost surely. Independent normalized Gaussian vectors give the uniform product-sphere law, which is absolutely continuous, so $\mathbb{P}[R] = 1$ under that law. □

**Positive scalar layer weights.** The same almost-sure convergence conclusion holds for every fixed metric $g_w = w_a g_a + w_b g_b + w_c g_c$ on the separate spheres, with $w_a, w_b, w_c > 0$. Analytic convergence and strict-saddle inertia are unchanged. The trivial coordinates now satisfy $\alpha' = -f\alpha/w_a$, $\beta' = -f\beta/w_b$ and $\gamma' = -f\gamma/w_c$. When nonzero, they conserve $w_a \log|\alpha| - w_b \log|\beta|$ and $w_a \log|\alpha| - w_c \log|\gamma|$. A fully trivial endpoint therefore requires $|\alpha(0)|^{w_a} = |\beta(0)|^{w_b} = |\gamma(0)|^{w_c}$; an initially zero coordinate cannot reach such an endpoint. For fixed $a$, the first equality restricts $\beta$ to at most two coordinate levels, each of sphere volume zero for $N \geq 2$. Fubini makes this exceptional set null under any absolutely continuous joint law. The same argument applies neuronwise at fixed finite width. These are scalar weights on entire layers, not arbitrary coordinatewise matrices. No constant is uniform as a weight tends to zero or infinity.

### B.6 The canonical score

**Corollary B.2** (Finite settling). *On the full-measure event of Proposition 4.1,* $\arg\max_\rho S_\rho$ *in* (7) *equals the limiting sector after a finite, initialization-dependent time, simultaneously for all neurons at finite width.*

Write $x = \|\Pi_\rho a\|^2$, $y = \|\Pi_\rho b\|^2$, and $z = \|\Pi_\rho c\|^2$. On the convex hull of $\mathcal{M}_1$, $x + y \leq 1$ and $z \leq 1$. Differentiating $S_\rho = xyz$ gives

$$\|\operatorname{grad} S_\rho\|^2 = 4xyz(yz + xz + xy) \leq \frac{5}{4}.$$

The maximum is approached by inward directions at the boundary, so the Lipschitz constant is $L_S = \sqrt{5}/2$. At a one-sector balanced maximum, $S_{\rho_*} = 1/4$ and every rival score is zero. If $\|X - Y\| < 1/(4\sqrt{5})$, then every score changes by less than $1/8$, which preserves the unique label.

Let $\varphi_T : \mathcal{M}_1 \to \mathcal{M}_1$ denote the time-$T$ map of the leading joint field, (2). For distinct sectors $\rho, \sigma$ and fixed $T$, the analytic function

$$X_0 \longmapsto S_\rho(\varphi_T(X_0)) - S_\sigma(\varphi_T(X_0))$$

is not identically zero: a pure $\rho$ equilibrium is a witness. On the connected manifold, its zero set has measure zero. Thus the label is unique almost surely at every fixed horizon. Convergence to a one-sector maximum eventually places the whole trajectory tail inside this label radius, proving Corollary B.2.

### B.7 The exact two-sector metric comparison

This subsection gives the reduced two-sector system, its saddle and the local reversal that Appendix C uses. The global metric threshold on the same square is an extension (Corollary F.7, Appendix F).

A finite group has at least two nontrivial real Wedderburn sectors if and only if $|G| \geq 4$. One direction is a list: the groups of order at most three are $C_1$, $C_2$ and $C_3$, whose nontrivial real sector counts are 0, 1 and 1. For the other, suppose $G$ has exactly one nontrivial conjugation orbit, carried by an irreducible $\chi$ of complex degree $d$. Evaluating the regular character at $g \neq e$ gives $1 + \sum_{\rho \neq 1} d_\rho \chi_\rho(g) = 0$. If that orbit is a singleton, this reads $d\chi(g) = -1$, so $\chi(g) = -1/d$ is a rational algebraic integer and hence $d = 1$. Thus $\chi$ is a homomorphism into $\{\pm 1\}$ with trivial kernel, so $|G| \leq 2$.

If the orbit is a conjugate pair, it reads $d(\chi(g) + \overline{\chi(g)}) = -1$. Since $\chi + \overline{\chi}$ is an algebraic integer, again $d = 1$, and $\chi(g) + \chi(g)^{-1} = -1$ for every $g \neq e$. Hence $\chi$ is injective with image in the cube roots of unity, so $|G| \leq 3$. A trivial group has no nontrivial sector at all. So one nontrivial sector forces $|G| \in \{2, 3\}$, and $|G| \geq 4$ forces at least two.

Throughout, $N = |G| \geq 4$ and $\rho_1 \neq \rho_2$ are two nontrivial real Wedderburn sectors with ceilings $p = \gamma_{\rho_1}$ and $q = \gamma_{\rho_2}$, as in Proposition 3.2. Let $P_j \in \mathcal{A}_{\rho_j}$ be a primitive self-adjoint idempotent, so $P_j^* = P_j = P_j^2$ and $\operatorname{tr} P_j = \kappa_{\rho_j}$. Let $r_j \in \mathbb{R}^G$ be the corresponding canonical unit ray, the real signal with $\widehat{(r_j)}_{\rho_j} = P_j/\sqrt{m_{\rho_j}\kappa_{\rho_j}}$ and every other block zero. Since blocks of distinct sectors never meet in (11),

$$\Omega(r_i, r_j, r_k) = \gamma_{\rho_i} \mathbf{1}_{\{i=j=k\}}, \qquad i, j, k \in \{1, 2\}, \tag{45}$$

so $\mathrm{span}\{r_1, r_2\}$ is invariant in all three slots and carries the reduced potential $p\,a_1b_1c_1 + q\,a_2b_2c_2$. When both sectors are real linear characters $\chi, \psi$, $r_1 = \chi/\sqrt{N}$, $r_2 = \psi/\sqrt{N}$ and $p = q = N^{-3/2}$. With $u_\alpha = \alpha/\sqrt{N}$, this recovers the character identity

$$\Omega(u_\alpha, u_\beta, u_\gamma) = \frac{1}{N^{7/2}} \sum_{g,h\in G} \alpha(g)\beta(h)\gamma(gh) = N^{-3/2}\mathbf{1}_{\{\alpha=\gamma\}}\mathbf{1}_{\{\beta=\gamma\}}, \qquad \alpha, \beta, \gamma \in \{\chi, \psi\}.$$

Let

$$\widetilde{\mathcal{M}} = S^{N-1} \times S^{N-1} \times S^{N-1}, \qquad \mathcal{B} = \{(a, b, c) \in \mathcal{M}_1 : \|a\| = \|b\| = 1/\sqrt{2}\},$$

and let $\Psi(\widetilde{a}, \widetilde{b}, c) = (\widetilde{a}/\sqrt{2}, \widetilde{b}/\sqrt{2}, c)$. Along the joint flow,

$$\frac{\mathrm{d}}{\mathrm{d}t}(\|a\|^2 - \|b\|^2) = -4f(\|a\|^2 - \|b\|^2),$$

so $\mathcal{B}$ is invariant and $\Psi : \widetilde{\mathcal{M}} \to \mathcal{B}$ identifies the separate state space with the balanced joint submanifold. For the full real spatial comparator $f_{\mathrm{sep}} = \Omega(Pa, Pb, Pc)$, linearity of $P$ gives

$$f_{\mathrm{joint}} \circ \Psi = \tfrac{1}{2} f_{\mathrm{sep}}, \qquad V := f_{\mathrm{sep}} = 2 f_{\mathrm{joint}} \circ \Psi,$$

whereas

$$g_{\mathrm{sep}} = g_a + g_b + g_c, \qquad \Psi^* g_{\mathrm{joint}} = \tfrac{1}{2} g_a + \tfrac{1}{2} g_b + g_c.$$

Thus the two systems follow the metric gradients of the same $V$, up to a global time rescaling.

With $R = \sqrt{p^2 + q^2}$ and $r_* = (q\,r_1 + p\,r_2)/R$, the separate-sphere mixed saddle is

$$\widetilde{a}_* = \widetilde{b}_* = r_*, \qquad c_* = r_*,$$

which corresponds under $\Psi$ to the joint-sphere point $a_* = b_* = r_*/\sqrt{2}$, $c_* = r_*$. It sits at angle $\theta_* = \phi_* = \arctan(p/q)$ in the chart below, and $r_*$ reduces to $(r_1 + r_2)/\sqrt{2}$ exactly when $p = q$.

**The reduced system for a general pair.** On the aligned two-sector surfaces,

$$g_{\mathrm{sep}} = 2\ \mathrm{d}\theta^2 + \mathrm{d}\phi^2, \qquad \Psi^* g_{\mathrm{joint}} = \mathrm{d}\theta^2 + \mathrm{d}\phi^2,$$

and $V(\theta, \phi) = p\cos^2\theta\cos\phi + q\sin^2\theta\sin\phi$. Taking the two metric gradients of $V$ and rescaling time gives

$$\theta' = \sin\theta\cos\theta(q\sin\phi - p\cos\phi), \qquad \phi' = \mu(-p\cos^2\theta\sin\phi + q\sin^2\theta\cos\phi), \tag{46}$$

with $\mu = 1/2$ for joint normalization and $\mu = 1$ for separate normalization. In the open square, $\theta' = 0$ forces $\tan\phi = p/q$. Substituting this into $\phi' = 0$ gives $\tan^2\theta = (p/q)\tan\phi = (p/q)^2$. Hence the unique interior equilibrium is $\theta_* = \phi_* = \arctan(p/q)$, with $\cos\theta_* = q/R$ and $\sin\theta_* = p/R$. There $\sin\theta_*\cos\theta_* = pq/R^2$ and $q\cos\phi_* + p\sin\phi_* = R$, so the Jacobian is

$$E\begin{pmatrix} 0 & 1 \\ 2\mu & -\mu \end{pmatrix}, \qquad E = \frac{pq}{R} > 0, \tag{47}$$

of determinant $-2\mu E^2 < 0$ and characteristic polynomial $x^2 + \mu E x - 2\mu E^2$. Writing a stable eigenvector as $(1, m_\mu)$, the common factor $E$ cancels and $m_\mu^2 + \mu m_\mu - 2\mu = 0$, hence the two slopes in Proposition 3.2. The ceilings move the saddle and the two eigenvalues $E(-\mu \pm \sqrt{\mu^2 + 8\mu})/2$, but not the slopes.

**The two rectangles are forward invariant for unequal ceilings.** Put $\mathcal{R}_- = [0, \theta_*] \times [0, \phi_*]$ and $\mathcal{R}_+ = [\theta_*, \pi/2] \times [\phi_*, \pi/2]$. On $\{\phi = \phi_*\}$, substituting $\sin\phi_* = p/R$ and $\cos\phi_* = q/R$ into (46) gives

$$\phi'\big|_{\phi=\phi_*} = \frac{\mu}{R}\big(q^2\sin^2\theta - p^2\cos^2\theta\big) = \frac{\mu\cos^2\theta}{R}\big(q^2\tan^2\theta - p^2\big), \tag{48}$$

negative for $\theta < \theta_*$ and positive for $\theta > \theta_*$. On $\{\theta = \theta_*\}$,

$$\theta'\big|_{\theta=\theta_*} = \frac{pq}{R^2}\big(q\sin\phi - p\cos\phi\big) = \frac{pq\cos\phi}{R^2}\big(q\tan\phi - p\big), \tag{49}$$

negative for $\phi < \phi_*$ and positive for $\phi > \phi_*$. Both faces therefore point into $\mathcal{R}_-$ on one side of the saddle coordinate and into $\mathcal{R}_+$ on the other. The faces $\theta \in \{0, \pi/2\}$ are invariant because $\theta'$ carries the factor $\sin\theta\cos\theta$, and on $\{\phi = 0\}$ one has $\phi' = \mu q \sin^2\theta \geq 0$ while on $\{\phi = \pi/2\}$ one has $\phi' = -\mu p \cos^2\theta \leq 0$, both inward. So the closed square and both rectangles are forward invariant for every $\mu > 0$ and $p, q > 0$. When $p = q$, $\theta_* = \pi/4$, and (48) reduces to the corresponding equal-ceiling expression $q^2 \tan^2\theta - p^2 = -p^2 \cos 2\theta / \cos^2\theta$.

Apart from the saddle, the closed square has exactly two equilibria, the pure corners, with linearizations $\mathrm{diag}(-p, -\mu p)$ at $(0, 0)$ and $\mathrm{diag}(-q, -\mu q)$ at $(\pi/2, \pi/2)$; both are hyperbolic sinks. On $\{\theta = 0\}$ one has $\phi' = -\mu p \sin\phi$ and on $\{\theta = \pi/2\}$ one has $\phi' = \mu q \cos\phi$, so there is no further boundary equilibrium.

**Where the orbits go.** Along (46), $\mathrm{d}V/\,\mathrm{d}t = V_\theta^2/2 + \mu V_\phi^2$, which vanishes only at equilibria. The closed square is compact, forward invariant and contains only the three equilibria above, so every orbit in it converges to one of them. The local stable curve of the saddle has slope $m_\mu < 0$, so near the saddle it meets $\mathcal{R}_+ \cup \mathcal{R}_-$ only at the saddle; an orbit of $\mathcal{R}_\pm$ converging to the saddle would, by forward invariance, eventually lie on that curve inside $\mathcal{R}_\pm$, so no orbit of $\mathcal{R}_\pm$ other than the saddle does. Hence every point of $\mathcal{R}_+$ except the saddle converges to $(\pi/2, \pi/2)$, the pure sector $\rho_2$, and every point of $\mathcal{R}_-$ except the saddle to $(0, 0)$, the pure sector $\rho_1$. The unstable eigenvector $(1, k_\mu)$, $k_\mu = (-\mu + \sqrt{\mu^2 + 8\mu})/2 > 0$, points into the interiors of $\mathcal{R}_\pm$, so the two unstable branches of the saddle converge to these two corners.

**The metric wedge.** Near the saddle the stable manifold is a smooth curve $W^{\mathrm{s}}_{\mathrm{loc}}$ with tangent $(1, m_\mu)$. By the two-dimensional case of the exit estimate in Appendix C, a point off $W^{\mathrm{s}}_{\mathrm{loc}}$ leaves a fixed small box through the unstable face on its own side, and as the point tends to the saddle its exit point tends to the corresponding branch point, which lies in the interior of $\mathcal{R}_+$, the side containing $(1, k_\mu)$, or of $\mathcal{R}_-$. So there is $r_\mu > 0$ such that within distance $r_\mu$ of the saddle, points above $W^{\mathrm{s}}_{\mathrm{loc}}$ converge to $(\pi/2, \pi/2)$ and points below it to $(0, 0)$. Writing $\delta = \theta - \theta_*$, $\psi = \phi - \phi_*$, and its local graph as $\psi = \varphi_\mu(\theta)$, a point at $\psi = m\delta$ satisfies

$$\psi - \varphi_\mu(\theta) = (m - m_\mu)\delta + o(|\delta|).$$

Since $m_{\mathrm{separate}} < m_{\mathrm{joint}} < 0$, on the side $\delta < 0$ the separate boundary lies strictly above the joint boundary. Every $m_{\mathrm{separate}} < m < m_{\mathrm{joint}}$ gives opposite signs for the two metrics. For $\delta < 0$ small enough, points with $\psi = m\delta$ and $m_{\mathrm{separate}} < m < m_{\mathrm{joint}}$ therefore lie above one local curve and below the other, so the two metrics send them to opposite corners. This local wedge is open relative to the aligned two-sector surface inside $\mathcal{B}$. The global continuation of $W^{\mathrm{s}}$ (Corollary F.7) is not used.

**The metric family.** For $\lambda > 0$ give $\widetilde{\mathcal{M}}$ the metric $g_\lambda = \lambda g_a + \lambda g_b + g_c$ of Corollary B.3. On the aligned surface it restricts to $2\lambda\,\mathrm{d}\theta^2 + \mathrm{d}\phi^2$, so after a global time rescaling the reduced flow is the system displayed above with $\mu = \lambda$, and its stable slope at the mixed saddle is the negative root of $m^2 + \lambda m - 2\lambda = 0$,

$$m(\lambda) = -\frac{\lambda + \sqrt{\lambda^2 + 8\lambda}}{2}, \qquad m'(\lambda) = -\frac{1}{2} - \frac{\lambda + 4}{2\sqrt{\lambda^2 + 8\lambda}} < 0. \tag{50}$$

**Corollary B.3** (Metric-only disagreement in full dimension)**.** *Let $G$ be as in Proposition 3.2 and give $\widetilde{\mathcal{M}}$ the metric $g_\lambda = \lambda g_a + \lambda g_b + g_c$, $\lambda > 0$. Its stable slope at the shared mixed saddle is $m(\lambda)$ in (50), which is strictly decreasing, so for every $\lambda_1 \neq \lambda_2$ there is a nonempty* open *$U_{\lambda_1,\lambda_2} \subset \widetilde{\mathcal{M}}$, of positive mass for the uniform law, from which the $g_{\lambda_1}$- and $g_{\lambda_2}$-gradient ascents of the same potential $f^{\mathrm{real}}_{\mathrm{sep}}$ reach aligned rank-one components of two different sectors. Separate normalization is $\lambda = 1$ and the ordinary joint metric pulls back to $\lambda = 1/2$, with $m(1) = -2$ and $m(\frac{1}{2}) = -(\sqrt{17} + 1)/4$, so both arms of the comparison are members of one family; write $U_{\mathrm{met}} = U_{1,1/2}$.*

**Corollary B.4** (Paired disagreement with positive probability)**.** *Let $G$ be as in Proposition 3.2 and define $T(a, b, c) = (a/\|a\|, b/\|b\|, c)$ on $\{a \neq 0,\, b \neq 0\} \subset \mathcal{M}_1$, a set of full measure. There is a nonempty* open *$U \subset \mathcal{M}_1$ such that for $x \in U$ the joint flow from $x$ and the separate flow from $T(x)$ reach aligned rank-one components of two different sectors. If $A, B, C$ are independent standard Gaussians on $\mathbb{R}^G$, then $X_J = \big((A, B)/\|(A, B)\|, C/\|C\|\big)$ is uniform on $\mathcal{M}_1$, $X_S = T(X_J)$ is uniform on $\widetilde{\mathcal{M}}$, and $\mathbb{P}[X_J \in U] > 0$.*

Each maximum component has a metric-independent trapping neighborhood, and finite-time entry makes both labelled basins open, which lifts the aligned-surface wedge to an open subset of the ambient manifold, including the input-imbalance direction. The map $T$ inverts $\Psi$ on $\mathcal{B}$ and elsewhere also renormalizes the two input halves, so Corollary B.4 changes the state space as well as the metric, whereas Corollary B.3 changes only the metric.

# C SECTOR-LOCAL SCORING: CHARACTERIZATION, FLOORS AND THE COMPENSATING WEIGHT

Let $\mathcal{I}$ index the nontrivial real Wedderburn sectors of $G$, and use $0$ for the trivial sector. The negative direction assumes $|\mathcal{I}| \geq 3$; the positive direction applies to every nontrivial finite group.

## C.1 THE SECTOR-LOCAL INFORMATION CLASS

This is the class of Definition 2.1, written in the notation the proofs use. For $X = (a, b, c)$, let

$$\mathfrak{m}(X) = \left(\|\Pi_k a\|^2, \|\Pi_k b\|^2, \|\Pi_k c\|^2\right)_{k \in \{0\} \cup \mathcal{I}}.$$

The allowed data $D_j(X)$ consist of $(\Pi_j a, \Pi_j b, \Pi_j c)$, $j$, $\gamma_j$, $G$, the three layer norms and $\mathfrak{m}(X)$. A sector-local rule has the form $\sigma(X) = \arg\max_{j \in \mathcal{I}} s_j(D_j(X))$, where the functions $s_j$ are otherwise unrestricted and one fixed total priority order resolves ties. In particular, a score may read every sector mass, but it may not read another sector's orientation. The rule may depend on the fixed metric parameter $\lambda$. No continuity or symmetry of the scores is assumed.

Appendix B.5 defines $S_\lambda$ almost everywhere for each fixed $\lambda > 0$; extend it arbitrarily on the exceptional null set. Unless another law is specified, probabilities are taken under normalized Riemannian volume $\mu_\lambda$ on $(S^{N-1})^3$, and rules $\sigma$ are assumed measurable. For each fixed $\lambda$, the negative result also rules out agreement outside a null set without assuming measurability of the scores. No common exceptional set over all $\lambda > 0$ is needed.

## C.2 AN INVARIANT THREE-SECTOR MODEL

Here the scalar reflection argument is embedded in any three real Fourier sectors, of any types and any ceilings, by exhibiting a four-dimensional invariant slice on which the flow is explicit. Choose distinct sectors $1, 2, 3$ and their canonical unit rays $r_j$ from Appendix B.7, with ceilings $\gamma_1 = p$, $\gamma_2 = q$, $\gamma_3 = r$, all positive. In each block the ray is a real multiple of a primitive self-adjoint idempotent $P_j$. The weighted Fourier partial gradients are $N^{-1}B_j^* C_j$, $N^{-1}C_j A_j^*$ and $N^{-1}B_j A_j$. Since $P_j^* = P_j = P_j^2$ and different blocks do not multiply, their spatial versions on the three-ray span are exactly

$$\nabla_a f = \sum_{j=1}^{3} \gamma_j b_j c_j r_j, \qquad \nabla_b f = \sum_{j=1}^{3} \gamma_j a_j c_j r_j, \qquad \nabla_c f = \sum_{j=1}^{3} \gamma_j a_j b_j r_j.$$

These are linear combinations within the span, generally with several nonzero coefficients. Sphere projection and the scalar layer weights preserve the span, and $a = b$ is invariant there. This proves contraction closure for real, complex and quaternionic sectors, independently of their ceilings or matrix sizes.

Put $\vartheta = \arctan(p/q)$ and, near $x = y = u = v = 0$, write

$$a = b = \sqrt{1 - u^2}\big(\cos(\vartheta + x) r_1 + \sin(\vartheta + x) r_2\big) + u r_3, \tag{51}$$

$$c = \sqrt{1 - v^2}\big(\cos(\vartheta + y) r_1 + \sin(\vartheta + y) r_2\big) + v r_3. \tag{52}$$

All omitted coordinates are zero. With $\theta = \vartheta + x$ and $\phi = \vartheta + y$, the restricted potential is

$$\begin{aligned} V &= (1 - u^2)\sqrt{1 - v^2}\, V_0(\theta, \phi) + r u^2 v, \\ V_0(\theta, \phi) &= p \cos^2\theta \cos\phi + q \sin^2\theta \sin\phi. \end{aligned} \tag{53}$$

The induced metric is

$$2\lambda\left((1-u^2)\,\mathrm{d}x^2+\frac{\mathrm{d}u^2}{1-u^2}\right)+(1-v^2)\,\mathrm{d}y^2+\frac{\mathrm{d}v^2}{1-v^2}. \tag{54}$$

At the two-sector saddle, $E=V_0(\vartheta,\vartheta)=pq/\sqrt{p^2+q^2}>0$. Put $\beta=r/E$. Dividing the vector field by $E$ gives the exact equations

$$x'=\frac{\sqrt{1-v^2}}{2\lambda E}\,\partial_\theta V_0, \tag{55}$$

$$y'=\frac{1-u^2}{E\sqrt{1-v^2}}\,\partial_\phi V_0, \tag{56}$$

$$u'=\frac{u(1-u^2)}{\lambda}\big(\beta v-\sqrt{1-v^2}V_0/E\big), \tag{57}$$

$$v'=\beta u^2(1-v^2)-v(1-u^2)\sqrt{1-v^2}V_0/E. \tag{58}$$

Here primes refer only to this rescaled time. For $C_2\times C_2$, take $r_j=\chi_j/2$ and $p=q=r=1/8$. Then $\vartheta=\pi/4$, $\beta=\sqrt{2}$, and $V_0/E=\cos y+\sin(2x)\sin y$, the simplest scalar example.

### C.3 The stable graph and its first orientation-sensitive term

We are looking for the first term of the rival boundary that changes sign when the third sector's output block is reflected; everything of lower order is even in that coordinate and invisible to the reflection. The origin is the same mixed saddle for every $\lambda>0$. Its linearization is

$$x'=y/\lambda,\qquad y'=2x-y,\qquad u'=-u/\lambda,\qquad v'=-v.$$

Write

$$m=-\frac{\lambda+\sqrt{\lambda^2+8\lambda}}{2},\qquad m^2+\lambda m-2\lambda=0,\qquad \kappa=-1-m/\lambda>0.$$

There are three negative eigenvalues, $m/\lambda,-1/\lambda,-1$, and one positive eigenvalue $\kappa$. The smooth local stable manifold is therefore a graph

$$y=h_\lambda(x,u,v),\qquad Dh_\lambda(0)=(m,0,0).$$

The hyperbolic stable-manifold theorem gives a smooth local graph (Shub, 1987, Chapter 5); we need only its finite Taylor expansion.

Changing both third input signs sends $u$ to $-u$ and is a symmetry of the field. On the invariant subspace $u=0$, changing $v$ to $-v$ is also a symmetry. Uniqueness of the local stable manifold consequently gives

$$h_\lambda(x,-u,v)=h_\lambda(x,u,v),\qquad h_\lambda(x,0,-v)=h_\lambda(x,0,v).$$

The plane $x=y=0$ is invariant because both derivatives of $V_0$ vanish at $(\vartheta,\vartheta)$, and it is locally attracting within itself, so $h_\lambda(0,u,v)=0$. Put $H(u,v)=\partial_x h_\lambda(0,u,v)$. Its allowed terms through degree three are

$$H(u,v)=m+Au^2+Bv^2+Cu^2v+O\big((|u|+|v|)^4\big). \tag{59}$$

To compute them, write

$$U=\frac{u(1-u^2)}{\lambda}\big(\beta v-\sqrt{1-v^2}\big),\qquad W=\beta u^2(1-v^2)-v(1-u^2)\sqrt{1-v^2}.$$

Differentiate the graph-invariance equation $h_x x'+h_u u'+h_v v'=y'$ with respect to $x$ at $x=y=0$. The result is the exact scalar equation

$$H_uU+H_vW+\frac{\sqrt{1-v^2}}{\lambda}H^2=\frac{1-u^2}{\sqrt{1-v^2}}(2-H). \tag{60}$$

Matching the coefficients of $u^2,v^2,u^2v$, respectively, gives

$$(\lambda+2m-2)A=-m^2,\qquad (2m-\lambda)B=m^2,\qquad (1-m)C=\beta(A+\lambda B).$$

In particular,

$$A = -\frac{m^2}{\lambda + 2m - 2}, \qquad B = \frac{m^2}{2m - \lambda},$$

and eliminating them gives the orientation coefficient

$$C = (\lambda - 1)\,\widetilde{C}_{p,q,r}(\lambda), \qquad \widetilde{C}_{p,q,r}(\lambda) = \frac{\beta m^2(\lambda + 2m)}{(1-m)(\lambda + 2m - 2)(2m - \lambda)}, \tag{61}$$

whose factorization uses $m^2 = \lambda(2-m)$. Since $\lambda + 2m = -\sqrt{\lambda^2 + 8\lambda} < 0$, all denominator factors are nonzero, so $\widetilde{C}_{p,q,r}(\lambda) \neq 0$ for every $\lambda > 0$ and $C \neq 0$ exactly when $\lambda \neq 1$.

### C.4 Reflected initial states lie on opposite sides

Fix $\lambda \neq 1$. The invariant plane and the two symmetries above imply, by smooth factorization in $x, u, v$, that

$$h_\lambda(x, u, v) - h_\lambda(x, u, -v) = x u^2 v\big(2C + O(|x| + u^2 + v^2)\big). \tag{62}$$

For sufficiently small $\epsilon > 0$, set $x = u = \epsilon$, take $v_\pm = \pm\epsilon$, and use the same output angle

$$y_\epsilon = \tfrac{1}{2}\big(h_\lambda(\epsilon, \epsilon, \epsilon) + h_\lambda(\epsilon, \epsilon, -\epsilon)\big)$$

in both states. If $\eta = y - h_\lambda(x, u, v)$ is the signed transverse coordinate, these states satisfy

$$\eta_\pm = \mp C\epsilon^4 + O(\epsilon^5).$$

Their complete blocks in sectors 1 and 2 are identical, as are all sector masses. Only the sign of the third output coordinate changes. Unequal ceilings need not preserve sector-exchange symmetry, hence the $O(\epsilon^5)$ remainder. When $p = q$, that symmetry improves it to $O(\epsilon^6)$. At $x = 0$ the reflected gap vanishes identically. The symmetry at $u = 0$ requires only evenness in $v$, not independence of $v$; indeed $B \neq 0$ in (59).

### C.5 The two sides reach different pure-sector traps

In a sufficiently small neighborhood of the saddle, invariance of the stable graph factors the transverse equation as $\eta' = b(x, u, v, \eta)\eta$, with $b(0) = \kappa > 0$. Choose local coordinates $(\xi, \zeta)$ that straighten both the stable and unstable manifolds, with $\zeta$ a positive smooth multiple of $\eta$. They can be obtained by first using $\eta$ and then subtracting the unstable graph from the three stable coordinates. In a fixed small coordinate cylinder, the equations have the form

$$\xi' = A_s\xi + R_s(\xi, \zeta), \qquad \zeta' = b_s(\xi, \zeta)\zeta, \quad R_s(0, \zeta) = 0, \quad DR_s(0, 0) = 0, \quad \kappa/2 < b_s < 2\kappa,$$

where $A_s$ has strictly negative eigenvalues. In an equivalent Lyapunov norm, shrinking the cylinder gives $D^+\,\|\xi\| \leq -\alpha\,\|\xi\|$ for some $\alpha > 0$. Fix the cylinder $\|\xi\| < d$, $|\zeta| < d$ once, independently of $\epsilon$. For an initial point in its interior with $\zeta_0 \neq 0$, the sign of $\zeta$ is preserved and $|\zeta_0|e^{\kappa t/2} \leq |\zeta(t)| \leq |\zeta_0|e^{2\kappa t}$ until exit. Meanwhile $\|\xi(t)\| \leq e^{-\alpha t}\,\|\xi_0\| < d$, so exit can occur only through the corresponding unstable face. Its time obeys

$$\frac{1}{2\kappa}\log\frac{d}{|\zeta_0|} \leq t_{\text{exit}} \leq \frac{2}{\kappa}\log\frac{d}{|\zeta_0|}.$$

For the reflected states $|\zeta_0| \asymp \epsilon^4$ and $\|\xi_0\| = O(\epsilon)$, so $t_{\text{exit}} \to \infty$ and $\xi(t_{\text{exit}}) \to 0$. Their exit points converge to the two fixed points $(0, +d)$ and $(0, -d)$ on the local unstable manifold.

The plane $u = v = 0$ is invariant and contains the unstable eigendirection; uniqueness of the local unstable manifold puts both branches in that plane. The positive branch enters $x, y > 0$, and the negative branch enters $x, y < 0$. The phase portrait of Appendix B.7, valid for arbitrary $p, q$, sends these two branches to pure sectors 2 and 1, respectively. Starting from each fixed exit point, choose a finite further time at which that branch lies in its open pure-sector trapping neighborhood. Continuity of these two finite-time flow maps gives neighborhoods of the exit points with the same trap entry. Now choose $\epsilon$ small enough that the reflected exit points belong to those neighborhoods. Both reflected trajectories then enter the respective traps and converge to opposite pure sectors, so a later transition to the third sector cannot occur. The cylinder, reference exit points, and trap-entry times are all fixed before $\epsilon$ is chosen.

### C.6 AMBIENT-OPEN LIFTING AND THE ERROR LOWER BOUND

The maximum components can be nonisolated in matrix-valued sectors. The trapping neighborhoods on the full product manifold from Appendix A.7 apply to each entire compact aligned component for every $g_\lambda$. Finite-time entry persists on an ambient neighborhood by continuous dependence. Thus the two component basins $B_1, B_2$ are open even when the pure rays have unequal ceilings or lie in nonisolated maximum components. Define the ambient reflection

$$\iota(a, b, c) = \big(a, b, c - 2\Pi_3 c\big).$$

It is an isometry for every $g_\lambda$ and preserves $D_1, D_2$. The reflection $\mathcal{R}$ of the main text is this $\iota$. After choosing the member of the witness pair in $B_1$, the set $U = B_1 \cap \iota^{-1}(B_2)$ is nonempty and open; $U$ and $\iota U$ are disjoint. Every $X \in U$ has $S_\lambda(X) = 1$ and $S_\lambda(\iota X) = 2$.

The scores of sectors 1 and 2 take identical values on each reflected pair. Their fixed priority also agrees. Hence a sector-local rule cannot be correct at both points: choosing a third sector already incurs an error, and choosing 1 at the first point and 2 at the second would require two opposite orderings of the same scores. For the measurable error set $E_\sigma = \{S_\lambda \neq \sigma\}$, volume preservation gives

$$\begin{aligned} \mu_\lambda(E_\sigma) &\geq \mu_\lambda(E_\sigma \cap U) + \mu_\lambda(E_\sigma \cap \iota U) \\ &= \int_U \big(\mathbf{1}_{E_\sigma}(X) + \mathbf{1}_{E_\sigma}(\iota X)\big)\, \mathrm{d}\mu_\lambda(X) \geq \mu_\lambda(U) > 0. \end{aligned} \tag{63}$$

Thus every measurable rule errs with probability at least $\mu_\lambda(U)$ at this fixed metric: $\mathrm{err}^*_\lambda \geq \mu_\lambda(U) > 0$. The same contradiction also rules out a possibly nonmeasurable rule that is correct outside a volume-null set $Z$: choose $X \in U \setminus (Z \cup \iota^{-1} Z)$. The proof gives no uniform lower bound as the metric varies.

*Remark* C.1 (The largest set the pairing argument covers, and its factor of two). Nothing in (63) used the construction of $U$. It used only that $S_\lambda(X)$ and $S_\lambda(\iota X)$ are two different sectors, neither of them the reflected one: the scores of every other sector are equal at $X$ and $\iota X$, so a rule's best non-reflected candidate is the same at both points, and it must either answer the reflected sector somewhere, which is wrong there, or answer one candidate at both, which is wrong at one of them. Writing $\iota_j$ for the reflection of any designated nontrivial sector $j$, the same argument applies on

$$W_j = \big\{X :\ S_\lambda(X) \neq S_\lambda(\iota_j X),\ S_\lambda(X) \neq j,\ S_\lambda(\iota_j X) \neq j\big\}, \tag{64}$$

for any number of sectors. But $W_j$ is $\iota_j$-invariant, and each reflected pair inside it carries two states and one guaranteed error, so every measurable rule $\sigma$ in the class has $\mu_\lambda(\sigma \neq S_\lambda) \geq \frac{1}{2}\mu_\lambda(W_j)$:

$$\mathrm{err}^*_\lambda \ \geq\ \tfrac{1}{2}\mu_\lambda(W_j), \qquad \tfrac{1}{2}\mu_\lambda(W_3) \ \geq\ \mu_\lambda(U),$$

with the factor of two, and not $\mu_\lambda(W_j)$; the second inequality holds because $U \cup \iota U \subseteq W_3$ and $U \cap \iota U = \emptyset$. Each half $U'$ of $W_j$ exchanged by $\iota_j$ satisfies $U' \cap \iota_j U' = \emptyset$ and $\mu_\lambda(U') = \frac{1}{2}\mu_\lambda(W_j)$, since $\iota_j$ preserves $\mu_\lambda$. The same applies on $\mathcal{M}_1$ with $\mu_J$ and $S_J$. The bound holds for each rule separately: the set of starts on which a rule errs depends on the rule, and no single start is misclassified by every rule. Unlike the constructed $U$, the set (64) is defined by endpoint labels alone and can therefore be sampled; Appendix H reports an estimate of its measure.

For the actual joint architecture, apply $\Psi$ to the slice at $\lambda = 1/2$. Its image is invariant, and the potential and metric identities of Appendix B.7 identify the trajectories up to time rescaling. Every endpoint component on $\mathcal{M}_1$ is Morse–Bott (Proposition 4.1, Appendix D.4), so the label-safe tube $U$ of Appendix D, with the value gap (71) on its boundary, traps the whole component in $\mathcal{M}_1$, input-imbalance direction included, for every $g^J_\omega$: a trajectory in $U$ above the boundary value never reaches $\partial U$ because $f$ increases, and the only critical points in $U$ lie on the component. Finite-time trap entry therefore extends the two basins from the balanced submanifold to open sets in $\mathcal{M}_1$. The same output reflection is an isometry of $\mathcal{M}_1$ and preserves the allowed data of sectors 1 and 2, so the pairing argument applies there too. Let $\mu_J$ be normalized Riemannian volume on $\mathcal{M}_1$, equivalently the law of $((A, B)/\|(A, B)\|, C/\|C\|)$ for independent standard real Gaussian vectors $A, B, C$. For the resulting nonempty open paired set $U_J$, every measurable rule in the class errs against the joint selector $S_J$ with probability at least $\mu_J(U_J)$; writing $\mathrm{err}^*_J := \inf_\sigma \mathbb{P}_{\mu_J}[\sigma \neq S_J]$, this is $\mathrm{err}^*_J \geq \mu_J(U_J) > 0$, a separate quantity from $\mathrm{err}^*_\lambda$ on the product of spheres.

*Remark* C.2 (Upper bounds from the two comparators). The comparator rates bound the optimal errors from above. On $\widetilde{\mathcal{M}}$ the $\lambda = 1$ clock scores $s_j = (1+T_j)^{-1}$ form a rule in the class, so $\mathrm{err}^*_\lambda \le \mathbb{P}[S_\lambda \neq S_1]$; at $\lambda = \frac{1}{2}$ this is $q_{\mathrm{met}}$. On $\mathcal{M}_1$ apply the same clocks to $T(X) = (a/\|a\|, b/\|b\|, c)$. Sector $j$'s blocks of $T(X)$ are $\Pi_j a/\|a\|$, $\Pi_j b/\|b\|$ and $\Pi_j c$, so each score reads its own blocks and two layer norms only, and the rule is sector-local on $\mathcal{M}_1$. Since $T$ pushes $\mu_J$ to normalized volume on $\widetilde{\mathcal{M}}$, the rule equals $S_1 \circ T$ $\mu_J$-almost surely, and its error against $S_J$ is the paired disagreement $q_{\mathrm{op}}$ of the shared Gaussian coupling. Hence

$$\tfrac{1}{2}\mu_J(W_j) \;\le\; \mathrm{err}^*_J \;\le\; q_{\mathrm{op}}, \qquad \mathrm{err}^*_{1/2} \;\le\; q_{\mathrm{met}},$$

and the measurements of Section 6 and Appendix H.2 bracket $\mathrm{err}^*_J$ without locating it. Appendix H.3 gives a much better rule in the class and, on $C_2 \times C_2$, estimates $\mathrm{err}^*_J$ itself.

*Remark* C.3 (Initialization laws equivalent to volume). Let $\nu = \varrho\,\mu$ on either manifold, where $\mu$ is its normalized volume, and retain its paired open set $U$ and volume-preserving reflection $\iota$. Disjointness of $U, \iota U$ and the pointwise error inequality give

$$\nu(E_\sigma) \ge \int_U \big[\mathbf{1}_{E_\sigma}(X)\varrho(X) + \mathbf{1}_{E_\sigma}(\iota X)\varrho(\iota X)\big]\,\mathrm{d}\mu(X) \ge \int_U \min\{\varrho(X), \varrho(\iota X)\}\,\mathrm{d}\mu(X).$$

If $\varrho > 0$ almost everywhere, this last integral is positive: its integrand is positive almost everywhere on a set of positive volume. Thus every law equivalent to volume has a law-dependent class-wide floor. More generally, positive paired overlap suffices. Absolute continuity alone does not: a density supported in one pure-sector basin admits a perfect constant leading-label rule. No lower bound is uniform over these laws.

### C.7 The positive direction at $\lambda = 1$

For $\lambda = 1$, the independent sector clocks give the required sector-local rule. The common radial removal underlying this argument is classical in homogeneous competition systems (Hofbauer, 2002); here it can be checked directly, including signed initial coordinates. In weighted real Fourier coordinates consider the independent ambient systems

$$\frac{\mathrm{d}A_j}{\mathrm{d}\xi} = N^{-1}B_j^* C_j, \qquad \frac{\mathrm{d}B_j}{\mathrm{d}\xi} = N^{-1}C_j A_j^*, \qquad \frac{\mathrm{d}C_j}{\mathrm{d}\xi} = N^{-1}B_j A_j, \qquad j \in \mathcal{I}, \tag{65}$$

initialized at the Fourier blocks of $(a, b, c)$; the trivial coordinates stay constant. The real forms are closed under products and adjoints. Let $T_j \in (0, \infty]$ be the maximal forward existence time of this finite-dimensional polynomial ODE. It depends only on sector $j$'s block. Continuous dependence up to any finite time below the maximal lifetime makes $T_j$ lower semicontinuous, hence Borel measurable.

The three total squared norms have the same derivative and all start at one, so they remain equal. Writing that common norm as $R$ and $(A, B, C) = R(a, b, c)$ gives $R' = R^2 f(a, b, c)$ and $\mathrm{d}t/\mathrm{d}\xi = R$. The normalized triple therefore follows exactly the $g_1$ flow. By Theorem B.1, almost every normalized trajectory tends to a positive pure-sector maximum. Along such a trajectory $R(t) = \exp(\int_0^t f)$ eventually grows exponentially, and $\xi_* = \int_0^\infty R(t)^{-1}\,\mathrm{d}t < \infty$. It is the first ambient blow-up time.

Write $f_\infty > 0$ for the limiting normalized value. Since $\mathrm{d}(1/R)/\mathrm{d}\xi = -f$, integration up to $\xi_*$ gives

$$R(\xi) \sim \frac{1}{f_\infty(\xi_* - \xi)}.$$

For $Y_j = (A_j, B_j, C_j)$, use the combined weighted Hilbert norm $q_j = \|Y_j\|$. Quadraticity gives $\|Y_j'\| \le L_j q_j^2$ for a finite $L_j > 0$. If $T_j < \infty$, the continuation criterion for a polynomial ODE and integration of $|(1/q_j)'| \le L_j$ up to blow-up imply

$$q_j(\xi) \ge \frac{1}{L_j(T_j - \xi)}.$$

The winning sector has unbounded $q_j$ at $\xi_*$, hence $T_j = \xi_*$. If another sector also had $T_j = \xi_*$, its normalized norm $q_j/R$ would have positive lower limit, contradicting the pure-sector limit. Thus $S_1 = \arg\min_j T_j$ almost surely. The real scores $s_j = (1 + T_j)^{-1}$, with $s_j = 0$ at $T_j = \infty$, give the required sector-local representation for every nontrivial finite group; they need not have a closed form. Together with the negative direction for $|\mathcal{I}| \ge 3$, this proves the separate form of Theorem 3.1; Corollary C.4 below gives the joint form.

### C.8 Compensating the actual joint normalization

**Corollary C.4** (Compensating relative update rates)**.** *On $\mathcal{M}_1$, give the joint input sphere weight $\omega > 0$ and the output sphere weight one: $g^J_\omega = \omega g_{ab} + g_c$. For a group with at least three nontrivial real sectors, its leading selector is sector-local almost everywhere exactly when $\omega = 2$, in the same information class with fixed tie priority, and at every fixed $\omega \neq 2$ every measurable rule in the class errs with probability at least $\mu_J(U_\omega) > 0$ for a nonempty open $U_\omega \subset \mathcal{M}_1$. The positive direction holds for every nontrivial finite group.*

The weighted joint field is

$$(\dot{a}, \dot{b}, \dot{c}) = \big((G_a - 2fa)/\omega, (G_b - 2fb)/\omega, G_c - fc\big).$$

Analytic convergence and strict-saddle avoidance apply as before. For nonzero trivial output coordinate, the trivial-coordinate equations conserve $(\alpha^2 + \beta^2)/|\gamma|^{4/\omega}$; a zero output coordinate cannot reach the flat stratum. A flat limit therefore requires $\alpha(0)^2 + \beta(0)^2 = |\gamma(0)|^{4/\omega}$. For each input state this restricts the output coordinate to at most two levels, of sphere volume zero. Fubini makes the basin of the flat stratum null, so the joint selector is almost surely defined for each fixed $\omega$.

On the balanced submanifold, $\Psi^* g^J_\omega = g_\lambda$ with $\lambda = \omega/2$. For $\omega \neq 2$, the reflected witness above together with the full joint trapping neighborhoods gives the negative direction, including a positive-volume error floor. For $\omega = 2$, use the independent ambient block equations

$$A' = \tfrac{1}{2}\nabla_A f, \qquad B' = \tfrac{1}{2}\nabla_B f, \qquad C' = \nabla_C f.$$

Trilinearity gives $\frac{\mathrm{d}}{\mathrm{d}\xi}(\|A\|^2 + \|B\|^2) = \frac{\mathrm{d}}{\mathrm{d}\xi}\|C\|^2 = 2f(A, B, C)$. Their initially unit normalization blocks therefore have a common radius $R$. Removing it, with $\mathrm{d}t/\,\mathrm{d}\xi = R$, gives exactly the $g^J_2$ field. The positive pure-sector limit, the reciprocal radial asymptotic, and the quadratic block-norm bound then imply a unique earliest finite ambient lifetime. Those lifetimes are Borel and depend on each sector's own blocks, so $(1 + T_j)^{-1}$ are the required scores on the full joint manifold. In update-rate language, input weight $\omega = 2$ multiplies both input gradient prefactors by $1/2$ and leaves the output prefactor alone; it does not say that the output trajectory moves twice as fast, which would also depend on the gradient magnitudes. This compensates the normalization-induced imbalance for the leading flow; it is not an exact-CE factorization theorem, and Appendix E shows that no such theorem holds for three real linear characters at width one.

*Remark* C.5 (No uniform floor near the separate metric). For almost every initial point, the $g_1$ flow enters a pure-sector trap after finite time. Continuous dependence on $\lambda$ preserves this entry for all sufficiently close metrics, and the same trap serves them all. Dominated convergence therefore gives $\mathbb{P}[S_\lambda \neq S_1] \to 0$ as $\lambda \to 1$ under any fixed absolutely continuous law. Since $S_1$ is sector-local, $\mathrm{err}^*_\lambda \leq \mathbb{P}[S_\lambda \neq S_1] \to 0$. Positivity at every fixed $\lambda \neq 1$ does not mean a uniform positive floor near 1.

## D Exact-CE endpoint proofs

This appendix supplies the estimates behind Theorem 4.2. The group is fixed throughout; width is fixed except in Appendix F.2, which proves uniform averaged persistence over all finite widths. Every statement is in exact real arithmetic.

### D.1 Quantitative endpoint estimates

In dimensionless leading time, the exact field differs from the leading field by at most

$$\epsilon_s = N^{1/2} s^3 \sqrt{M(N+4)}. \tag{66}$$

For every fixed horizon $T$,

$$\sup_{0 \leq t \leq T} \big\| X^{\mathrm{CE}}_s(t) - X^{\mathrm{sh}}(t) \big\| \leq \frac{\epsilon_s}{5}(e^{5T} - 1), \tag{67}$$

$$\max_{k\delta_s \leq T} \big\| X^{\mathrm{CE}}_{s,k} - X^{\mathrm{sh}}(k\delta_s) \big\| \leq \frac{\epsilon_s + C(s)\delta_s}{5}(e^{5T} - 1). \tag{68}$$

Fix a connected leading maximum component $K_{\boldsymbol{\rho}}$ of value $f_{\boldsymbol{\rho}}$. Its Morse–Bott gap gives a tube $U_0$ and $c, c_0 > 0$ such that

$$\mathcal{F}(x) \le f_{\boldsymbol{\rho}} - c\,\mathrm{dist}(x, K_{\boldsymbol{\rho}})^2, \tag{69}$$

$$\|\mathrm{grad}\,\mathcal{F}(x)\| \ge c_0\,\mathrm{dist}(x, K_{\boldsymbol{\rho}}). \tag{70}$$

Choose a label-safe tube $U$ with $\overline{U} \subset U_0$. Compactness gives $\Delta > 0$ such that

$$\sup_{\partial U} \mathcal{F} \le f_{\boldsymbol{\rho}} - 5\Delta. \tag{71}$$

For small $s$, uniform convergence of the exact potential gives

$$\Phi_s \le f_{\boldsymbol{\rho}} - 4\Delta \text{ on } \partial U, \qquad \Phi_s \ge f_{\boldsymbol{\rho}} - \Delta \text{ on } K_{\boldsymbol{\rho}}. \tag{72}$$

Finally, the retracted CE iteration satisfies

$$\Phi_s(X_{k+1}) - \Phi_s(X_k) \ge \frac{\eta\lambda_s}{2}\,\|\mathrm{grad}\,\Phi_s(X_k)\|^2\,. \tag{73}$$

### D.2 An absolute Lipschitz constant for the leading field

For one neuron, use the dimensionless field

$$V = N^{3/2}(A - 2fa,\ B - 2fb,\ C - fc).$$

For a group-space tensor $W_{gh} = w_{gh}$, $\|W\|_F^2 = N\,\|w\|^2$, hence $\|W\|_{\mathrm{op}} \le \sqrt{N}\,\|w\|$. It follows that

$$|\Omega(u, v, w)| \le N^{-3/2}\,\|u\|\,\|v\|\,\|w\|$$

and the corresponding first and second partial derivatives have the same multilinear bounds.

Put $\alpha = \|a\|$, $\beta = \|b\|$, with $\alpha^2 + \beta^2 \le 1$, and $\|c\| \le 1$. After the $N^{3/2}$ time scaling, the operator norms of the nine derivative blocks are bounded entrywise by

$$M(\alpha, \beta) = \begin{pmatrix} 4\alpha\beta & 1 + 2\alpha^2 & \beta(1 + 2\alpha^2) \\ 1 + 2\beta^2 & 4\alpha\beta & \alpha(1 + 2\beta^2) \\ 2\beta & 2\alpha & 2\alpha\beta \end{pmatrix}.$$

The entries are nonnegative and monotone in the two norms. On $\alpha^2 + \beta^2 = 1$, its Frobenius norm satisfies

$$\|M(\alpha, \beta)\|_F^2 = 15 + 40\alpha^2\beta^2 \le 25.$$

At $\alpha = \beta = 1/\sqrt{2}$ the matrix is

$$\begin{pmatrix} 2 & 2 & \sqrt{2} \\ 2 & 2 & \sqrt{2} \\ \sqrt{2} & \sqrt{2} & 1 \end{pmatrix},$$

whose top eigenvalue is exactly five. Hence $\|DV\| \le 5$ throughout the convex hull of $\mathcal{M}_1$. The segment between two sphere points stays in that convex hull, so the mean-value inequality gives a global 5-Lipschitz bound. Separability carries the same constant to $\mathcal{M}_M$ in the product metric.

On $\mathcal{M}_1$, Euler's identity and the same multilinear bounds give

$$\|V\|^2 = N^3\big(\|A\|^2 + \|B\|^2 + \|C\|^2 - 5f^2\big) \le \alpha^2 + \beta^2 + \alpha^2\beta^2 \le \tfrac{5}{4}$$

per neuron. Below we use only the coarser bound $\|V\|^2 \le 5$. Thus

$$\|V(X)\| \le B := \sqrt{5M} \tag{74}$$

for the width-$M$ field.

### D.3 Continuous and discrete finite-horizon estimates

From (39), after division by $\lambda_s$ and the dimensionless time scaling,

$$\sup_X \left\| V_s^{\mathrm{CE}}(X) - V(X) \right\| \le \epsilon_s = N^{1/2} s^3 \sqrt{M(N+4)}.$$

If $D(t) = \left\| X_s^{\mathrm{CE}}(t) - X^{\mathrm{sh}}(t) \right\|$, then $D'(t) \le 5D(t) + \epsilon_s$ in the upper Dini sense. Gronwall gives (67).

For one unit sphere and tangent $v$,

$$\|\mathrm{Ret}_x(\delta v) - (x + \delta v)\| = \sqrt{1 + \delta^2 \|v\|^2} - 1 \le \frac{\delta^2}{2} \|v\|^2 .$$

On the product of spheres, summing the squared block defects and using $\sum_i p_i^2 \le (\sum_i p_i)^2$ gives a total one-step retraction defect at most $\delta^2 \|v\|^2 / 2$. The exact-CE field has norm at most $B + \epsilon_s$. The local truncation error of the leading flow is at most

$$C_F \delta^2, \qquad C_F = \frac{5}{2} B,$$

because the field is 5-Lipschitz and bounded by $B$. Consequently one recurrence step obeys

$$D_{k+1} \le (1 + 5\delta) D_k + \delta \epsilon_s + C(s) \delta^2,$$

where

$$C(s) = \frac{(B + \epsilon_s)^2}{2} + \frac{5}{2} B. \tag{75}$$

Discrete Gronwall proves (68). The two error sources add, so $s \to 0$ and $\delta \to 0$ can be taken independently at a fixed horizon.

**The explicit envelope is qualitative.** Collar entry in the endpoint proof requires the right side of (67) to stay below a fixed margin. Since $\epsilon_s \propto s^3$, the resulting scale threshold contains the factor $(e^{5T(X_0)} - 1)^{-1/3}$, and may therefore decay exponentially with the leading-flow entry time $T(X_0)$. The proof therefore yields no initialization-uniform threshold on the full-measure basin set and no rate for the label-error probability. The envelope is also quantitatively vacuous for the reported $D_4$ experiment: at its horizon $T = 60$ the factor $e^{5T} = e^{300}$ makes the bound uninformative even at the smallest tested scale. At $N = 8$, $M = 1$, $s = 0.20$ and $\eta = 0.25$, the right side of (68) is about $6.7 \times 10^{128}$, against a state-space diameter of $2\sqrt{2}$. These estimates establish fixed-horizon convergence as $s \downarrow 0$; they do not quantitatively explain the finite-scale agreement in Section 6. The exponential factor comes from Gronwall's inequality and may be far from sharp.

### D.4 The positive maximum components are Morse–Bott

In this subsection we use intrinsic $D$-matrix conventions: $\mathrm{Re}\,\mathrm{tr}\, X = \sum_i \mathrm{Re}\, X_{ii}$ and $\|X\|_F^2 = \sum_{ij} |X_{ij}|^2$. For $D = \mathbb{H}$, both quantities are half their standard-complex-embedding counterparts; that factor is absorbed into the omitted positive block scale.

It is enough to analyze one active $M_n(D)$ block, since width separates and Proposition 4.1 gives strict negative curvature in the balance, trivial-coordinate, and inactive-sector directions. Gauge an aligned rank-one point to $A = B = C = E_{11} := e_1 e_1^*$. For feasible $D$-linear tangent perturbations $X, Y, Z$, the intrinsic constrained Hessian, up to one positive block factor, is

$$\begin{aligned} Q(X,Y,Z) = {} & 2\,\mathrm{Re}\,\mathrm{tr}(Z^* Y E_{11}) + 2\,\mathrm{Re}\,\mathrm{tr}(Z^* E_{11} X) + 2\,\mathrm{Re}\,\mathrm{tr}(E_{11} Y X) \\ & - \|X\|_F^2 - \|Y\|_F^2 - \|Z\|_F^2 . \end{aligned}$$

Splitting $D^n = De_1 \oplus (De_1)^\perp$ yields the exact identity

$$\begin{aligned} Q = {} & - \|X_{\perp\perp}\|_F^2 - \|Y_{\perp\perp}\|_F^2 - \|Z_{\perp\perp}\|_F^2 \\ & - \sum_{i>1} \left( |y_{i1} - z_{i1}|^2 + |z_{1i} - x_{1i}|^2 + |y_{1i} - x_{i1}^*|^2 \right) \\ & - |z_{11} - y_{11} - x_{11}|^2 . \end{aligned} \tag{76}$$

For $D = \mathbb{H}$, this uses the real inner product $\mathrm{Re}(q^*p)$, norm polarization, and cyclicity of the real part of the quaternionic trace. No unrestricted complex tangent direction enters.

The zero equations in (76) are precisely the tangent equations obtained by differentiating

$$(p, q, r) \longmapsto (pq^*, rp^*, rq^*).$$

If $\beta_D = \dim_{\mathbb{R}} D$, the orbit is the quotient

$$(S^{\beta_D n-1})^3/S^{\beta_D-1}$$

and has dimension $3\beta_D n - \beta_D - 2$, equal to the nullity of the square map. Therefore

$$\ker \mathrm{Hess}\, \mathcal{F}(x) = T_x K_{\boldsymbol{\rho}}.$$

Compactness of $K_{\boldsymbol{\rho}}$ supplies a uniform negative normal gap. Applying the Morse–Bott lemma in finitely many tubular charts gives a single tube $U_0$ on which both (69) and (70) hold. Choose the label-safe collar $U$ with $\overline{U} \subset U_0$; shrinking $U$ and using compactness gives (71).

### D.5 Continuous trapping and convergence

The uniform value bound follows directly from (39):

$$\|\Phi_s - \mathcal{F}\|_\infty \le \frac{Ms^3}{2N}.$$

Choose $s$ so this is at most $\Delta$. Equations (72) follow. Define

$$A_s := U \cap \{\Phi_s \ge f_{\boldsymbol{\rho}} - 2\Delta\}.$$

By the boundary gap, $A_s \Subset U$. Monotonicity of $\Phi_s$ prevents a continuous trajectory in $A_s$ from reaching $\partial U$.

For almost every initialization, the leading trajectory converges to one $K_{\boldsymbol{\rho}}$. Choose nested inner tubes $K_{\boldsymbol{\rho}} \subset W' \Subset W \Subset U$, with $\mathcal{F} \ge f_{\boldsymbol{\rho}} - \Delta/4$ on $W$, and a finite time $T$ such that the leading state lies in $W'$ for every $t \ge T$. By uniform continuity of $\mathcal{F}$, choose the finite-horizon shadowing error small enough that the exact state remains in $U$ and

$$\mathcal{F}(X_s^{\mathrm{CE}}(T)) \ge f_{\boldsymbol{\rho}} - \Delta.$$

Since $\|\Phi_s - \mathcal{F}\|_\infty \le \Delta$, it follows that $\Phi_s(X_s^{\mathrm{CE}}(T)) \ge f_{\boldsymbol{\rho}} - 2\Delta$. Thus the exact state lies in $A_s$ at time $T$ and remains there.

For fixed $s > 0$, normalize every nonzero ambient sphere block by $R(z) = z/\|z\|$ and define $\widetilde{\Phi}_s = \Phi_s \circ R$ on a neighborhood of the manifold. On the manifold,

$$\nabla \widetilde{\Phi}_s = \mathrm{grad}_{\mathcal{M}_M} \Phi_s.$$

Apply the Euclidean Łojasiewicz gradient inequality to $\widetilde{\Phi}_s$ at an accumulation point. The standard finite-length argument then gives one critical endpoint in $U$.

### D.6 Discrete ascent, trapping, and convergence

For one sphere, the normalization retraction obeys

$$R_x(hv) = \frac{x + hv}{\sqrt{1 + h^2\|v\|^2}}, \qquad \|R_x(hv) - x\| \le h\|v\|.$$

Taking the tangent projection of $R_x(hv) = x$ gives $hv = 0$; hence a fixed point of the retracted CE update is exactly a critical point of $\Phi_s$.

Analyticity gives a uniform $C^2$ bound on $\Phi_s$ for all small $s$. The pullback

$$t \longmapsto \Phi_s(\mathrm{Ret}_X(t\,\mathrm{grad}\,\Phi_s(X)))$$

has second derivative bounded by $C_2\|\mathrm{grad}\,\Phi_s(X)\|^2$. If $h_s C_2 \le 1$, Taylor's theorem gives (73). The increase vanishes only when $\mathrm{grad}\,\Phi_s(X) = 0$.

The set $A_s$ stays a uniform positive distance from the complement of $U$. If $G_0$ bounds $\|\operatorname{grad}\Phi_s\|$ for small $s$, the condition $h_s G_0 < \operatorname{dist}(A_s, \mathcal{M}_M \setminus U)$ prevents an iterate from crossing the collar. Since $h_s \le 2\eta_{\max} s^3 N^2/M$, this holds uniformly for $0 < \eta \le \eta_{\max}$ when $s$ is small.

Use the same $W' \Subset W$ and $T$. For each $\delta_s > 0$, let $k_s = \lceil T/\delta_s \rceil$. Since $\delta_s \le 2\eta_{\max} s^3\sqrt{N}/M$, decreasing $s_0$ makes $\delta_s \le 1$ uniformly over $0 < \eta \le \eta_{\max}$; hence $k_s\delta_s \in [T, T+1]$. Apply (68) on the fixed horizon $T+1$. Its right-hand side tends to zero uniformly over that learning-rate interval. Once the finite-horizon error bound is below the spatial and energy margins of $W' \Subset W$, the iterate satisfies

$$\mathcal{F}(X^{\mathrm{CE}}_{s,k_s}) \ge f_{\boldsymbol{\rho}} - \Delta, \qquad \Phi_s(X^{\mathrm{CE}}_{s,k_s}) \ge f_{\boldsymbol{\rho}} - 2\Delta.$$

Thus $X^{\mathrm{CE}}_{s,k_s} \in A_s$, and the ascent estimate (73) keeps every later iterate there. The iteration count need not be uniform as $\eta \downarrow 0$, but the compared leading time is always at most $T+1$.

The compact trapped sequence has an accumulation point $x_*$. Normalize every nonzero ambient sphere block analytically by $R(z) = z/\|z\|$ and set $\bar\phi = -\Phi_s \circ R$ on a neighborhood of $x_*$. On the manifold its ambient gradient is the Riemannian gradient of $-\Phi_s$. Equation (73) and the step-length bound give the strong-descent inequality with constant $1/2$; equality of consecutive objective values can occur only at a fixed point. Boundedness of $\Phi_s$ together with (73) gives

$$\sum_k h_s \|\operatorname{grad}\Phi_s(X_k)\|^2 < \infty.$$

At fixed positive $s$ and $\eta$, $h_s > 0$, so the gradient tends to zero and $x_*$ is critical. Moreover, $\|X_{k+1} - X_k\| \le h_s \|\operatorname{grad}\Phi_s(X_k)\| \to 0$. Because $\bar\phi(X_k)$ is decreasing and $x_*$ is an accumulation point, $\bar\phi(X_k) \downarrow \bar\phi(x_*)$. The accumulation-point theorem and local Łojasiewicz finite-length argument of Absil et al. (2005, Theorem 3.4) therefore give $X_k \to x_*$, proving the discrete endpoint statement.

### D.7 Endpoint distance and uniform probability control

The collar was chosen inside $U_0$, so (70) holds throughout it. Direct differentiation of the exact remainder gives $\|\operatorname{grad}(\Phi_s - \mathcal{F})\| \le Cs^3$. At an exact critical endpoint,

$$\operatorname{dist}(X^{\mathrm{CE}}_\infty, K_{\boldsymbol{\rho}}) \le (C/c_0)s^3.$$

The estimate controls only normal distance; no tangent distance to the particular leading-flow endpoint follows.

Let $\mathcal{B}$ be the union of the leading-flow basins of all maximum components. It has full initialization measure. Fix $\zeta > 0$. Inner regularity gives a compact

$$K_\zeta \subset \mathcal{B}, \qquad \mathbb{P}[K_\zeta] \ge 1 - \zeta.$$

For every $x \in K_\zeta$, write $K_{\boldsymbol{\rho}(x)}$ for its selected maximum component. Choose nested leading-flow neighborhoods

$$K_{\boldsymbol{\rho}} \subset V'_{\boldsymbol{\rho}} \Subset V_{\boldsymbol{\rho}} \Subset U_{\boldsymbol{\rho}}$$

so that $V_{\boldsymbol{\rho}}$ is forward invariant and its closure lies strictly inside the corresponding exact-CE collar $U_{\boldsymbol{\rho}}$. Leading convergence gives a finite $T_x$ with $\mathrm{Fl}^{T_x}_V(x) \in V'_{\boldsymbol{\rho}(x)}$. By continuous dependence, shrink an open neighborhood $O_x$ within the same basin so that

$$\mathrm{Fl}^{T_x}_V(O_x) \subset V'_{\boldsymbol{\rho}(x)}.$$

Compactness supplies a finite subcover $K_\zeta \subset O_{x_1} \cup \cdots \cup O_{x_J}$. Put $T_\zeta = \max_j T_{x_j}$. Forward invariance then gives, for every $y \in K_\zeta$ and its label $\boldsymbol{\rho}(y)$,

$$\mathrm{Fl}^t_V(y) \in V_{\boldsymbol{\rho}(y)} \qquad (T_\zeta \le t \le T_\zeta + 1). \tag{77}$$

At fixed $G$ and $M$, only finitely many label components meet $K_\zeta$. Over those components, take the minimum spatial and energy margins, where the spatial margin is the distance from $\overline{V}_{\boldsymbol{\rho}}$ to the complement of $U_{\boldsymbol{\rho}}$, and the maximum perturbation, gradient, and finite-horizon shadowing constants. The resulting margins are positive and the constants finite. For $0 < \eta_{\max} < \infty$, choose one $s_0 = s_0(K_\zeta, \eta_{\max})$ so that, whenever $0 < s < s_0$,

$$\delta_s = \frac{2\eta s^3\sqrt{N}}{M} \le 1 \quad \text{for all } 0 < \eta \le \eta_{\max},$$

and the continuous and discrete shadowing errors on $[0, T_\zeta + 1]$ are below the common spatial and energy margins. The discrete choice

$$k_s = \left\lceil \frac{T_\zeta}{\delta_s} \right\rceil$$

satisfies $k_s \delta_s \in [T_\zeta, T_\zeta + 1]$ for every positive $\eta$, including arbitrarily small ones. Equation (77) and the uniform shadowing estimate therefore put every such iterate in the appropriate $A_{\boldsymbol{\rho},s}$. The uniform ascent and step-length bounds then keep every subsequent iterate in the appropriate trap.

Thus, for every $s < s_0$ and simultaneously every $0 < \eta \le \eta_{\max}$, the retracted endpoint label equals the leading label on all of $K_\zeta$. In particular,

$$\sup_{0<\eta\le\eta_{\max}} \mathbb{P}[S_{\mathrm{CE},s,\eta} \neq S_{\mathrm{joint}}] \le \mathbb{P}[K_\zeta^c] \le \zeta.$$

### D.8 Paired labels, finite-width counts, and symmetry

Three questions are answered in turn: how a label is made total on every trajectory, what law the disagreement count follows at fixed width, and why the automorphism group forces equal sector marginals.

To make every label total, adjoin an unresolved symbol $\perp$. An exact-CE trajectory receives the unique nontrivial maximizer of the canonical scores (7) at its limit, and $\perp$ if it does not converge or if that maximum is tied; each leading flow receives its sector at an aligned maximum and $\perp$ on its exceptional null set. These labels are measurable, since they are read from finitely many continuous scores at a pointwise limit, and on a collar covered by Theorem 4.2 the label is unique and is the collar's sector.

**Corollary D.1** (Exact-CE mismatch against the separate selector)**.** *Let $G$ be as in Proposition 3.2, let $M = 1$, and draw $X_J$ and $X_S = T(X_J)$ from the coupling of Corollary B.4. Write $S_{\mathrm{joint}}$ and $S_{\mathrm{sep}}^{\mathrm{real}}$ for the joint leading label of $X_J$ and the separate leading label of $X_S$, and $S_{\mathrm{CE},s,\eta}$ and $S_{\mathrm{CE},s}$ for the total retracted and continuous exact-CE labels from $X_J$. Then $q_{\mathrm{op}} = \mathbb{P}[S_{\mathrm{joint}} \neq S_{\mathrm{sep}}^{\mathrm{real}}] \ge \mathbb{P}[X_J \in U] > 0$ and, for every $0 < \eta_{\max} < \infty$, as $s \downarrow 0$,*

$$\sup_{0<\eta\le\eta_{\max}} \left|\mathbb{P}[S_{\mathrm{CE},s,\eta} \neq S_{\mathrm{sep}}^{\mathrm{real}}] - q_{\mathrm{op}}\right| \to 0,$$
$$\inf_{0<\eta\le\eta_{\max}} \mathbb{P}\big[S_{\mathrm{CE},s,\eta} = S_{\mathrm{joint}} \;\big|\; S_{\mathrm{joint}} \neq S_{\mathrm{sep}}^{\mathrm{real}}\big] \to 1,$$

*and likewise for continuous exact CE.*

**Corollary D.2** (Exactly equal leading-selector marginals on $C_2 \times C_2$)**.** *For $G = C_2 \times C_2$ under the same coupling, put $S_J = S_{\mathrm{joint}}$ and $S_S = S_{\mathrm{sep}}^{\mathrm{real}}$. For the three nontrivial characters $i, j$,*

$$\mathbb{P}[S_J = i] = \mathbb{P}[S_S = i] = \frac{1}{3}, \qquad \mathbb{P}[S_J = i, S_S = j] = \begin{cases} (1 - q_{\mathrm{op}})/3, & i = j, \\ q_{\mathrm{op}}/6, & i \neq j. \end{cases} \tag{78}$$

*The population marginal total variation is exactly zero despite $q_{\mathrm{op}} > 0$.*

For Corollary D.1, both leading labels exist almost surely, by Proposition 4.1 and Theorem B.1. Write $D_{\mathrm{op}} = \{S_{\mathrm{joint}} \neq S_{\mathrm{sep}}^{\mathrm{real}}\}$, so that $q_{\mathrm{op}} = \mathbb{P}[D_{\mathrm{op}}]$, and $D_{s,\eta} = \{S_{\mathrm{CE},s,\eta} \neq S_{\mathrm{sep}}^{\mathrm{real}}\}$. With the total labels just defined,

$$D_{s,\eta} \triangle D_{\mathrm{op}} \subseteq \{S_{\mathrm{CE},s,\eta} \neq S_{\mathrm{joint}}\}, \qquad 1 - \mathbb{P}[S_{\mathrm{CE},s,\eta} = S_{\mathrm{joint}} \mid D_{\mathrm{op}}] \le \frac{\mathbb{P}[S_{\mathrm{CE},s,\eta} \neq S_{\mathrm{joint}}]}{q_{\mathrm{op}}}.$$

The open witness set gives $q_{\mathrm{op}} > 0$, and the preceding compact-set argument makes both errors vanish uniformly over $0 < \eta \le \eta_{\max}$. The continuous statement follows in the same way. These events include the unresolved label $\perp$.

At fixed finite $M$, the leading field and the shared Gaussian initialization both separate across neurons, so the $M$ leading disagreement indicators are i.i.d. Bernoulli variables with parameter $q_{\mathrm{op}}$. Let $K_M$ be the sum of these indicators. Define $K_{M,s,\eta}^{\mathrm{CE}}$ by replacing, for each neuron, the joint leading label in that indicator with its total exact-CE label while keeping the same separate-flow label. The softmax

couples neurons at finite $s$; no independence is used for this second count. Under the shared Gaussian draw, equality of the whole exact-CE and joint leading sector vectors implies equality of the counts. The coupling inequality gives

$$d_{\mathrm{TV}}\big(\mathcal{L}(K^{\mathrm{CE}}_{M,s,\eta}), \mathrm{Binomial}(M, q_{\mathrm{op}})\big) \le \mathbb{P}[K^{\mathrm{CE}}_{M,s,\eta} \neq K_M] \le \mathbb{P}[\boldsymbol{S}_{\mathrm{CE},s,\eta} \neq \boldsymbol{S}_{\mathrm{joint}}].$$

The fixed-$M$ compact-set argument above applies to the whole sector vector, so the right side tends to zero uniformly over the bounded positive learning-rate interval. The same holds for continuous CE. This is a fixed-width small-scale limit, not a joint limit in $M$ and $s$.

For Corollary D.2, write $G = \mathbb{F}_2^2$ additively. Each automorphism $A \in GL(2, \mathbb{F}_2)$ induces the orthogonal coordinate permutation $(U_A x)_g = x_{A^{-1}g}$ on $\mathbb{R}^G$. It fixes $\mathbf{1}$ and commutes with $P$. Applied simultaneously in all three slots, it preserves each sphere norm and the joint input norm, hence both $\widetilde{\mathcal{M}}$ and $\mathcal{M}_1$. It is an isometry for the product round metrics and for every constant layer-weighted metric $\lambda g_a + \lambda g_b + g_c$ used in the joint-versus-separate comparison. Changing variables $g = Ag'$, $h = Ah'$ and using $A(g' + h') = Ag' + Ah'$ gives

$$f(U_A a, U_A b, U_A c) = f(a, b, c).$$

An invariant potential and an isometry have an equivariant gradient field; uniqueness of solutions therefore makes both leading flows equivariant.

Orthogonal invariance of the independent standard Gaussians makes their joint law invariant under $U_A$. Both normalization maps commute with $U_A$, and $T(U_A a, U_A b, U_A c) = U_A T(a, b, c)$ wherever $T$ is defined, a full-measure set. Thus the *paired* initialization law is invariant, not only its two marginals. If $A \cdot \chi = \chi \circ A^{-1}$, the real character projectors obey $U_A \Pi_\chi = \Pi_{A\cdot\chi} U_A$. The canonical scores and hence the selected labels are permuted accordingly. By Proposition 4.1 and Theorem B.1, no probability mass is left outside the three nontrivial labels.

Finally $GL(2, \mathbb{F}_2) \cong S_3$ acts as all permutations of the three nonzero dual vectors. Transitivity forces each marginal probability to be $1/3$. On ordered label pairs there are exactly two orbits: the three diagonal pairs and the six distinct ordered pairs, since the action is 2-transitive. The total masses of these orbits are $1 - q_{\mathrm{op}}$ and $q_{\mathrm{op}}$, proving (78). Each row and column sums to $(1 - q_{\mathrm{op}})/3 + 2q_{\mathrm{op}}/6 = 1/3$, and the off-diagonal entries sum to $q_{\mathrm{op}}$.

*Remark* D.3 (Field closeness alone gives no population rate). The preceding argument is qualitative in initialization, and an $O(\epsilon)$ analytic field perturbation need not move a degenerate basin boundary by $O(\epsilon)$. On $S^1$, fix an integer $m \ge 2$ and consider gradient ascent for

$$F_0(\theta) = \sin^{2m}\theta, \qquad F_\epsilon(\theta) = F_0(\theta) + \epsilon \sin\theta, \qquad 0 < \epsilon < 2m.$$

Both flows have maxima at $\pi/2$ and $3\pi/2$. The two minima, hence the two basin boundaries, move from $0, \pi$ to

$$-\alpha_\epsilon, \quad \pi + \alpha_\epsilon, \qquad \alpha_\epsilon = \arcsin\Big((\epsilon/(2m))^{1/(2m-1)}\Big),$$

because $F_\epsilon'(\theta) = \cos\theta(2m\sin^{2m-1}\theta + \epsilon)$. Although $\|F_\epsilon - F_0\|_{C^r} = O(\epsilon)$ for every fixed $r$, the uniform-initialization probability that the two flows choose different maxima is exactly $\alpha_\epsilon/\pi = \Theta(\epsilon^{1/(2m-1)})$. Thus setting $\epsilon = s^3$ rules out inferring an $O(s^3)$ label-error rate from field closeness alone. The example only shows that field closeness by itself gives no population rate; a model-specific rate would require quantitative boundary regularity or tube-mass control.

### D.9 Every fixed input weight

**Proposition D.4** (Every fixed input weight)**.** *Fix a nontrivial finite group, a finite width and $\omega > 0$, and let the initialization law be absolutely continuous with respect to Riemannian volume on $\mathcal{M}_M$. For almost every $X_0$ there is $s_0(X_0, \omega) > 0$ such that, for every $0 < s < s_0$, continuous exact CE in the metric $g^J_\omega$ converges to an endpoint with the neuronwise labels of the leading $g^J_\omega$ endpoint, at normal distance $O(s^3)$ from its maximum component. At width one, under $\mu_J$, the optimal error of measurable sector-local rules against the total continuous exact-CE endpoint label $S^{\mathrm{CE}}_{\omega,s}$ converges to* $\mathrm{err}^*(S^J_\omega)$ *as $s \downarrow 0$; at $\omega = 2$ its limit is zero.*

Theorem 4.2 is stated for the ordinary joint metric, and its continuous half uses the metric only through four facts. Each holds for $g^J_\omega$ at every fixed $\omega > 0$, with constants that depend on $\omega$.

1. *Comparison.* The metrics $g^J_\omega$ and $g^J_1$ are uniformly equivalent, with ratio between $\min(\omega, 1)$ and $\max(\omega, 1)$, and exact CE is the same function $\Phi_s$ in both. The $g^J_\omega$-gradients of $\Phi_s$ and $\mathcal{F}$ therefore differ by $O(s^3)$ uniformly on $\mathcal{M}_M$, the leading $g^J_\omega$ field is bounded and Lipschitz on the compact manifold, and the finite-horizon estimate (67) holds with $\omega$-dependent constants.
2. *Collars.* Critical sets do not depend on the metric, and at a critical point the Hessian is one bilinear form in every metric, on which a positive weight acts by congruence. Each leading maximum component therefore stays Morse–Bott, with the same kernel and normal inertia; (69) and (70) hold with $\omega$-dependent constants, and the label-safe tubes and value gaps (71) of Appendix D.4 can be kept.
3. *Leading convergence.* For fixed $\omega$, almost every leading $g^J_\omega$ trajectory converges to a maximum component, simultaneously for all neurons at finite width: analytic convergence and strict-saddle avoidance do not depend on the metric, and the flat stratum is handled by the conserved ratio $(\alpha^2 + \beta^2)/\,|\gamma|^{4/\omega}$ of Appendix C.
4. *Trapping and endpoints.* $\Phi_s$ increases along its own $g^J_\omega$-gradient flow, so a trajectory inside a collar above its boundary value never leaves it; analytic convergence gives an endpoint, and the critical equation with the gradient-distance gap gives normal distance $O(s^3)$.

With these, the continuous half of the proof of Theorem 4.2 and its compact-set argument (Appendix D.7) apply under $g^J_\omega$. This proves the first assertion of Proposition D.4, and shows that the probability that the total continuous exact-CE endpoint label $S^{\mathrm{CE}}_{\omega,s}$, defined as in Appendix D.8, differs from the leading label $S^J_\omega$ tends to zero as $s \downarrow 0$. No weighted retracted iteration is analyzed here.

For the second assertion, fix the width-one class $\mathcal{C}$ and write $R^*(S) = \inf_{\sigma\in\mathcal{C}} \mathbb{P}_{\mu_J}[\sigma \neq S]$ for a total label $S$. For every rule $\sigma$, pointwise,

$$\big|\mathbf{1}\{\sigma \neq S\} - \mathbf{1}\{\sigma \neq S'\}\big| \leq \mathbf{1}\{S \neq S'\},$$

so taking expectations, and then a near-minimizing rule for each target in turn, gives

$$|R^*(S) - R^*(S')| \leq \mathbb{P}[S \neq S'];$$

no minimizer need exist. With $S = S^{\mathrm{CE}}_{\omega,s}$ and $S' = S^J_\omega$ the right side tends to zero, so $R^*(S^{\mathrm{CE}}_{\omega,s}) \to \mathrm{err}^*(S^J_\omega)$. At $\omega = 2$ the clock rule of Section 3 lies in $\mathcal{C}$ and is exact almost everywhere (Corollary C.4), so the limit is zero; at every other fixed $\omega$, for groups with at least three nontrivial real sectors, it is positive by Theorem 3.1. Under the hypotheses of Theorem 5.1 the optimal error at $\omega = 2$ is nevertheless positive at every sufficiently small $s > 0$.

## E Finite-scale failure of leading compensation

This appendix proves Theorem 5.1 and Corollary E.1. Throughout, $G$ has three distinct nontrivial real linear characters $\chi_1, \chi_2, \chi_3$, the width is $M = 1$, and $\mathcal{M}_1 = S^{2N-1} \times S^{N-1}$ carries the compensated metric $g^J_2 = 2g_{ab} + g_c$ of Corollary C.4. The dynamics is continuous exact CE, the $g^J_2$-gradient flow of $\Phi_s$ in (40). Write $r_j = \chi_j/\sqrt{N}$, $\varepsilon = s^3$ and

$$\alpha = \frac{\varepsilon}{N^{3/2}}, \tag{79}$$

the size of the first finite-scale correction in the rescaled potential below. It is not the dimensionless step $\delta_s$ of the retracted iteration. The three characters are distinct and nontrivial; their product may be trivial or may be a fourth nontrivial character, and nothing below distinguishes the two cases. Section E.8 extends the conclusion to every input weight in a fixed compact subset of $(0, \infty)$, and so to a weight $\omega(s)$ retuned with the scale inside such a set. Nothing here is asserted for retracted descent, for sector triples that are not real linear characters, for $M \geq 2$, or for weights degenerating to 0 or $\infty$.

The proof deforms the reflected stable-graph argument of Appendix C in the scale. Its coefficient (61) vanishes at the compensated weight, consistently with the independence of the leading clocks there; that independence comes from the cubic potential being homogeneous and sector separable at matched rates, not from the vanishing of this one coefficient. Two separate objects are involved, the exact-CE stable graph and the stable graph of a truncated chart model, and Section E.4 says which identities belong to which and exactly what passes from the second to the first.

At zero scale the compensated rival boundary is unchanged by the third-character reflection. We differentiate that boundary with respect to the scale, show the first derivative does change under the reflection, and follow the two sides of the displaced boundary into different pure-character endpoints.

The mechanism at positive scale is a loss of homogeneity, not a mismatch of radial rates. The first exact-CE correction $Q_2$ in (80) is quartic in the joint input and quadratic in the output, so its ambient bidegree is $(4, 2)$, and at the compensated weight $\omega = 2$ its two ratios $4/2$ and $2/1$ agree with each other, just as the cubic term's ratios $2/2$ and $1/1$ do. In the unconstrained $g_2^J$ lift of $f - \varepsilon Q_2$, Euler's identity therefore gives both squared radii the same derivative $2f - 4\varepsilon Q_2$, so equal input and output radii persist, as they do for any loss of the logits. What fails is homogeneity: the two terms have total degrees 3 and 6, so on the equal-radius set of radius $R$ the normalized blocks follow the constrained flow of $f - \varepsilon R^3 Q_2$ rather than of $f - \varepsilon Q_2$. A common radius turns the fixed scale into a running one, and no single radial time change reduces the corrected potential at a fixed scale to independent sector systems, which is the construction the leading clocks use. Losing the construction is not yet losing the labels, which is why the rest of this appendix is about a boundary that moves.

### E.1 The first finite-scale correction

Let $\psi(w) = \log\big(N^{-1}\sum_k e^{w_k}\big)$ and, for centered $w \in \mathbb{R}^N$, $\mu_j = N^{-1}\sum_k w_k^j$. Since $\psi(0) = 0$ and the first cumulant of a centered vector vanishes, $\psi(w) = \mu_2/2 + O(\|w\|^3)$ near the origin. At width one the centered scale-free logits are $z^{(g,h)} = (a_g + b_h)^2 Pc$, so $R_s = \sum_{g,h} \psi(\varepsilon z^{(g,h)})$ with $\lambda_s = 2\varepsilon N^2$ in (38). Dividing by $\lambda_s$,

$$\Phi_s = f - \varepsilon Q_2 + O_{C^k}(\varepsilon^2), \qquad Q_2 = \frac{\|Pc\|^2}{4N^3}\sum_{g,h}(a_g + b_h)^4. \tag{80}$$

The logits are polynomial in the state and $\mathcal{M}_1$ is compact, so $\varepsilon z^{(g,h)}$ stays in a fixed ball on which $\psi$ is analytic; every term is a polynomial in the state times an analytic function of $\varepsilon z^{(g,h)}$, and the expansion therefore holds in each fixed $C^k$ norm, as already used for (40). Only (80) is needed below.

### E.2 Three real linear characters

Assuming the three sectors are real linear characters reduces the fourth-moment sum in $Q_2$ to character orthogonality, which is what makes the chart potential explicit below.

Write a balanced three-character state as

$$a = b = \frac{A}{\sqrt{2}}, \qquad A = \sum_{j=1}^{3} A_j r_j, \qquad c = \sum_{j=1}^{3} C_j r_j, \qquad \|A\| = \|C\| = 1, \tag{81}$$

and use the chart (51)–(52) of Appendix C with the two rivals $\chi_1, \chi_2$ and the third character $\chi_3$. For real linear characters $p = q = r = N^{-3/2}$, so $\vartheta = \pi/4$, $\beta = \sqrt{2}$, and the compensated joint weight pulls back to the family value $\lambda = 1$ in (54). Explicitly,

$$\begin{aligned}(A_1, A_2, A_3) &= \big(\sqrt{1-u^2}\cos\theta, \sqrt{1-u^2}\sin\theta, u\big), \quad \theta = \tfrac{\pi}{4} + x,\\ (C_1, C_2, C_3) &= \big(\sqrt{1-v^2}\cos\phi, \sqrt{1-v^2}\sin\phi, v\big), \quad \phi = \tfrac{\pi}{4} + y.\end{aligned} \tag{82}$$

Reflection means $v \mapsto -v$: it is the restriction of the ambient $\iota(a, b, c) = (a, b, c - 2\Pi_3 c)$ of Appendix C, an isometry of $g_2^J$ that preserves every sector mass and both rivals' complete blocks.

Character orthogonality gives the two sums the correction needs. Among four indices from $\{1, 2, 3\}$, the product $\chi_j\chi_k\chi_l\chi_{l'}$ is trivial exactly when they pair off: every other pattern cancels two equal characters and leaves $\chi_j\chi_k$ with $j \neq k$, which is nontrivial because the three characters are distinct. Hence

$$\sum_g A_g^4 = \frac{\sum_j A_j^4 + 3\big(1 - \sum_j A_j^4\big)}{N} = \frac{3 - 2\sum_j A_j^4}{N}. \tag{83}$$

This uses only $\sum_j A_j^2 = 1$, not $\chi_1\chi_2\chi_3 = \mathbf{1}$. With $a = b = A/\sqrt{2}$, $\sum_g A_g = 0$ and $\sum_g A_g^2 = 1$,

$$\sum_{g,h}(a_g + b_h)^4 = \frac{N}{2}\sum_g A_g^4 + \frac{3}{2} = 3 - \sum_j A_j^4. \tag{84}$$

The same orthogonality gives $f = T/(2N^{3/2})$ with $T = \sum_j A_j^2 C_j$, and $\|Pc\| = 1$, so (80) becomes

$$f - \varepsilon Q_2 = \frac{1}{2N^{3/2}}\Big(T + \frac{\alpha}{2}\sum_j A_j^4\Big) - \frac{3\varepsilon}{4N^3}. \tag{85}$$

Dropping the constant and the positive factor, the rescaled potential on the chart is

$$W_\alpha = T + \frac{\alpha}{2}\sum_j A_j^4. \tag{86}$$

On the chart $\|Pc\| = 1$, so the correction appears as the input-only quartic term of (86). The rest of this appendix shows that it moves a basin boundary rather than only a speed.

### E.3 The saddle is hyperbolic and does not move at first order

The whole argument runs at one saddle, so two things have to hold uniformly in the scale: that the saddle stays hyperbolic, and that it does not drift at first order, which would confound a boundary displacement with a moved reference point.

Let $p_0$ be the equal two-character mixed point, $x = y = u = v = 0$ in (82). It is a critical point of $f$ on $\mathcal{M}_1$ with value $E = f(p_0) > 0$, and by Appendix B.2 the constrained Hessian of $f$ there splits orthogonally into the input-balance direction, one trivial direction per slot, and a reduced block congruent to one half of the separate-sphere Hessian, with curvatures $-4E, -2E, -2E, -E$ on the first four summands. Trilinearity and block orthogonality decouple distinct sectors, so on the reduced block the three angular directions of the two active characters carry the amplitude form with curvatures proportional to $1, -2, -2$, a real linear character has no orientation direction, and every remaining nontrivial direction carries a negative multiple of the identity. Hence the inertia on $\mathcal{M}_1$ is

$$(n_+, n_-, n_0) = (1,\ 3N-3,\ 0), \tag{87}$$

so $p_0$ is an isolated hyperbolic saddle of index one. A positive metric weight acts on the Hessian by congruence and preserves (87).

The first correction does not move it. Write $Q_2 = \|Pc\|^2\, S(a,b)/(4N^3)$ with $S(a,b) = \sum_{g,h}(a_g + b_h)^4$. Then $\operatorname{grad}_c Q_2 = S\,Pc/(2N^3)$ is radial at $p_0$, because $c \perp \mathbf{1}$ there. At $p_0$ every $a_g \in \{0, \pm N^{-1/2}\}$, so $a_g^3 = a_g/N$, while $\sum_h b_h = \sum_h b_h^3 = 0$ and $\sum_h b_h^2 = 1/2$; hence

$$\partial_{a_g} S = 4\sum_h (a_g + b_h)^3 = 4\Big(N a_g^3 + 3a_g \sum_h b_h^2\Big) = 10\, a_g, \tag{88}$$

and symmetrically $\partial_{b_h} S = 10\, b_h$. Both input gradients are the same multiple of their layer, so the gradient is radial for the joint input sphere as well and the full constrained gradient of $Q_2$ vanishes at $p_0$. Therefore $\operatorname{grad}\Phi_s(p_0) = O(\varepsilon^2)$, and since $\operatorname{Hess} f(p_0)$ is invertible on $T_{p_0}\mathcal{M}_1$ by (87), the implicit-function theorem gives a unique nearby exact-CE critical point $p_\varepsilon = p_0 + O(\varepsilon^2)$, hyperbolic of index one for all small $\varepsilon$.

### E.4 Two stable graphs, and what transfers between them

Two different objects appear from here on, and the argument depends on not conflating them. The chart (82) is invariant under the leading flow but *not* under exact CE, because $\operatorname{grad} Q_2$ has components outside the three-character span. So we carry both:

$$\begin{aligned} h_\alpha^{\mathrm{CE}} &: \quad \text{the intersection with the chart of the exact-CE stable hypersurface } W_\alpha^s \text{ of } p_\varepsilon, \\ \widehat{h}_\alpha &: \quad \text{the stable graph of the chart gradient flow of the truncated potential } W_\alpha \text{ of (86).} \end{aligned} \tag{89}$$

For small $\alpha$ each of $W_\alpha^s$ and the truncated chart stable manifold meets the $y$ direction transversally near the origin, so both are graphs $y = h_\alpha^{\mathrm{CE}}(x,u,v)$ and $y = \widehat{h}_\alpha(x,u,v)$. The exact symmetries and the exact scalar equation below are properties of $\widehat{h}_\alpha$. What transfers to $h_\alpha^{\mathrm{CE}}$ is $h_0^{\mathrm{CE}} = \widehat{h}_0$ together with one derivative in $\alpha$, and that is all the proof uses.

**Regularity.** Fix a finite $k$. After shrinking the neighborhood if necessary, both graphs are jointly $C^k$ in the spatial variables and in $\alpha$, and jointly $C^k$ in $(\alpha, \lambda)$ as well when the metric weight varies over a compact subset of $(0, \infty)$, as it does in Section E.8. We apply the stable manifold theorem to a fixed positive-time map of the field, which is analytic in the state and in the parameters. Abbondandolo & Majer (2006, Theorem A) state it for a single map; their proof (Abbondandolo & Majer, 2006, Sections 2–3) obtains the local stable manifold from the implicit-function theorem on a space of sequences, and running that argument with $(\alpha, \lambda)$ as additional variables gives a local stable manifold of the same finite regularity class jointly in the state and the parameters. The one input it needs is an inverse of the linearized sequence equation bounded uniformly in the parameters, which the uniform hyperbolic gap at $p_\varepsilon$ recorded in (87) supplies for $(\alpha, \lambda)$ in a compact set. Transversality then yields the graph by the implicit-function theorem with the same regularity. We use $k \geq 8$: the coefficient extracted in (97) is a mixed derivative of total order five in $(x, u, v, \alpha)$, the one in (98) has total order seven, and every Taylor remainder below is the corresponding $C^k$ remainder on a fixed compact neighborhood. Lower regularity would not license those coefficients.

**Symmetries of the truncated graph.** Both symmetries of Appendix C survive the correction *in the truncated chart model*. Simultaneous sign change of the third-character component of $a$ and $b$ sends $u \mapsto -u$ and leaves $T$ and $\sum_j A_j^4$ unchanged; on the invariant subspace $u = 0$, $v \mapsto -v$ leaves both unchanged as well. Exchanging the two rivals, $(x, y) \mapsto (-x, -y)$, also preserves $W_\alpha$, because $\cos^4\theta + \sin^4\theta$ and $T$ are invariant under $\theta \mapsto \pi/2 - \theta$, $\phi \mapsto \pi/2 - \phi$. Hence the plane $x = y = 0$ is invariant for the truncated chart flow at every $\alpha$, and uniqueness of the local stable manifold gives

$$\widehat{h}_\alpha(0, u, v) = 0, \qquad \widehat{h}_\alpha(x, -u, v) = \widehat{h}_\alpha(x, u, v), \qquad \widehat{h}_\alpha(x, 0, -v) = \widehat{h}_\alpha(x, 0, v). \tag{90}$$

Put $\widehat{H}_\alpha(u, v) = \partial_x \widehat{h}_\alpha(0, u, v)$.

None of (90) is asserted for $h_\alpha^{\mathrm{CE}}$ at a fixed positive $\alpha$, and the last identity is in general false for it. Restricted to the chart, the exact-CE potential carries at order $\alpha^2$ a term proportional to $C_1C_2C_3$ whenever $\chi_1\chi_2\chi_3 = \mathbf{1}$, through the third cumulant $N^{-1}\sum_k (Pc)_k^3 = 6C_1C_2C_3N^{-3/2}$ of the centered output; that term is odd in $v$ and survives on $u = 0$. An exact reflection symmetry of the full exact-CE graph would need a separate argument, and nothing below uses one.

**First-order transfer.** At $\alpha = 0$ the two flows agree on the chart, so

$$h_0^{\mathrm{CE}} = \widehat{h}_0. \tag{91}$$

Their first variations in the scale agree as well. Split the ambient tangent directions at a chart point into the four chart directions and the transverse ones $w$, and write the exact-CE stable hypersurface near $p_\varepsilon$ as a graph $y = \mathcal{H}(x, u, v, w, \alpha)$, so that $h_\alpha^{\mathrm{CE}}(x, u, v) = \mathcal{H}(x, u, v, 0, \alpha)$. Two factors vanish on $\{\alpha = 0,\ w = 0\}$.

- $\dot{w} = 0$ there, because the chart is invariant for the leading flow.
- $\partial_w \mathcal{H} = 0$ there. Take $w$ to be Fermi coordinates of $g_\omega^J$ normal to the chart, so that the isometries below act linearly on $w$ and, at $w = 0$, the chart components of the ambient gradient are the gradient of the restriction in the induced metric. For each sector $\rho$ absent from the chart, the trivial one included, the leading potential is invariant under the sign changes $(A_\rho, B_\rho) \mapsto -(A_\rho, B_\rho)$ and $(A_\rho, C_\rho) \mapsto -(A_\rho, C_\rho)$. The remaining transverse directions are the three input-imbalance directions $(r_j, -r_j, 0)$. For them use $\Theta(a, b, c) = (b^\star, a^\star, c^\star)$ with $x^\star(g) = x(g^{-1})$: the substitution $(g, h) \mapsto (h^{-1}, g^{-1})$ gives $\Omega(b^\star, a^\star, c^\star) = \Omega(a, b, c)$, and $P$ commutes with $x \mapsto x^\star$, so $f \circ \Theta = f$; $\Theta$ permutes coordinates within the joint input sphere and within the output sphere, so it is an isometry of $g_\omega^J$ for every $\omega$; and real linear characters satisfy $\chi_j(g^{-1}) = \chi_j(g)$, so $\Theta$ fixes the chart pointwise and sends $(r_j, -r_j, 0)$ to its negative. (For Abelian $G$ the exchange $(a, b, c) \mapsto (b, a, c)$ also works; for $D_4$ and $Q_8$ it does not preserve $f$.) All these maps fix $p_0$ and the chart pointwise, act linearly on $w$ and preserve the leading stable hypersurface, and no nonzero transverse vector is fixed by all of them; hence $\partial_w \mathcal{H} = 0$.

Differentiate the graph-invariance equation

$$(\partial_x \mathcal{H})\dot{x} + (\partial_u \mathcal{H})\dot{u} + (\partial_v \mathcal{H})\dot{v} + (\partial_w \mathcal{H}) \cdot \dot{w} = \dot{y}$$

once in $\alpha$ at $\alpha = 0$, $w = 0$. The transverse term contributes $(\partial_\alpha \partial_w \mathcal{H}) \cdot \dot{w} + (\partial_w \mathcal{H}) \cdot \partial_\alpha \dot{w} = 0$: the first summand meets the zero velocity, and the second meets the zero derivative, so the transverse $Q_2$

forcing $\partial_\alpha \dot{w}$ drops out. What remains is the first-order equation of the truncated chart model, whose forcing is the tangential projection of $\operatorname{grad} Q_2$, by the choice of Fermi coordinates.

That equation has only one solution near the origin. Write $z = (x, u, v)$ and $h_0 = h_0^{\mathrm{CE}} = \widehat{h}_0$, and let $b(z)$ be the $z$-components of the leading chart field at $(z, h_0(z))$, the leading flow on its stable graph. Both $\varphi = \partial_\alpha h_\alpha^{\mathrm{CE}}|_{\alpha=0}$ and $\widehat{\varphi} = \partial_\alpha \widehat{h}_\alpha|_{\alpha=0}$ are $C^{k-1}$ and solve

$$D\varphi \cdot b = c\,\varphi + F, \qquad c = \partial_y \dot{y} - Dh_0 \cdot \partial_y \dot{z},$$

with the leading field in $c$ and the same forcing $F$. The linearization of $b$ at the origin is diagonal with the three stable rates of the chart flow at $p_0$, all negative, so a small ball $B$ is forward invariant under $b$ and every $b$-orbit in $B$ tends to the origin, while $c(0) = \kappa_* > 0$ is the unstable rate of $p_0$. The difference $\delta = \varphi - \widehat{\varphi}$ satisfies $\frac{\mathrm{d}}{\mathrm{d}t}\delta(z(t)) = c(z(t))\,\delta(z(t))$ along every $b$-orbit. At $z = 0$, where $b$ vanishes, this gives $\delta(0) = 0$, so $\delta(z(t)) \to 0$; but $\delta(z(0)) \neq 0$ would make $|\delta(z(t))|$ eventually grow at least like $e^{\kappa_* t/2}$. Hence $\delta = 0$ on $B$, that is,

$$\partial_\alpha h_\alpha^{\mathrm{CE}}\big|_{\alpha=0} = \partial_\alpha \widehat{h}_\alpha\big|_{\alpha=0}, \tag{92}$$

as functions on the fixed ball $B$, and therefore coefficient by coefficient in their Taylor expansions. This is a statement about one derivative in $\alpha$, not a claim that the chart is exact-CE invariant, and it holds whether $\chi_1\chi_2\chi_3$ is trivial or is an additional nontrivial character.

### E.5 The reflection-odd coefficients of the truncated graph

Everything in this subsection is a computation on $\widehat{h}_\alpha$; it is carried to $h_\alpha^{\mathrm{CE}}$ only through (91)–(92).

Take the $g_2^J$-gradient of (86) in the chart and rescale time by a constant. With $e = \sqrt{2}\alpha$, the linearization in the rival coordinates and the base field in the third-character coordinates are

$$\begin{aligned} L_{xx} &= e(1-u^2), & L_{xy} &= \sqrt{1-v^2}, \\ L_{yx} &= \frac{2(1-u^2)}{\sqrt{1-v^2}}, & L_{yy} &= -\frac{1-u^2}{\sqrt{1-v^2}}, \\ U &= u(1-u^2)\Big[\sqrt{2}v - \sqrt{1-v^2} + \tfrac{e}{2}(3u^2-1)\Big], \\ V &= \sqrt{2}u^2(1-v^2) - (1-u^2)v\sqrt{1-v^2}. \end{aligned} \tag{93}$$

At $e = 0$ these are (55)–(58) with $\lambda = 1$, $p = q = r$ and $\beta = \sqrt{2}$, up to that constant time change; the terms in $e$ are the tangential $Q_2$ forcing allowed by (92). Differentiating the invariance equation $\widehat{h}_x x' + \widehat{h}_u u' + \widehat{h}_v v' = y'$ in $x$ at $x = 0$, where (90) gives $\widehat{h}_\alpha = 0$ and $\partial_u \widehat{h}_\alpha = \partial_v \widehat{h}_\alpha = 0$, leaves the exact scalar equation for the truncated model,

$$\widehat{H} L_{xx} + \widehat{H}^2 L_{xy} + \partial_u \widehat{H}\, U + \partial_v \widehat{H}\, V = L_{yx} + L_{yy}\widehat{H}. \tag{94}$$

Writing $\widehat{H} = m + Au^2 + Bv^2 + Cu^2 v + O((|u| + |v|)^4)$ as in (59) and matching the constant, $u^2$, $v^2$ and $u^2 v$ coefficients gives

$$\begin{aligned} m^2 + (1+e)m - 2 = 0, \qquad m(0) &= -2, \\ (2m-1)A &= -m^2, \\ 2(e+2m-1)B &= m^2 - m + 2, \\ (2m-2)C &= -2\sqrt{2}(A+B). \end{aligned} \tag{95}$$

At $e = 0$ they give $(A, B, C) = (4/5, -4/5, 0)$, the compensated case of (61), and differentiating in $e$ at zero gives

$$(m', A', B', C') = \Big(-\frac{2}{3}, \frac{8}{25}, -\frac{7}{25}, \frac{\sqrt{2}}{75}\Big). \tag{96}$$

Since $e = \sqrt{2}\alpha$, the first scale derivative of the reflection-odd coefficient of the truncated graph is

$$[\alpha u^2 v]\, \partial_x \widehat{h}_\alpha(0, u, v) = \frac{2}{75} \neq 0, \tag{97}$$

the bracket being a Taylor coefficient without factorials. Carrying the same matching through degree five in $(u, v)$ gives the next reflection-odd coefficient,

$$[\alpha u^2 v^3]\, \partial_x \widehat{h}_\alpha(0, u, v) = \frac{1327}{73500}, \tag{98}$$

which Section E.8 needs and the rest of this section does not. These are identities between finitely many Taylor coefficients; no fit and no small-denominator estimate enters. Both belong to this chart at this saddle and are not coefficients at an arbitrary section.

By (92) the same two numbers are the first scale derivatives of the corresponding coefficients of $h_\alpha^{\mathrm{CE}}$. In particular

$$\partial_\alpha \left[x u^2 v\right] h_\alpha^{\mathrm{CE}}\big|_{\alpha=0} = \frac{2}{75}, \tag{99}$$

so the reflected exact-CE graph difference $h_\alpha^{\mathrm{CE}}(x, u, v) - h_\alpha^{\mathrm{CE}}(x, u, -v)$ has first scale derivative $4xu^2v/75 + \cdots$ at $\alpha = 0$.

### E.6 Reflected pairs at every sufficiently small positive scale

At $\alpha = 0$ the rival boundary is unchanged by the reflection: on a neighborhood of the origin,

$$h_0^{\mathrm{CE}}(x, u, -v) = h_0^{\mathrm{CE}}(x, u, v). \tag{100}$$

This is not a symmetry of the leading field, since $\iota$ is not one; the clocks force it. Write $z = (x, u, v)$ and $\bar{z} = (x, u, -v)$, and suppose that $h_0^{\mathrm{CE}}(z_n) \neq h_0^{\mathrm{CE}}(\bar{z}_n)$ for points $z_n \to 0$. With $y_n$ strictly between the two values, the chart states $X_n = (z_n, y_n)$ and $\iota X_n = (\bar{z}_n, y_n)$ lie on opposite sides of the leading stable hypersurface of $p_0$, at transverse distance at most $\left|h_0^{\mathrm{CE}}(z_n) - h_0^{\mathrm{CE}}(\bar{z}_n)\right| \to 0$. By the exit argument of Appendix C and the branch convergence of Appendix B.7, for large $n$ the leading flow at $\omega = 2$ carries $X_n$ and $\iota X_n$ into the traps of the two different rivals, and by continuity of the finite-time flow the same holds on an open neighborhood $U_n \subset \mathcal{M}_1$ of $X_n$ and on its image $\iota U_n$. The pairing argument of (63) then gives every sector-local rule an error of at least $\mu_J(U_n) > 0$ against the leading selector at $\omega = 2$, contradicting the clock rule of Corollary C.4, which is exact almost everywhere.

By (99) the function

$$D(x, u, v) = \partial_\alpha \left[h_\alpha^{\mathrm{CE}}(x, u, v) - h_\alpha^{\mathrm{CE}}(x, u, -v)\right]\big|_{\alpha=0}$$

is not identically zero on any neighborhood of the saddle: its Taylor expansion contains $4xu^2v/75$. Fix one sufficiently small triple $(x_0, u_0, v_0)$ with $D(x_0, u_0, v_0) \neq 0$, and only then let $\alpha$ decrease. By (100) and joint $C^1$ dependence on $(x, u, v, \alpha)$, the reflected difference at that triple is $\alpha D(x_0, u_0, v_0) + o(\alpha)$, which gives $\alpha_0 > 0$ such that

$$h_\alpha^{\mathrm{CE}}(x_0, u_0, v_0) \neq h_\alpha^{\mathrm{CE}}(x_0, u_0, -v_0) \qquad \text{for all } 0 < \alpha < \alpha_0, \tag{101}$$

and shows that the difference tends to zero with $\alpha$. The chart point is chosen before the scale, so no limit is exchanged, and nothing above assumes that the displayed monomial is the whole displacement. Choosing $y_0$ strictly between the two graph values places the state at $(x_0, u_0, v_0, y_0)$ and its reflection $(x_0, u_0, -v_0, y_0) = \iota X$ on opposite local sides of $W_\alpha^s$, each within transverse distance $O(\alpha)$ of it. Only $h_\alpha^{\mathrm{CE}}$ enters here; $\widehat{h}_\alpha$ has done its work in (92).

### E.7 Labels and the class-wide floor

The two sides keep their labels. Near the hyperbolic saddle $p_\varepsilon$, straighten the stable and unstable manifolds as in Appendix C: in a fixed small cylinder the transverse coordinate satisfies $\zeta' = b\zeta$ with $b$ bounded away from zero, the stable coordinates contract, and a trajectory off the graph can leave only through the corresponding unstable face, points nearer the graph merely staying longer. No uniform residence time is needed. The exact-CE field is $C^2$-close to the leading one on a fixed neighborhood by (80), so the cylinder, its constants and the two exit faces can be fixed once for all small $\alpha$.

Both leading unstable branches lie in the invariant plane $u = v = 0$ and converge to the two pure-character maxima, by the branch convergence of Appendix B.7. Fix one exit point on each

branch, a finite later time at which it lies in the label-safe collar of that maximum component, and a neighborhood of the exit point, of radius independent of $\alpha$, whose exact-CE trajectories lie in the collar at that time for all small $\alpha$, by finite-time continuous dependence on that one fixed horizon. Those collars are built in Appendix D from the Morse–Bott value gap and the uniform bound $\|\Phi_s - f\|_\infty \le s^3/(2N)$, and their forward invariance uses only monotonicity of $\Phi_s$ along the flow, so the same collar construction applies for the fixed positive weight $\omega = 2$, with finite-horizon comparison constants that depend on it. The witness and its reflection lie in the cylinder within transverse distance $O(\alpha)$ of $W^s_\alpha$, so by the exit estimate of Appendix C, in the cylinder fixed above for all small $\alpha$, their exit points converge as $\alpha \downarrow 0$ to the two exact-CE branch points, and hyperbolic persistence keeps those within $O(\alpha)$ of the leading ones. For all sufficiently small $s$ both exits therefore lie in the two neighborhoods and enter the two collars, after which monotonicity of $\Phi_s$ prevents escape; the endpoint labels are the two collars' characters. Shrinking $\alpha_0$ if necessary, fix $s_*$ with $s_*^3 = N^{3/2}\alpha_0$.

Continuity of the flow and of the graph makes the two label sets open. Let $U_s \subset \mathcal{M}_1$ be a full-dimensional open neighborhood of the witness carrying the first character's label, small enough that $\iota U_s$ carries the second character's label; $U_s$ and $\iota U_s$ are disjoint. The trap entry is a finite-time open condition on $\mathcal{M}_1$, so this is an open set of the full joint manifold, not of the chart alone, and it includes unbalanced inputs and perturbations in other sectors.

Let $\mathcal{C}$ be the sector-local class of Theorem 3.1 on $\mathcal{M}_1$, with its fixed tie priority, and $S^{\mathrm{CE}}_{2,s}$ the total exact-CE endpoint label of Appendix D.8 in the metric $g^J_2$. On each pair $X \in U_s$, $\iota X$, the two rivals' scores and their priority are identical, because $\iota$ preserves both rivals' complete blocks and the entire mass vector; the true labels differ. No rule in $\mathcal{C}$ can therefore be correct at both points: selecting a third sector is already an error at both, and selecting the first at $X$ and the second at $\iota X$ would require two opposite orderings of the same scores. Both witnesses have resolved endpoints in strict pure-character traps, so the bound does not depend on how $S^{\mathrm{CE}}_{2,s}$ is completed elsewhere. Since $\iota$ preserves $\mu_J$, the pointwise inequality

$$\mathbf{1}_{\{\sigma \neq S^{\mathrm{CE}}_{2,s}\}}(X) + \mathbf{1}_{\{\sigma \neq S^{\mathrm{CE}}_{2,s}\}}(\iota X) \ge 1, \qquad X \in U_s, \tag{102}$$

integrates to $\mathbb{P}_{\mu_J}[\sigma \neq S^{\mathrm{CE}}_{2,s}] \ge \mu_J(U_s) > 0$ for every $\sigma \in \mathcal{C}$, which proves Theorem 5.1. The witness, the open set and the floor may all depend on $s$, and the rule may be chosen after $s$; the argument supplies no value for $\mu_J(U_s)$, no lower bound uniform in $s$ and no rate as $s \downarrow 0$. Allowing the scores to read $s$ and the metric does not evade it, since the contradiction is at a fixed scale.

### E.8 Other input weights, uniformly on compact sets

The preceding subsections treat the compensated weight. Repeating the argument at one other weight at a time would still leave open the possibility that a weight *retuned with the scale* restores exact sector-local rankability. It does not, as long as the retuning stays in a fixed compact set of positive weights. The reason is that two independent reflection-odd coefficients have to vanish at once, and one scalar knob cannot arrange both.

**Two coefficients as functions of the weight and the scale.** Give $\mathcal{M}_1$ the metric $g^J_\omega$ and write $\lambda = \omega/2$, so the chart metric is (54). All of Sections E.4–E.5 goes through with $\lambda$ in place of 1: the saddle $p_0$ is the same point for every $\lambda > 0$, its inertia is (87) because a positive weight acts on the Hessian by congruence, the constrained gradient of $Q_2$ still vanishes there, and both graphs are jointly $C^k$ in $(x, u, v, \alpha, \lambda)$ on a compact weight set. Set

$$c_1(\lambda, \alpha) = [xu^2 v]\, h^{\mathrm{CE}}_{\lambda,\alpha}, \qquad c_2(\lambda, \alpha) = [xu^2 v^3]\, h^{\mathrm{CE}}_{\lambda,\alpha}, \tag{103}$$

Taylor coefficients of the exact-CE graph at the actual parameter values. Both monomials are odd in $v$, so if either coefficient is nonzero then $h^{\mathrm{CE}}_{\lambda,\alpha}(x, u, v) - h^{\mathrm{CE}}_{\lambda,\alpha}(x, u, -v)$ is not the zero germ at the origin.

Four values determine the pair near the compensated point. At $\alpha = 0$ the exact-CE flow *is* the leading flow, so $c_1(\lambda, 0)$ and $c_2(\lambda, 0)$ are leading quantities; $c_1(\lambda, 0) = C$ of (61) at $p = q = r$, and both vanish at $\lambda = 1$, because by (100) every Taylor coefficient of $h^{\mathrm{CE}}_{1,0}$ odd in $v$ does; $\iota$ is not a symmetry of the leading field, only of this rival boundary. Their $\lambda$-derivatives there are

$$\partial_\lambda c_1(1, 0) = -\frac{4\sqrt{2}}{25}, \qquad \partial_\lambda c_2(1, 0) = -\frac{901\sqrt{2}}{6125}, \tag{104}$$

the first being $\widetilde{C}_{p,p,p}(1)$. The $\alpha$-derivatives are (97) and (98), carried to the exact-CE graph by (92): $\partial_\alpha c_1(1,0) = 2/75$ and $\partial_\alpha c_2(1,0) = 1327/73500$. Hence

$$\det \frac{\partial(c_1,c_2)}{\partial(\lambda,\alpha)}\Big|_{(1,0)} = \begin{vmatrix} -\dfrac{4\sqrt{2}}{25} & \dfrac{2}{75} \\ -\dfrac{901\sqrt{2}}{6125} & \dfrac{1327}{73500} \end{vmatrix} = \frac{19\sqrt{2}}{18375} \neq 0. \tag{105}$$

Every entry is a derivative of the exact-CE coefficients: the first column at $\alpha = 0$, where exact CE coincides with the leading flow, and the second by first-order transfer at $\lambda = 1$. No second-order property of $h^{\mathrm{CE}}$ is used.

**Why one knob is not enough.** Retuning the weight can cancel $c_1$: along $\lambda = 1 + k\sqrt{2}\alpha$ the two coefficients read, to first order,

$$c_1 = \frac{\sqrt{2}(1-12k)}{75}\sqrt{2}\alpha, \qquad c_2 = \frac{\sqrt{2}(1327-21624k)}{147000}\sqrt{2}\alpha,$$

so $k = 1/12$ annihilates $c_1$ and leaves $c_2 = -(19\sqrt{2}/5880)\sqrt{2}\alpha \neq 0$. The determinant (105) is the coordinate-free version of that observation.

**Corollary E.1** (No input weight in a compact set restores exactness at all small scales)**.** *Assume the hypotheses of Theorem 5.1 except for the choice of weight, and let $I \subset (0,\infty)$ be compact. There is $s_0(I) > 0$ such that for every $0 < s < s_0(I)$ and every $\omega \in I$, no rule in $\mathcal{C}$ reproduces the continuous exact-CE endpoint label $S^{\mathrm{CE}}_{\omega,s}$ in the metric $g^J_\omega = \omega g_{ab} + g_c$ almost everywhere under $\mu_J$: the class has a positive error floor, $\inf_{\sigma\in\mathcal{C}} \mathbb{P}[\sigma \neq S^{\mathrm{CE}}_{\omega,s}] \geq \mu_J(U_{s,\omega}) > 0$ for a nonempty open set $U_{s,\omega} \subset \mathcal{M}_1$, at that scale and weight, and the rule may be chosen after both. In particular, no scale-dependent weight $\omega(s)$ taking values in $I$ makes the exact-CE selector sector-local for all sufficiently small positive scales. The threshold $s_0(I)$ depends on $I$; no floor uniform in $s$, and no rate as $s \downarrow 0$, is asserted, and the statement says nothing about weights approaching $0$ or $\infty$.*

*Proof.* Write $J = \{\omega/2 : \omega \in I\}$, a compact subset of $(0,\infty)$, and work with $\lambda \in J$. It suffices to produce $\alpha_0 > 0$ with

$$\big(c_1(\lambda,\alpha), c_2(\lambda,\alpha)\big) \neq (0,0) \qquad \text{for all } \lambda \in J,\ 0 < \alpha < \alpha_0, \tag{106}$$

since the rest of the argument is unchanged. Indeed, a nonzero pair makes the reflected difference a nonzero germ, so triples $(x_0,u_0,v_0)$ with $h^{\mathrm{CE}}_{\lambda,\alpha}(x_0,u_0,v_0) \neq h^{\mathrm{CE}}_{\lambda,\alpha}(x_0,u_0,-v_0)$, which is (101) at these parameters, exist arbitrarily close to the origin; there is no limit to exchange, because (106) is a statement at the actual $(\lambda,\alpha)$. Taking the triple close enough to the origin makes the transverse offsets of the witness and its reflection, at most the reflected difference, small enough for their exits to lie in the two trap-entry neighborhoods. The exit argument, the two collars and the openness of the label sets then give paired open sets $U_{s,\omega}$ and $\iota U_{s,\omega}$ exactly as above, and (102) integrates to the floor $\mu_J(U_{s,\omega}) > 0$. Those steps use only the hyperbolicity of $p_\varepsilon$, the convergence of the two leading unstable branches to the two pure-character maxima, and finite-time comparison; each holds for every $\lambda > 0$, with constants continuous in $\lambda$ and therefore bounded on the compact set $J$.

For (106), the map $(\lambda,\alpha) \mapsto (c_1,c_2)$ is $C^1$ on a neighborhood of $J \times \{0\}$, by the joint regularity of Section E.4. It sends $(1,0)$ to $(0,0)$ and has invertible differential there by (105), so the inverse function theorem gives $\delta > 0$ and a neighborhood $V = [1-\delta, 1+\delta] \times [0,\delta]$ on which the map is injective; hence $(c_1,c_2) \neq (0,0)$ on $V \setminus \{(1,0)\}$, and in particular whenever $|\lambda - 1| \leq \delta$ and $0 < \alpha \leq \delta$. On the complementary compact set $J_\delta = \{\lambda \in J : |\lambda - 1| \geq \delta\}$ the leading coefficient $c_1(\lambda,0) = (\lambda-1)\widetilde{C}_{p,p,p}(\lambda)$ is continuous and nowhere zero, so $|c_1(\lambda,0)| \geq c > 0$ there; by uniform continuity of $c_1$ on $J_\delta \times [0,\delta]$ there is $\alpha_1 \in (0,\delta]$ with $|c_1(\lambda,\alpha)| \geq c/2$ for all $\lambda \in J_\delta$ and $0 \leq \alpha \leq \alpha_1$. Taking $\alpha_0 = \alpha_1$ and $s_0(I)^3 = N^{3/2}\alpha_0$ gives (106).

Finally, a scale-dependent weight $\omega(s) \in I$ is covered because the conclusion holds at every pair $(s,\omega)$ with $0 < s < s_0(I)$ and $\omega \in I$ separately; no continuity or monotonicity of $\omega(\cdot)$ is needed. □

Taking $I = \{\omega\}$ recovers the fixed-weight statement, with its weight-dependent threshold. What the compact hypothesis excludes is a retuning that escapes to $0$ or $\infty$ as the scale falls; nothing here addresses that case, a nonscalar or sector-dependent metric, an exceptional larger scale, or compensation in another model.

### E.9 SCOPE

The hypotheses are used as follows. Three distinct nontrivial real linear characters supply two rivals with equal ceilings and a third sector whose output orientation can be reflected while both rivals' blocks and every sector mass stay fixed; equivalently $G$ has a $C_2 \times C_2$ quotient, so $C_2 \times C_2$, $D_4$ and $Q_8$ qualify, while $C_4$, $C_5$, $C_6$, $C_7$, $S_3$ and $A_4$ do not. Width one makes (80) a single-neuron computation; at larger width $Q_2$ carries cross-neuron terms at the same order, which this argument does not control. Continuous flow is used through the hyperbolic saddle, its stable hypersurface and the exit argument; the retracted iteration displaces the same boundary by $O(s^3 + \delta_s)$, which does not establish that this nonzero coefficient survives a discrete update. The initialization law is $\mu_J$, and the conclusion is a floor at each fixed sufficiently small scale, not an asymptotic floor: it is consistent with, and does not follow from, Theorem 4.2 and Corollary F.2, whose labels agree with the leading ones as $s \downarrow 0$.

Two further boundaries. Nothing above transfers a symmetry, a Riccati equation or a higher coefficient of the truncated chart model to the exact-CE graph; only (91) and (92) cross between them, and both concern the value and one derivative at $\alpha = 0$. And the compact-weight extension in Section E.8 rests on the single determinant (105); a metric family with more than one scalar knob is not covered by it.

# PART II EXTENSIONS

## F EXTENSIONS

No numbered result of the main text depends on this appendix; the main text cites its results only in sentences that name them. Each part strengthens a main result in a direction the narrative names but does not pursue: growing width, the transfer of the leading floor to exact cross entropy, the separately normalized comparator network, the full conditional endpoint law behind the equal label marginals on $C_2^d$, and the global metric threshold on the two-sector square.

### F.1 THE EXPECTED FRACTION OF CHANGED LABELS, UNIFORMLY OVER WIDTH

This states and proves the width-uniform extension that Section 4 mentions in passing. It controls an average over neurons rather than the entire label vector, and the two error quantities are kept apart throughout: *exact-CE versus leading-label error* is the one that vanishes, while the *sector-local prediction error* is the one that keeps a positive floor.

Theorem 4.2 preserves the whole label vector at a fixed width, and a union bound over $M$ neurons cannot be uniform in $M$. A weaker but width-uniform statement is nevertheless available, and it is what the error floor needs in order to survive growing width.

The product-norm field error grows as $\sqrt{M}$, but the perturbation seen by one neuron does not. Direct differentiation of the exact remainder gives

$$\|\mathrm{grad}_m \, \Phi_s - \mathrm{grad}\, f(X_m)\| \le \frac{\sqrt{N+4}}{N} s^3 \qquad (M \ge 1,\ m \le M,\ X \in \mathcal{M}_M). \tag{107}$$

Here $\mathrm{grad}_m$ is the gradient in the $m$th neuron factor. On an energy band around its selected maximum component, the leading ascent of $f(X_m)$ dominates this perturbation regardless of the other neurons. For sufficiently small $s$, the retracted update can neither cross the band nor jump over it, with a threshold independent of $M$.

Fix $G$, $0 < \eta_{\max} < \infty$, and a one-neuron law $\nu$ absolutely continuous with respect to joint volume. Under $\nu^{\otimes M}$,

$$\sup_{\substack{M \ge 1 \\ 0 < \eta \le \eta_{\max}}} \mathbb{E}\left[ \frac{1}{M} \sum_{m=1}^{M} \mathbf{1}_{\{S_{\mathrm{CE},m,s,\eta} \ne S_{J,m}\}} \right] \longrightarrow 0 \qquad (s \downarrow 0), \tag{108}$$

and likewise for continuous exact CE. The labels include $\perp$ for a nonconvergent or tied endpoint, as specified above. Appendix F.2 proves the uniform trapping, convergence and probability estimates. Thus (108) permits any finite width $M = M(s)$ when measuring the expected error fraction. The two statements should be kept apart: the whole-vector union bound is $Mr(s)$, where $r(s) \to 0$, so whole-vector persistence in Theorem 4.2 still fixes $M$.

**Corollary F.1** (Expected neuronwise error uniformly over width)**.** *Fix a group with at least three nontrivial real sectors and $0 < \eta_{\max} < \infty$. For every finite $M$, initialize the neurons independently with normalized joint volume $\mu_J$. For neuron $m$, allow a measurable sector-local ranking $\sigma_m$ to depend additionally on the complete initial states of the other neurons, retaining the information restriction within neuron $m$ and a fixed tie priority. Then*

$$\liminf_{s\downarrow 0} \inf_{\substack{M\geq 1,\ \sigma_1,\dots,\sigma_M \\ 0<\eta\leq\eta_{\max}}} \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} \mathbf{1}_{\{S_{\mathrm{CE},m,s,\eta}\neq\sigma_m\}}\right] \geq \mathrm{err}_J^* > 0,$$

*where $\mathrm{err}_J^*$ is the optimal joint leading-flow error of Appendix C. The same holds for continuous exact CE. For i.i.d. initialization from any fixed law $\nu$ equivalent to $\mu_J$, the bound holds with the law-dependent overlap constant of Remark C.3.*

Conditioning on the other initial neurons leaves a sector-local rule for one neuron, so its leading-label error is at least $\mathrm{err}_J^*$. Subtracting the error in (108) proves the exact-CE floor uniformly over widths, rules, and bounded rates (Appendix C). Finite-scale CE labels need not be independent.

The $O(s^3)$ field and normal-endpoint estimates do not imply an $O(s^3)$ label-error rate. At fixed group and width, the $O_{C^r}(s^3)$ perturbation of the rescaled potential gives an $O(s^3)$ gradient error, and with the Morse–Bott gap the normal endpoint estimate $\mathrm{dist}(X_\infty^{\mathrm{CE}}, K_{\boldsymbol{\rho}}) = O(s^3)$. These estimates do not supply a rate for the label-error probability $r(s) \to 0$, which depends on how initializations approach their attracting components (Remark D.3). Transferring the population floor requires only the qualitative limit. Normal endpoint distance does not control basin-boundary displacement.

### F.2 Width-uniform neuronwise persistence

We strengthen the preceding fixed-width argument by retaining the neuron blocks rather than their aggregate norm. Fix $G$, write $\epsilon = s^3$, $\theta = \epsilon/M$, and stack all sample logits into $Z = \sum_{m=1}^{M} Z_m$. Let $\mathcal{P}$ apply $P$ to every sample and $\mathcal{L}(w) = \sum_{g,h} \psi(w^{(g,h)})$. Then $R_s = \mathcal{L}(\theta\mathcal{P}Z)$ exactly, $\|\mathrm{Hess}\,\mathcal{L}\|_{\mathrm{op}} \leq 1/2$ and $\mathcal{P}\nabla\mathcal{L}(0) = 0$.

**The per-neuron bound.** For $t_{gh} = a_g + b_h$ in one neuron, $\max t_{gh}^2 \leq 2$ and $\sum_{g,h} t_{gh}^2 \leq 2N$, hence $\sum t_{gh}^4 \leq 4N$. Direct differentiation of $Z_m = c_m t^2$ gives

$$\|Z_m\| \leq 2\sqrt{N}, \qquad \|D_m Z\|_F^2 = 8\sum_{g,h} t_{gh}^2 + N\sum_{g,h} t_{gh}^4 \leq 4N(N+4).$$

Consequently $\|\mathcal{P}Z\| \leq 2\sqrt{N}M$ and $\|\mathcal{P}\nabla\mathcal{L}(\theta\mathcal{P}Z)\| \leq \epsilon\sqrt{N}$. The chain rule and tangent projection give

$$\|\mathrm{grad}_m\, R_s\| \leq \frac{2N\sqrt{N+4}}{M}\epsilon^2, \qquad \|\mathrm{grad}_m\, \Phi_s - \mathrm{grad}\, f(X_m)\| \leq \frac{\sqrt{N+4}}{N}\epsilon. \tag{109}$$

These estimates hold at every state of every finite-width manifold, without an independence assumption on the neurons.

**A one-neuron finite-horizon comparison.** In dimensionless leading time the $m$th exact field is $V(X_m) + E_m(X)$, with

$$\|E_m\| \leq e_s := \sqrt{N(N+4)}\,\epsilon.$$

The one-neuron leading field is 5-Lipschitz and bounded by $\sqrt{5}$. Thus (67) holds for the distance between the $m$th exact trajectory and its leading trajectory with $\epsilon_s$ replaced by $e_s$. The discrete recurrence likewise holds with

$$B = \sqrt{5}, \qquad C(s) = \tfrac{1}{2}(\sqrt{5} + e_s)^2 + \tfrac{5}{2}\sqrt{5}.$$

This recurrence uses only the two sphere blocks of that neuron. The forcing may depend on every other neuron at every step, but its displayed bound is uniform. Since $\delta_s \le 2\eta_{\max}\epsilon\sqrt{N}$, the finite-horizon errors vanish uniformly over $M$ and $0 < \eta \le \eta_{\max}$.

**Global convergence with a width-independent step threshold.** To obtain a width-independent step bound, put $D_1 = 2\sqrt{N(N+4)}$. On the product of the closed input and output unit balls, a one-neuron second derivative satisfies $\left\|D^2 Z_m[v,v]\right\| \le D_2 \|v\|^2$, with $D_2 = 8\sqrt{N}$. Indeed, writing $v = (v_{ab}, v_c)$, $w_{gh} = v_{a,g} + v_{b,h}$ gives

$$D^2 Z_m[v,v] = 2c_m w^2 + 4v_c t w.$$

Consequently,

$$\left\|D^2 Z_m[v,v]\right\| \le 4\sqrt{N}\,\|v_{ab}\|^2 + 8\sqrt{N}\,\|v_c\|\,\|v_{ab}\| \le 8\sqrt{N}\,\|v\|^2 .$$

Summing over neurons therefore yields $\|DZ\|_{\mathrm{op}} \le D_1\sqrt{M}$ and $\left\|D^2 Z[v,v]\right\| \le D_2\|v\|^2$. Differentiating $R_s$ twice gives

$$\left\|\nabla^2 R_s\right\|_{\mathrm{op}} \le \frac{\epsilon^2}{M}\left(\tfrac{1}{2}D_1^2 + \sqrt{N}D_2\right) = \frac{2N(N+8)}{M}\epsilon^2.$$

The sphere-curvature term is bounded by the largest ambient sphere-block gradient, which obeys the same bound as the first term of (109). After division by $\lambda_s = 2\epsilon N^2/M$,

$$\|\mathrm{Hess}(\Phi_s - \mathcal{F})\|_{\mathrm{op}} \le \frac{N+8+\sqrt{N+4}}{N}\epsilon.$$

Thus both $\|\mathrm{Hess}\,\Phi_s\|_{\mathrm{op}} \le H_*$ and each sphere-block $\|\mathrm{grad}_i\,\Phi_s\| \le B_*$ have bounds independent of $M$ for $s \le 1$.

For $v = \mathrm{grad}\,\Phi_s(X)$, the normalization-retraction curve $R_X(tv)$ has block speed at most $\|v_i\|$ and covariant acceleration at most $2t\,\|v_i\|^3$. Hence for $0 \le t \le 1$,

$$\left|\frac{\mathrm{d}^2}{\mathrm{d}t^2}\Phi_s(R_X(tv))\right| \le (H_* + 2B_*^2)\,\|v\|^2 .$$

Here $\sum_i \|v_i\|^3 \le B_* \sum_i \|v_i\|^2$ avoids a product-norm factor $\sqrt{M}$. Since $h_s \le 2\eta_{\max}\epsilon N^2$, a single small-scale threshold makes (73) hold for all finite $M$ and bounded rates. The compact analytic convergence argument above therefore applies for every finite $M$, every sufficiently small admissible $s$, every $0 < \eta \le \eta_{\max}$, and every initial state. Its local Lojasiewicz constants need not be uniform in $M$. The continuous analytic flow also converges on each compact $\mathcal{M}_M$.

**One-neuron trapping under the coupled field.** Let $K \subset \mathcal{M}_1$ be a connected leading maximum component of value $f_K$. Choose a tube $U$ inside both the strict label radius and the Morse–Bott neighborhood, and then choose $\Delta > 0$ so that $\sup_{\partial U} f < f_K - 4\Delta$. The closed energy band

$$B_K = \{x \in \overline{U} : f_K - 2\Delta \le f(x) \le f_K - \Delta\}$$

is disjoint from $K$ and from $\partial U$. The normal gradient gap therefore gives $\|\mathrm{grad}\,f\| \ge b_K > 0$ on $B_K$. Along the $m$th coupled exact flow,

$$\frac{\mathrm{d}}{\mathrm{d}\tau}f(X_m) = N^{3/2}\,\|\mathrm{grad}\,f(X_m)\|^2 + \langle \mathrm{grad}\,f(X_m), E_m(X)\rangle > 0 \qquad (X_m \in B_K)$$

when $e_s < N^{3/2}b_K/2$. Thus $A_K = U \cap \{f \ge f_K - 2\Delta\} \Subset U$ traps that neuron once entered.

For a discrete step, Taylor expansion of the *one-neuron* function $f$ along its two retraction blocks has remainder $O_G(\delta_s^2)$, uniform over $M$ and all other states. On $B_K$ the first-order increase is at least $\delta_s N^{3/2} b_K^2/2$, so the increment is positive at small $s$. Globally, the one-neuron step length is at most $\delta_s(\sqrt{5} + e_s)$, and the corresponding change in $f$ is at most $C_G\delta_s$. For sufficiently small $s$, these are smaller than $\mathrm{dist}(A_K, \mathcal{M}_1 \setminus U)$ and $\Delta$, respectively. A step starting above the band cannot jump below it; a step starting in the band increases $f$; neither can jump out of $U$. Hence $A_K$ traps the discrete neuron as well. The argument uses monotonicity of $f(X_m)$ only while the neuron is in the collar band, so it applies to the coupled CE field.

**Entry, normal distance, and uniform probability control.** For each one-neuron initialization $x$ in a leading maximum basin, its leading trajectory enters $\{f > f_K - \Delta\} \cap U$ after finite time. The one-neuron comparison puts the exact neuron in $A_K$ for all sufficiently small $s$, uniformly over width, other initial neurons and bounded rates. For the discrete dynamics use $k_s = \lceil T/\delta_s \rceil$ and the single horizon $[0, T+1]$. At the resulting exact-CE endpoint, (109) and the normal gradient gap give $\operatorname{dist}(X_{\mathrm{CE},m,\infty}, K) = O_G(s^3)$. There are only finitely many one-neuron maximum components, so the constant can be chosen uniformly among them.

For a fixed law $\nu$ absolutely continuous with respect to $\mu_J$ and $\zeta > 0$, choose a compact set of one-neuron leading basin starts with $\nu$-mass at least $1 - \zeta$. A finite cover by neighborhoods with the same entry and trapping margins supplies one $s_\zeta > 0$, independent of $M$, all other initial states and $0 < \eta \le \eta_{\max}$. For $s < s_\zeta$ a neuron's label is correct whenever its initial state belongs to this compact set. Therefore, under $\nu^{\otimes M}$,

$$\sup_{\substack{M \ge 1 \\ 0 < \eta \le \eta_{\max}}} \mathbb{E}\left[\frac{1}{M}\sum_{m=1}^{M} \mathbf{1}_{\{S_{\mathrm{CE},m,s,\eta} \ne S_{J,m}\}}\right] \le \zeta, \tag{110}$$

and likewise for the flow. This proves the averaged limit. Equivalently, there is a deterministic envelope $r(s) \to 0$ that bounds every one-neuron error uniformly in width and learning rate; it follows by applying the compact argument at $\zeta = 2^{-j}$ and taking decreasing thresholds. The compact-set construction uses one threshold on each $K_\zeta$, uniformly over all $0 < \eta \le \eta_{\max}$; it therefore requires neither a trajectory-dependent measurable selection nor independence of the finite-scale CE labels. The whole-vector union bound remains $Mr(s)$, and no convergence rate for $r(s)$ or uniformity in group size is asserted.

### F.3 The floor transfers to exact cross entropy

The obstruction is not confined to the leading flow. Write $\mathcal{C}$ for the class of measurable sector-local rules on the width-one joint manifold $\mathcal{M}_1$, let $S_J$ be the joint leading selector there, and let $\mathrm{err}_J^* = \inf_{\sigma \in \mathcal{C}} \mathbb{P}[\sigma \ne S_J] > 0$ be the optimal joint leading-flow error of Appendix C. All probabilities in the following corollary use $\mu_J$. Let $S_{\mathrm{CE},s,\eta}$ be the total exact-CE endpoint label of Appendix D.8, with its unresolved symbol, and put

$$\bar{r}(s) = \sup_{0 < \eta \le \eta_{\max}} \mathbb{P}\big[S_{\mathrm{CE},s,\eta} \ne S_J\big].$$

**Corollary F.2** (Sector-local error floor for joint exact CE)**.** *Let $G$ have at least three nontrivial real Wedderburn sectors, let $M = 1$, initialize under $\mu_J$, and fix $0 < \eta_{\max} < \infty$. Then*

$$\liminf_{s \downarrow 0} \inf_{\substack{\sigma \in \mathcal{C} \\ 0 < \eta \le \eta_{\max}}} \mathbb{P}\big[S_{\mathrm{CE},s,\eta} \ne \sigma\big] \;\ge\; \mathrm{err}_J^* > 0,$$

*and the same holds for continuous exact CE. In particular there is $s_0 > 0$ such that every $0 < s < s_0$, every $0 < \eta \le \eta_{\max}$ and every $\sigma \in \mathcal{C}$ satisfy $\mathbb{P}[S_{\mathrm{CE},s,\eta} \ne \sigma] \ge \mathrm{err}_J^*/2$. The rule $\sigma$ may be chosen after $s$ and $\eta$. The optimal error itself converges:*

$$\sup_{0 < \eta \le \eta_{\max}} \left| \inf_{\sigma \in \mathcal{C}} \mathbb{P}\big[S_{\mathrm{CE},s,\eta} \ne \sigma\big] - \mathrm{err}_J^* \right| \;\le\; \bar{r}(s) \;\longrightarrow\; 0,$$

*and likewise for continuous exact CE.*

*Proof.* Pointwise, $\sigma \ne S_J$ forces $\sigma \ne S_{\mathrm{CE},s,\eta}$ or $S_{\mathrm{CE},s,\eta} \ne S_J$, so

$$\mathbb{P}[S_{\mathrm{CE},s,\eta} \ne \sigma] \;\ge\; \mathbb{P}[\sigma \ne S_J] - \mathbb{P}[S_{\mathrm{CE},s,\eta} \ne S_J] \;\ge\; \mathrm{err}_J^* - \bar{r}(s)$$

for every $\sigma \in \mathcal{C}$ and every $0 < \eta \le \eta_{\max}$. The second inequality combines the pairing bound (63), applied on $\mathcal{M}_1$ at $\lambda = 1/2$, with the definition of $\bar{r}(s)$. To show $\bar{r}(s) \to 0$, use the compact-set argument from the proof of Theorem 4.2(ii) rather than a pointwise threshold. Given $\zeta > 0$ it supplies a compact $K_\zeta$ with $\mathbb{P}[K_\zeta] \ge 1 - \zeta$ and a single $s_\zeta = s_0(K_\zeta, \eta_{\max}) > 0$ such that $S_{\mathrm{CE},s,\eta} = S_J$ simultaneously for every $X_0 \in K_\zeta$, $0 < s < s_\zeta$, and $0 < \eta \le \eta_{\max}$. Each fixed-$\eta$ error event lies outside $K_\zeta$, so $\bar{r}(s) \le \zeta$ for $s < s_\zeta$ and $\bar{r}(s) \to 0$. Choose $s_0$ with $\bar{r}(s) \le \mathrm{err}_J^*/2$ for $s < s_0$. Part (i) gives the continuous statement. An unresolved exact-CE label only enlarges the mismatch

event. The argument is pointwise in $\sigma$, so it does not interchange an optimization over rules with integration. For the convergence, the same pointwise comparison with the roles of the two labels exchanged gives $\mathbb{P}[S_{\mathrm{CE},s,\eta} \neq \sigma] \leq \mathbb{P}[\sigma \neq S_J] + \bar{r}(s)$ for every $\sigma \in \mathcal{C}$; choosing $\sigma$ within $\zeta$ of $\mathrm{err}_J^*$ and letting $\zeta \downarrow 0$ bounds the infimum by $\mathrm{err}_J^* + \bar{r}(s)$, while the display above bounds it below by $\mathrm{err}_J^* - \bar{r}(s)$. No optimal rule needs to exist. □

*Proof of Corollary F.1.* Condition on the other neurons' initial states. Independence leaves the $m$th initial law unchanged, and the conditional ranking remains sector-local, so $\mathbb{P}[S_{J,m} \neq \sigma_m] \geq \mathrm{err}_J^*$. Appendix F.2 gives $r(s) \to 0$ such that, uniformly in $M$ and $0 < \eta \leq \eta_{\max}$, the probability that neuron $m$'s exact-CE label differs from its leading label is at most $r(s)$. Therefore

$$\mathbb{P}[S_{\mathrm{CE},m,s,\eta} \neq \sigma_m] \geq \mathrm{err}_J^* - r(s)$$

for every allowed predictor. Averaging and taking the infimum over widths, rules and rates proves the claim. Under $\nu^{\otimes M}$, the same argument uses the overlap constant of Remark C.3 and the corresponding one-neuron compact-set argument for $\nu$. The continuous statement is identical. Independence is used only when conditioning on the other initial neurons; no independence of finite-scale CE labels or errors is asserted. □

The infimum includes rules chosen separately at each scale and learning rate. The result supplies no numerical value for $\mathrm{err}_J^*$. Positive-scale exact-CE representability at $\lambda = 1$ fails for three real linear characters at width one under continuous flow (Theorem 5.1) and is open outside that setting.

*Remark* F.3 (Examples and the remaining two-sector case). Groups with a $C_2 \times C_2$ quotient have three nontrivial sign characters, so they are examples of the theorem; this includes $D_4$ and $Q_8$. The theorem also covers $C_6$, with a real sign sector and two complex sectors, and $C_7$, with three complex sectors and no nontrivial real linear character. For groups with exactly two nontrivial real sectors, including $C_4$, $C_5$, $S_3$ and $A_4$, the $\lambda = 1$ clock construction still applies and the metric-reversal theorem still holds. Whether the selector is sector-local at $\lambda \neq 1$ remains unresolved here: preserving both rivals' complete blocks leaves no third nontrivial block for this reflection construction.

*Remark* F.4 (The information restriction is necessary). Theorem 3.1 restricts what a score may read, and that restriction cannot be dropped. On $C_2 \times C_2$, use the orthonormal character coordinates $e_j = \chi_j/2$ and $\gamma = 1/8$, and write $r_j = \gamma\,|a_j b_j c_j|$ and $t_j = \mathrm{sign}(a_j b_j c_j)$ for the three nontrivial characters, so that $f = \sum_j t_j r_j$. The sector masses determine every $r_j$. Two sign vectors $t \neq t'$ give the same value of $\sum_j t_j r_j$ exactly on the hyperplane $\sum_{t_j \neq t'_j} t_j r_j = 0$, and finitely many such hyperplanes are null in $(0,\infty)^3$. This nullness pulls back to initialization: conditional on generic $b, c$, the three nontrivial coordinates of $a \in S^3$ have a density in the unit ball, and taking their absolute values and multiplying by the nonzero $\gamma|b_j c_j|$ preserves absolute continuity. Thus the exact value of $f$ and the masses determine $t$ almost surely. Flipping the three layer signs of one sector with product $+1$ is an isometry of every $g_\lambda$ preserving $f$, hence a symmetry of the flow that fixes labels, and the trivial coordinates do not enter the nontrivial field, so their sign choices do not affect the selected label. The masses together with $(t_1, t_2, t_3)$ therefore determine $S_\lambda$. A predictor granted the masses and the single real number $f$ consequently recovers $S_\lambda$ almost surely for every $\lambda$. Thus one additional scalar can remove the information obstruction in exact arithmetic. The decoding gives no noise-stability guarantee.

### F.4 Exact CE on the separately normalized network

**Corollary F.5** (Exact CE on the separately normalized network)**.** *Theorem 4.2*(i)–(iii) *holds on* $\widetilde{\mathcal{M}}_M = \big((S^{N-1})^3\big)^M$ *with the separate leading flow of Theorem B.1 in place of the joint one and its aligned rank-one components* $\widetilde{K}_{\boldsymbol{\rho}}$ *in place of* $K_{\boldsymbol{\rho}}$*. Consequently, in the setting of Corollary D.1, write* $S^J_{\mathrm{CE},s,\eta}$ *for the exact-CE label of the jointly normalized network from* $X_J$ *and* $S^S_{\mathrm{CE},s,\eta'}$ *for that of the separately normalized network from* $X_S$*. For every* $0 < \eta_{\max} < \infty$*,*

$$\sup_{0<\eta,\eta'\leq\eta_{\max}} \Big|\mathbb{P}\big[S^J_{\mathrm{CE},s,\eta} \neq S^S_{\mathrm{CE},s,\eta'}\big] - q_{\mathrm{op}}\Big| \longrightarrow 0 \qquad (s \downarrow 0), \tag{111}$$

*with the same constant* $q_{\mathrm{op}}(G) > 0$*, and likewise for the two continuous exact-CE flows.*

The proof of Theorem 4.2 carries over to $\widetilde{\mathcal{M}}_M$ with modified constants. The required ingredients are the same.

*The decomposition.* The exact identity (38) is unchanged. Its leading term follows from cancellation of the quadratic logit terms because $h \mapsto gh$ and $g \mapsto gh$ are bijections, independently of the layer normalization. Thus the decomposition holds on $\widetilde{\mathcal{M}}_M$ with the same $\mathcal{F}$ and $\lambda_s = 2s^3N^2/M$. The leading potential of exact CE on the separately normalized network is thus $\sum_m f^{\mathrm{real}}_{\mathrm{sep}}(a_m, b_m, c_m)$, the potential of Theorem B.1, and the metric is the product round metric used there.

*The perturbation.* The map $\Psi(\widetilde{a}, \widetilde{b}, c) = (\widetilde{a}/\sqrt{2}, \widetilde{b}/\sqrt{2}, c)$ is a fixed linear diffeomorphism of $\widetilde{\mathcal{M}}$ onto the balanced subset of $\mathcal{M}_1$, and the scale-free logits $Z_k^{(g,h)} = \sum_m (c_m)_k (a_{m,g} + b_{m,h})^2$ are quadratic in the inputs, so $Z^{\mathrm{sep}} = 2\, Z^{\mathrm{joint}} \circ \Psi$. The two estimates behind (39) therefore hold on $\widetilde{\mathcal{M}}_M$ with their constants multiplied by 4, giving $|R_s| \le 4Ns^6$ and $\|\mathrm{grad}\, R_s\| \le 8Ns^6\sqrt{(N+4)/M}$ there, hence $\Phi_s - \mathcal{F} = O(s^3)$ in value and gradient, and $O(s^3)$ in every finite derivative order on a compact tube. The separate leading field is a polynomial field on a compact manifold, so it is bounded and Lipschitz; the argument therefore changes only the absolute Lipschitz and boundedness constants.

*Morse–Bott.* Let $x$ be a separate aligned intrinsic-rank-one maximum of sector $\rho$, of value $\gamma_\rho > 0$. Since $f^{\mathrm{real}}_{\mathrm{sep}}$ is trilinear in $(Pa, Pb, Pc)$ and the three multipliers all equal $\gamma_\rho$, the constrained Hessian at $x$ is

$$2\big[\Omega(PX, PY, Pc) + \Omega(PX, Pb, PZ) + \Omega(Pa, PY, PZ)\big] - \gamma_\rho\big(\|X\|^2 + \|Y\|^2 + \|Z\|^2\big).$$

Split the tangent space into the nontrivial sectors and the three trivial-coordinate directions. Blockwise multiplicativity of the Fourier transform makes the trilinear bracket vanish unless all three arguments lie in the same sector, while $P\mathbf{1} = 0$ makes it vanish whenever an argument is trivial. Hence the Hessian is block diagonal under this splitting, equal to $-\gamma_\rho$ times the identity on each trivial direction and on each inactive sector. On the active block this is exactly the reduced Hessian (76), now without the input-balance direction, whose kernel is the tangent space of the aligned orbit $(p, q, r) \mapsto (pq^*, rp^*, rq^*)$. Hence $\ker \mathrm{Hess}\, f^{\mathrm{real}}_{\mathrm{sep}}(x) = T_x \widetilde{K}_{\boldsymbol{\rho}}$, and compactness supplies the uniform normal gap, so (69)–(71) hold as before.

*Leading convergence.* Theorem B.1 gives almost-sure convergence to such a maximum under any law absolutely continuous on $\widetilde{\mathcal{M}}$, and the separate field and the product law both separate across neurons, so the same holds at finite width.

*Trapping and convergence.* The remaining arguments of this appendix use only the collar, uniform closeness of $\Phi_s$ to $\mathcal{F}$, monotonicity, sufficient ascent with no overshoot, and the blockwise analytic extension $R(z) = z/\|z\|$. The same arguments apply with three sphere blocks per neuron in place of two. Hence Theorem 4.2(i)–(iii) carries over to the separately normalized network, including the $O(s^3)$ normal-distance bound and uniformity over $0 < \eta \le \eta_{\max}$.

For (111), $X_S$ is uniform on $\widetilde{\mathcal{M}}$ by Corollary B.4, so the separate endpoint result applies. Writing $E_J = \{S^J_{\mathrm{CE},s,\eta} \ne S_{\mathrm{joint}}\}$ and $E_S = \{S^S_{\mathrm{CE},s,\eta'} \ne S^{\mathrm{real}}_{\mathrm{sep}}\}$,

$$\big\{S^J_{\mathrm{CE},s,\eta} \ne S^S_{\mathrm{CE},s,\eta'}\big\} \triangle \big\{S_{\mathrm{joint}} \ne S^{\mathrm{real}}_{\mathrm{sep}}\big\} \subseteq E_J \cup E_S,$$

so the left side of (111) is at most $\mathbb{P}[E_J] + \mathbb{P}[E_S]$. The compact-set argument above, applied once on $\mathcal{M}_1$ and once on $\widetilde{\mathcal{M}}$, sends both probabilities to zero uniformly over the bounded learning-rate interval; $\eta$ and $\eta'$ need not be equal. The continuous statement follows in the same way.

### F.5 The conditional leading endpoint law on a full aligned orbit

Write $\mathcal{A}^S_\rho$ for the full set of separate aligned intrinsic-rank-one positive maxima in sector $\rho$, an entire orbit including every connected component, and $\mathcal{A}^J_\rho = \Psi(\mathcal{A}^S_\rho)$ for its joint counterpart.

**Proposition F.6** (Metric-independent conditional leading endpoint law)**.** *Fix a nontrivial finite group, width one and $\lambda > 0$, and draw $X_0$ from normalized Riemannian volume on $\widetilde{\mathcal{M}} = (S^{N-1})^3$. Let $X^\lambda_\infty$ be its leading $g_\lambda$-endpoint and $S_\lambda$ its sector. For every $\rho$ with $\mathbb{P}[S_\lambda = \rho] > 0$,*

$$\mathcal{L}(X^\lambda_\infty \mid S_\lambda = \rho) = \nu_\rho, \qquad \mathcal{L}(X^\lambda_\infty) = \sum_\rho \mathbb{P}[S_\lambda = \rho]\, \nu_\rho,$$

*where $\nu_\rho$ is the unique Borel probability on $\mathcal{A}_\rho^S$ invariant under the slotwise action of $\mathcal{G}_\rho = \mathrm{O}(n_\rho)^3$, $\mathrm{U}(n_\rho)^3, \mathrm{Sp}(n_\rho)^3$ for $D_\rho = \mathbb{R}, \mathbb{C}, \mathbb{H}$. It does not depend on $\lambda$. For the joint leading flow under $\mu_J$, replace $\nu_\rho$ by $\Psi_*\nu_\rho$ on $\mathcal{A}_\rho^J$. Equal sector marginals therefore give equal leading endpoint laws after this matching.*

On $C_2^d$, $d \geq 2$, the sector marginals are uniform, so the entire leading endpoint laws agree after applying $\Psi$, despite positive paired disagreement. The identification matters because the raw input norms differ: joint leading endpoints have $\|a\| = 1/\sqrt{2}$ and separate endpoints have $\|a\| = 1$. Exact CE need not have the same conditional law, since Theorem 4.2(iii) controls only normal distance to the maximum component.

Fix $\rho$, write $D = D_\rho$ and $n = n_\rho$, and let $\mathcal{G}_\rho = U_D(n)^3$ act by

$$(A, B, C) \longmapsto (PAQ^*, RBP^*, RCQ^*).$$

Here $U_D(n)$ is $\mathrm{O}(n), \mathrm{U}(n)$ or $\mathrm{Sp}(n)$ according to the real form. The action is the identity on every other Fourier sector.

**A product isometry preserving the potential.** Under this action,

$$C^*BA \longmapsto Q(C^*BA)Q^*.$$

Cyclicity of the real trace preserves the block trilinear potential. Left and right unitary multiplication preserve the block Frobenius norms; the weighted Fourier isometry therefore makes the action a product of three orthogonal spatial maps. It preserves every slot norm and sector mass, acts isometrically for every $g_\lambda$ and $g_\omega^J$, fixes the sector labels and preserves normalized volume on each manifold.

**The full aligned set is one orbit.** Suppress the fixed positive norm scales in each slot and start at $A = B = C = E_{11}$. The orbit consists of

$$(pq^*, rp^*, rq^*), \qquad p, q, r \in D^n, \quad \|p\| = \|q\| = \|r\| = 1.$$

Each $U_D(n)$ is transitive on the unit sphere, so this is exactly $\mathcal{A}_\rho^S$, including all its connected components. The vector parametrization has the simultaneous right-unit ambiguity $(p, q, r) \sim (pu, qu, ru)$, $u \in U_D(1)$. The stabilizer in the *unreduced acting group* $U_D(n)^3$ is larger:

$$\{(\mathrm{diag}(u, P_0), \mathrm{diag}(u, Q_0), \mathrm{diag}(u, R_0)) : u \in U_D(1),\ P_0, Q_0, R_0 \in U_D(n-1)\}.$$

Thus it is isomorphic to $U_D(1) \times U_D(n-1)^3$, with $U_D(0)$ trivial. The orbit dimension is $3\beta_D n - \beta_D - 2$, $\beta_D = \dim_{\mathbb{R}} D$, as in Appendix D.4. For $D = \mathbb{R}, n = 1$ the orbit is four points, not one connected component.

**Equivariance determines the conditional law.** An isometry preserving the potential commutes with its gradient flow, hence with the endpoint map wherever it exists. The endpoint map is Borel as a pointwise limit of continuous flow maps on its full-measure domain. Since $\gamma \in \mathcal{G}_\rho$ preserves initialization and the event $\{S_\lambda = \rho\}$, for every Borel $E \subseteq \mathcal{A}_\rho^S$,

$$\mathbb{P}[X_\infty^\lambda \in \gamma E \mid S_\lambda = \rho] = \mathbb{P}[X_\infty^\lambda \in E \mid S_\lambda = \rho].$$

A compact group acting transitively on a compact Hausdorff space has a unique invariant probability: the pushforward of normalized Haar measure by $\gamma \mapsto \gamma x$. Call it $\nu_\rho$. Neither the action nor the orbit depends on $\lambda$. Almost-sure convergence to a pure positive maximum (Theorem B.1) gives the mixture over sectors.

For the full joint flow, the same action preserves $\mu_J$, its metric, potential and selected label. It commutes with $\Psi$, which rescales only the input slots. The joint positive maxima are balanced by Proposition 4.1, so $\mathcal{A}_\rho^J = \Psi(\mathcal{A}_\rho^S)$ is the joint orbit and $\Psi_*\nu_\rho$ its unique invariant measure. This argument uses symmetry of the full joint initialization law; it does not identify a random unbalanced joint start with $\Psi(X_0)$.

At finite width the leading endpoint map separates across neurons, giving the stated product law under i.i.d. initialization. Exact CE is outside this symmetry argument: its remainder reads individual logits, and is not in general invariant under $\mathcal{G}_\rho$.

**Symmetry constraints on sector marginals.** Proposition F.6 determines the law inside each sector; symmetries of the leading problem constrain the weights between sectors. An automorphism of $G$ permutes the nontrivial real sectors and preserves both geometries (Appendix D.8). So does a map that no automorphism need induce: for two nontrivial real linear characters $\chi \neq \chi'$, the orthogonal map of $\mathbb{R}^G$ that swaps $\chi/\sqrt{N}$ and $\chi'/\sqrt{N}$ and fixes their orthogonal complement, $\mathbf{1}$ included, swaps the two one-dimensional terms of (11), which have the same form, and fixes the others. Applied in all three slots it preserves $\Omega$ and $f$, both geometries, the map $T$ and the Gaussian law, and it exchanges the two labels. Exact CE is not invariant under it, since its remainder reads individual logits. Each kind of symmetry forces equal marginal leading weight within its orbits. For $C_2^d, C_4, Q_8, D_4$ the combined orbits on nontrivial real sectors number $1, 2, 2, 2$, leaving $0, 1, 1, 1$ free marginal parameters. For $C_4$, the sign and complex sectors form separate orbits; for $Q_8$, the three sign sectors form one orbit and the quaternionic sector another; for $D_4$, the three sign sectors form one orbit, although automorphisms alone give two, and the two-dimensional sector another. These constraints permit, but do not establish, metric dependence of the remaining weights. Among these examples, only $C_2^d$ has a single orbit of nontrivial sectors, so symmetry fixes all mixture weights; the other examples retain undetermined weights.

### F.6 The global metric threshold on the two-sector square

This extends the two-sector comparison of Appendix B.7, whose reduced system, rectangles and saddle it uses; nothing else depends on it.

**Corollary F.7** (Global metric threshold on the positive two-sector square)**.** *In the aligned chart of Proposition 3.2, put $Q = (0, \pi/2)^2$ and $\vartheta = \theta_* = \phi_* = \arctan(p/q)$. For every $x = (\theta, \phi) \in Q$ with $(\theta - \vartheta)(\phi - \vartheta) < 0$, exactly one $\lambda_\star(x) \in (0, \infty)$ places $x$ on the stable separatrix of the $g_\lambda$-ascent. At that metric its trajectory tends to the mixed saddle; at every other metric it tends to a pure-sector corner. The threshold is smooth on each of the two opposed-coordinate rectangles, with labels*

| | $\lambda < \lambda_\star(x)$ | $\lambda > \lambda_\star(x)$ |
|---|---|---|
| $\theta < \vartheta < \phi$ | $\rho_2$ | $\rho_1$ |
| $\phi < \vartheta < \theta$ | $\rho_1$ | $\rho_2$ |

*If both coordinates are below $\vartheta$, the label is always $\rho_1$; if both are above, it is always $\rho_2$. On either coordinate line through the saddle, the off-saddle coordinate determines the label according to the adjacent aligned rectangle; the saddle itself is fixed for every metric.*

*Locally, uniformly when $\lambda_\star(x)$ stays in a fixed compact subset of $(0, \infty)$,*

$$\lambda_\star(x) = \frac{m_x^2}{2 - m_x} + O(|\theta - \vartheta|), \qquad m_x = \frac{\phi - \vartheta}{\theta - \vartheta}, \tag{112}$$

*as $x$ approaches the saddle through either opposed rectangle.*

*Proof.* Write the time-rescaled field as $(F, \lambda H)$, where

$$F = \sin\theta\cos\theta(q\sin\phi - p\cos\phi), \qquad H = -p\cos^2\theta\sin\phi + q\sin^2\theta\cos\phi.$$

The closed square $\overline{Q} = [0, \pi/2]^2$ is compact and forward invariant, and the calculation of Appendix B.7 shows that its only equilibria are the two pure corners and the mixed saddle. Since $\dot{V} = 2F^2 + \lambda H^2 > 0$ away from these equilibria, the limit set of every trajectory in $\overline{Q}$ is contained in them; being connected, it is one of them, so every trajectory converges to a pure corner or to the saddle.

**Continuation to the edges.** In $L = \{0 < \theta < \vartheta < \phi < \pi/2\}$ one has $F > 0$ and $H < 0$. The upper-left stable branch stays in $L$ backward in time, with strictly decreasing $\theta$ and increasing $\phi$. It cannot reach $\theta = 0$ in finite time, since that edge is invariant for a smooth field. If it never left $\overline{Q}$, both coordinates would have backward limits and that limit would be an equilibrium in $[0, \theta_0] \times [\phi_0, \pi/2]$ for any point $(\theta_0, \phi_0)$ on the branch; there is none. Thus it meets $\phi = \pi/2$ in finite backward time at an angle in $(0, \vartheta)$, transversely since $H < 0$. The transformation $(\theta, \phi, p, q) \mapsto (\pi/2 - \theta, \pi/2 - \phi, q, p)$ gives the other branch, ending on $\phi = 0$ at an angle in $(\vartheta, \pi/2)$. Strict coordinate monotonicity excludes self-intersections; the two branches occupy disjoint rectangles. Their union with the saddle

is a decreasing simple arc joining the horizontal edges. Its complement in $Q$ has two connected components. Every point of the complement reaches a pure sink, and the two open basins are identified by the invariant aligned rectangles.

**Strict metric ordering.** Parameterize the upper-left branch by $\theta = T_\lambda(\phi)$ for $\vartheta < \phi \le \pi/2$ and put $A = F/(-H) > 0$. The graph equation is

$$\partial_\phi T_\lambda = -\lambda^{-1} A(T_\lambda, \phi).$$

The stable manifold theorem with $\lambda$ adjoined, obtained as in the regularity paragraph of Appendix E.4, gives, locally and smoothly in $\lambda > 0$, $T_\lambda(\vartheta + y) = \vartheta + y/m(\lambda) + O(y^2)$. Hence $w = \partial_\lambda T_\lambda > 0$ for small $y > 0$, because $m'(\lambda) < 0$. On every further compact interior segment,

$$w' = \lambda^{-2} A - \lambda^{-1} A_\theta w.$$

The integrating-factor formula keeps $w$ strictly positive. The transverse edge hit and smooth continuation extend the graph to every $\vartheta < \phi \le \pi/2$. Consequently two different metric branches cannot cross away from the saddle.

**The extreme metrics sweep the rectangle.** Let $R = \sqrt{p^2 + q^2}$ and $E = pq/R$. Since $H_\theta > 0$, $H(\vartheta, \vartheta + y) = -E \sin y$, and $F = R \sin\theta \cos\theta \sin y$, throughout $L$

$$0 < A \le \frac{R}{2E}, \qquad 0 < \vartheta - T_\lambda(\vartheta + y) \le \frac{Ry}{2E\lambda}.$$

Thus $T_\lambda(\phi) \to \vartheta$ as $\lambda \to \infty$. For the opposite limit fix $d \in (0, \vartheta)$ and put $c_d = \min_{d \le t \le \vartheta} \sin t \cos t > 0$. If $T_\lambda(\vartheta + y) \ge d$, the whole preceding graph segment has $\theta \in [d, \vartheta]$, and $-H \le p + q$ gives

$$\vartheta - d \;\ge\; \vartheta - T_\lambda(\vartheta + y) \;\ge\; \frac{Rc_d}{\lambda(p+q)}(1 - \cos y).$$

For fixed $y > 0$ this is impossible when $\lambda$ is small enough. Hence $T_\lambda(\phi) \to 0$ as $\lambda \downarrow 0$. For each fixed $\phi > \vartheta$, the map $\lambda \mapsto T_\lambda(\phi)$ is therefore a smooth increasing bijection from $(0, \infty)$ onto $(0, \vartheta)$, with nonzero derivative. The implicit function theorem gives the unique smooth threshold throughout $L$. Points to the right of this graph have label $\rho_2$, those to the left $\rho_1$, proving the first row of the table. The stated reflection exchanges $p, q$ and the two labels, proving the second row for all unequal ceilings too. The coordinate lines enter their adjacent aligned rectangles immediately. The opposite label orientations show that the two entire basins are not nested as $\lambda$ varies.

Finally the local stable graph satisfies $\phi - \vartheta = m(\lambda)(\theta - \vartheta) + O((\theta - \vartheta)^2)$ uniformly on compact positive metric intervals. Dividing by $\theta - \vartheta$ and inverting $m$ gives (112). The compact-interval qualification is essential; that formula is not a global expression for the threshold. □

# PART III COMPARISONS AND CORRECTIONS

## G COMPARISONS AND CORRECTIONS

This appendix compares the results with directly related prior statements and records one technical correction to a previously published claim. No proof in Appendices A–E and no statement of the main text depends on it. The intrinsic $D_\rho$-rank conventions those proofs use are established in Appendix A. What is here is the comparison of the feasible and unrestricted stability classifications, which qualifies one clause of He et al. (2026b).

### G.1 THE QUATERNIONIC REAL FORM: FEASIBILITY AND STABILITY

For $X \in M_n(D)$ write $\operatorname{rank}_D X$ for matrix rank over $D$. Intrinsic rank one has complex ranks $1, 1, 2$ for $D = \mathbb{R}, \mathbb{C}, \mathbb{H}$ and realified ranks $1, 2, 4$. In the minimal quaternionic case,

$$q = z + w\mathbf{j} \longmapsto \begin{pmatrix} z & w \\ -\overline{w} & \overline{z} \end{pmatrix}, \qquad q^* q = (|z|^2 + |w|^2) I_2, \tag{113}$$

so every nonzero feasible coefficient has two equal complex singular values. The same embedding gives $\operatorname{rank}_{\mathbb{C}} X = 2\operatorname{rank}_{\mathbb{H}} X$ for every $X \in M_n(\mathbb{H})$: a nonzero feasible quaternionic coefficient of any size has *even* complex rank, and (113) is that statement at $n = 1$. Intrinsic $D_\rho$-rank one, not ordinary complex rank one, is therefore the feasible conclusion in a quaternionic sector.

An unrestricted-complex perturbation that raises one singular value and lowers the other leaves the real form. With

$$S = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \qquad Q = \begin{pmatrix} z & w \\ -\overline{w} & \overline{z} \end{pmatrix} \in M_1(\mathbb{H}),$$

$$\operatorname{Re}\operatorname{tr}(S^* Q) = \operatorname{Re}(z - \overline{z}) = 0, \tag{114}$$

so $S$ is orthogonal to the entire quaternionic real form and its projection onto the feasible sphere tangent vanishes. In the enlarged complex space the same direction can increase the objective, and the following two statements separate the two effects: the enlargement does not move a feasible trajectory, but it changes the stability classification of a feasible point.

**Lemma G.1** (The real form is flow-invariant)**.** *Let $\mathcal{A} = M_n(D) \subset M_d(\mathbb{C})$ be a real Wedderburn form, and give $\mathcal{E}(A, B, C) = \operatorname{Re}\operatorname{tr}(C^* BA)$ the gradient flow of the enlarged product of unit Frobenius spheres $S(M_d(\mathbb{C}))^3$. Then $S(\mathcal{A})^3$ is an invariant submanifold of that flow.*

*Proof.* $\mathcal{A}$ is an $\mathbb{R}$-subalgebra of $M_d(\mathbb{C})$ closed under the conjugate transpose. The Euclidean gradients of $\mathcal{E}$ in the real inner product $\langle U, V\rangle = \operatorname{Re}\operatorname{tr}(U^* V)$ are

$$\nabla_A \mathcal{E} = B^* C, \qquad \nabla_B \mathcal{E} = CA^*, \qquad \nabla_C \mathcal{E} = BA, \tag{115}$$

each a product of two elements of $\mathcal{A}$ or of their adjoints, hence in $\mathcal{A}$. The Riemannian gradient on the sphere subtracts $\langle \nabla_A \mathcal{E}, A\rangle\, A$, a *real* multiple of a point of $\mathcal{A}$, so it lies in $\mathcal{A}$ as well; likewise in the other two slots. The field is tangent to $S(\mathcal{A})^3$ at each of its points, so that set is invariant. □

**Proposition G.2** (Opposite stability classifications at one feasible point)**.** *Take $A = B = C = I_2/\sqrt{2}$ in the minimal quaternionic form $M_1(\mathbb{H}) \subset M_2(\mathbb{C})$. This is a critical point of $\mathcal{E}$ on three unit Frobenius spheres in both spaces, and the constrained Hessian there has inertia $(n_+, n_-, n_0)$ equal to*

$$(0, 3, 6) \ \text{on} \ S(M_1(\mathbb{H}))^3, \qquad (3, 10, 8) \ \text{on} \ S(M_2(\mathbb{C}))^3. \tag{116}$$

*The point is a Morse–Bott local maximum of the feasible problem, on a six-dimensional critical manifold, and a strict saddle of the enlarged one.*

*Proof.* Criticality holds in both spaces because $\nabla_A \mathcal{E} = B^* C = I_2/2$ is the real multiple $2^{-1/2} A$ of the point, so the sphere projection of the gradient vanishes; the same holds in the other two slots by symmetry. The common value is $\mathcal{E} = 1/\sqrt{2}$ and the multiplier in each slot is $\langle \nabla_A \mathcal{E}, A\rangle = 1/\sqrt{2}$.

*Feasible side.* Identify $M_1(\mathbb{H})$ with $\mathbb{H}$ by (113). Then $\|A\|_F = \sqrt{2}\,|a|$ and $\mathcal{E} = 2\operatorname{Re}(\bar{c}ba)$, so on the three spheres $|a| = |b| = |c| = 2^{-1/2}$ and

$$\mathcal{E} = 2\operatorname{Re}(\bar{c}ba) \le 2\,|a|\,|b|\,|c| = \frac{1}{\sqrt{2}}, \tag{117}$$

with equality exactly when $\bar{c}ba$ is a nonnegative real, that is when $c = \sqrt{2}\,ba$. The maximizer set is the graph $\{(a, b, \sqrt{2}\,ba)\}$ over two copies of the radius-$2^{-1/2}$ sphere in $\mathbb{H}$, a six-dimensional manifold through the point on which $\mathcal{E}$ is constant at its maximum. Hence $n_+ = 0$ and $n_0 \ge 6$.

*Enlarged side.* Put $\hat{q} = I_2/\sqrt{2}$ and $\hat{S} = S/\|S\|_F$, and move all three slots along the same great circle $A_\varepsilon = B_\varepsilon = C_\varepsilon = \cos\varepsilon\,\hat{q} + \sin\varepsilon\,\hat{S}$. Then

$$\mathcal{E}(\varepsilon) = \frac{\cos\varepsilon\,(2 - \cos 2\varepsilon)}{\sqrt{2}} = \frac{1}{\sqrt{2}}\left(1 + \frac{3}{2}\varepsilon^2\right) + O(\varepsilon^4), \tag{118}$$

so the curvature is $3/\sqrt{2} > 0$ along a direction of squared length 3, and the point is a strict saddle of the enlarged problem. By (114) that direction is orthogonal to the feasible form, so it is invisible to the feasible Hessian.

*Both inertias from one block.* At $\hat{q}$ the sphere tangent of $M_2(\mathbb{C})$ splits orthogonally as

$$T_{\hat{q}}S(M_2(\mathbb{C})) = \mathrm{Herm}_0(2) \oplus \mathbb{R}\, i\hat{q} \oplus i\,\mathrm{Herm}_0(2), \tag{119}$$

three Hermitian traceless directions and four skew-Hermitian ones. Fix a unit $H$ among these seven and vary all three slots along it, $\dot{A} = \alpha H$, $\dot{B} = \beta H$, $\dot{C} = \gamma H$. With the multiplier $1/\sqrt{2}$ in each slot the constrained Hessian is

$$\sqrt{2}\big(t\,\alpha\beta + \alpha\gamma + \beta\gamma\big) - \tfrac{1}{\sqrt{2}}\big(\alpha^2 + \beta^2 + \gamma^2\big), \qquad t = \mathrm{Re}\,\mathrm{tr}(H^2), \tag{120}$$

and $t = +1$ for Hermitian $H$, $t = -1$ for skew-Hermitian $H$. Distinct directions of (119) have both $\langle H_i, H_j\rangle = 0$ and $\mathrm{Re}\,\mathrm{tr}(H_i H_j) = 0$, the latter because the product of a Hermitian and a skew-Hermitian matrix has purely imaginary trace. The 21-dimensional Hessian is therefore block diagonal with seven blocks

$$\frac{1}{\sqrt{2}}\begin{pmatrix} -1 & t & 1 \\ t & -1 & 1 \\ 1 & 1 & -1 \end{pmatrix}. \tag{121}$$

At $t = 1$ this is $(J - 2I_3)/\sqrt{2}$ with $J$ the all-ones matrix, of spectrum $\{1/\sqrt{2}, -\sqrt{2}, -\sqrt{2}\}$; at $t = -1$ it is $-ww^{\mathsf{T}}/\sqrt{2}$ with $w = (1, 1, -1)^{\mathsf{T}}$, of spectrum $\{-3/\sqrt{2}, 0, 0\}$. Three Hermitian and four skew directions give $(3, 10, 8)$, and the single positive eigenvalue $1/\sqrt{2}$ is the curvature $3/\sqrt{2}$ of (118) divided by the squared length 3 of that direction. The feasible tangent $T_{\hat{q}}S(M_1(\mathbb{H}))^3$ is exactly the three traceless skew-Hermitian directions in each slot, so the feasible Hessian consists of three copies of the $t = -1$ block and gives $(0, 3, 6)$: the transverse Hessian is negative definite with the single eigenvalue $-3/\sqrt{2}$, so the maximum (117) is Morse–Bott nondegenerate on its six-dimensional critical manifold. □

# PART IV REPRODUCIBILITY

## H EXPERIMENTAL PROTOCOLS

### H.1 DISAGREEMENT-RATE TEST: PROTOCOL AND RESULTS

The sampling design was prespecified. Because the same three groups were used during theory development, the experiment estimates the disagreement rate rather than generalization to held-out groups. Width one and float64 throughout. The fixed settings were:

- $n = 20{,}000$ paired draws per group, seed stream 20260902; batched RK4 with a retraction at every stage, step $0.01$ in $\tau$;
- resolution rule: integrate to $\tau = 60$, continue any draw whose relative sector-score gap is below $0.999$ in blocks of 60 up to $\tau = 180$, then mark the draw unresolved and use unresolved draws to bracket each reported rate;
- Wilson $95\%$ intervals on every rate; total variations are reported as point estimates;
- exact-CE ladder: 1000 further draws, seed stream 20260903, $s \in \{0.50, 0.40, 0.30, 0.20\}$, $\eta = 0.25$, the same draws at every scale.

The protocol also prespecified bootstrap intervals for the total variations, not reported here, and Stuart–Maxwell tests of marginal homogeneity; none rejected at the $5\%$ level (minimum $p = 0.176$).

For the exact-CE ladder, continuous CE is integrated from the coupled joint initialization to the fixed horizon $\tau = 60$ by fixed-step RK4 with step $0.01$ and joint-sphere retraction at every RK4 stage; it has no adaptive tolerance, convergence stop, or extension to $\tau = 180$. The retracted arm takes exactly

$$K_s = \mathrm{round}(60/\delta_s), \qquad \delta_s = 2\eta s^3\sqrt{N}/M,$$

updates and does not stop early. Here $N = 8$, $M = 1$, and the four update counts are 339, 663, 1571, and 5303 in decreasing scale order. Thus the discrete horizon is $K_s\delta_s$, with $|K_s\delta_s - 60| \le \delta_s/2$, rather than exactly 60: the four realized horizons are 59.9273, 60.0079, 59.9867, and 59.9966 in the same order.

The continuous arm reaches $\tau = 60$ exactly. At the terminal state, the label is the nontrivial canonical-score maximizer only when its relative gap $(S_{(1)} - S_{(2)})/S_{(1)}$ is at least 0.999; otherwise it is the unresolved label $\perp$. No ladder endpoint was unresolved.

The settling index is descriptive only. Labels are sampled after every $\max(1, \lfloor K_s/240 \rfloor)$ updates and at the final update; the stored value is the zero-based index of the last observed label change and does not affect stopping. The empirical tables report resolved labels at nominal horizon $\tau \approx 60$; they do not estimate the infinite-time endpoint directly. The asymptotic disagreement probability $q_{\mathrm{op}}$ is defined in Corollary D.1. Table 2 records the rates behind Figure 2(a); multiplying by 20,000 gives the disagreement counts.

Table 2: Width-one Monte Carlo estimates at $n = 20{,}000$ per group, with Wilson 95% intervals rounded outward. $q_{\mathrm{op}}$ is the paired disagreement probability under the shared Gaussian coupling; every trajectory was numerically resolved. $q_{\mathrm{met}}$ is the disagreement rate when only the metric changes.

| group | $N$ | $q_{\mathrm{op}}$ | $q_{\mathrm{met}}$ | unresolved |
|---|---|---|---|---|
| $C_2 \times C_2$ | 4 | 0.1843 [0.1789, 0.1898] | 0.1131 [0.1087, 0.1176] | 0 |
| $D_4$ | 8 | 0.14465 [0.1398, 0.1496] | 0.1086 [0.1043, 0.1130] | 0 |
| $Q_8$ | 8 | 0.15845 [0.1534, 0.1636] | 0.12035 [0.1159, 0.1250] | 0 |

The following five numerical checks were prespecified:

1. On 64 random points per group, the analytic ambient partials of the centered potential were required to agree with coordinatewise central finite differences to relative error below $10^{-7}$; the largest observed error was $2.04 \times 10^{-10}$. Separately, the projected joint, separate, and metric-only fields were required to be tangent to their sphere constraints to $10^{-12}$; the largest normal component was $6.3 \times 10^{-17}$.
2. At $\theta_0 = 0.70$ and the midpoint of each group's own joint and separate boundaries, which is 0.925803534 on all three groups at the $5 \times 10^{-10}$ resolution of the root solve, the joint and separate labels, and the separate and metric-only labels, were required to disagree.
3. At the two outside-wedge points $\phi = 0.64906212$ and $1.20254495$, all three flows were required to agree with the named pure character.
4. On 256 draws per group and all three flows, DOP853 with $\mathrm{rtol} = 10^{-11}$ and $\mathrm{atol} = 10^{-13}$ was required to resolve every draw and match the RK4 label; all 2304 comparisons agreed.
5. Every resolved endpoint was required to lie within $10^{-6}$ of its sector ceiling. The maximum over all three groups and all three flows was $4.38 \times 10^{-11}$, attained by the $Q_8$ joint arm; the maximum for the $Q_8$ separate arm was $5.17 \times 10^{-15}$.

We also checked the stable-slope formula directly against the spatial field for $\lambda = 0.25$, 0.5, 0.75, 1, 1.5, 2, and 4 on all three groups. The required relative residual was below $10^{-12}$; the maximum was $6.66 \times 10^{-16}$.

A second implementation uses $n = 1000$ draws per group and the same group tables. It constructs the sector projectors by SVD of each orbit's matrix coefficients instead of reading central idempotents off the characters, computes the output gradient by scattering over the multiplication table, and integrates with DOP853 at $\mathrm{rtol} = 10^{-10}$ and $\mathrm{atol} = 10^{-12}$. Its draws reuse the primary seed stream with a different batch size, so its initialization triples are neither paired with nor independent of the primary sample.

No draw of the second implementation was unresolved. Across the nine group/comparison rates, the largest absolute difference from the primary $n = 20{,}000$ estimate was 0.0129. Using $\sqrt{p(1-p)/1000}$, with $p$ the primary rate, as a descriptive scale, the largest standardized difference

was 1.36. This is not an independent-sample test. Per-draw labels were not retained, so only aggregate rates can be compared.

Two summaries were added after the run: the 0.01 descriptive tolerance for marginal agreement and the conditional $\mathbb{P}[S_{\mathrm{CE},s} = S_{\mathrm{joint}} \mid D_{\mathrm{op}}]$.

Table 3 records that conditional at every scale as a count out of the 148 contested draws, together with its Wilson interval. At the three larger scales, every contested draw that does not follow the joint selector follows the separate selector except for one draw in a third sector. By $s = 0.20$, that draw carries the joint label. Across all 1000 draws, the continuous and retracted labels agreed on 996, 997, 999, and 999 draws at $s = 0.50$, 0.40, 0.30, and 0.20, respectively. The prespecified agreement criterion required the continuous and retracted labels to agree on at least 99% of the 1000 draws at every ladder scale.

The all-draw comparison has one exception at the smallest ladder scale. Counted over all 1000 draws rather than the contested 148, continuous exact CE at $s = 0.20$ carries the joint leading label on 999 of 1000 draws; the retracted arm carries it on 1000. The single exception is one draw whose continuous label is the two-dimensional sector, which is neither leading label. The exceptional label persists under numerical refinement: adaptive reruns retain it at normalized times 60 and 240 using both the primary tolerances $(\mathrm{rtol}, \mathrm{atol}) = (10^{-11}, 10^{-13})$ and the refined tolerances $(10^{-12}, 10^{-14})$. At the smaller scales $s = 0.10$ and $s = 0.05$, both tolerance settings return the joint leading label at normalized times 60 and 240; the leading flow has that same label. These follow-up runs were added after the original scale sweep. They are consistent with a finite-scale effect, but their finite integration times do not establish endpoint agreement.

Table 3: Exact-CE inheritance on the 148 of 1000 $D_4$ draws whose joint and separate leading labels differ, a conditional summary chosen after the run, with Wilson 95% intervals rounded outward. "Follows $J$" ("follows $S$") counts draws whose resolved exact-CE label at nominal $\tau \approx 60$ is the joint (separate) leading label.

| $s$ | mode | follows $J$ | follows $S$ | neither | conditional |
|---|---|---|---|---|---|
| 0.50 | continuous | 143/148 | 4 | 1 | 0.9662 [0.9233, 0.9855] |
| 0.50 | retracted | 142/148 | 5 | 1 | 0.9595 [0.9143, 0.9813] |
| 0.40 | continuous | 144/148 | 3 | 1 | 0.9730 [0.9325, 0.9895] |
| 0.40 | retracted | 144/148 | 3 | 1 | 0.9730 [0.9325, 0.9895] |
| 0.30 | continuous | 147/148 | 0 | 1 | 0.9932 [0.9627, 0.9989] |
| 0.30 | retracted | 147/148 | 0 | 1 | 0.9932 [0.9627, 0.9989] |
| 0.20 | continuous | 148/148 | 0 | 0 | 1.0000 [0.9747, 1.0000] |
| 0.20 | retracted | 148/148 | 0 | 0 | 1.0000 [0.9747, 1.0000] |

### H.2 A REFLECTED-WITNESS LOWER BOUND ON THE OPTIMAL SECTOR-LOCAL ERROR

Theorem 3.1 and Corollary C.4 assert that the optimal sector-local error is positive and give no value; their proof bounds it below by the volume of a constructed neighborhood, which the proof does not quantify. Remark C.1 removes the construction: for a designated nontrivial sector $j$ with reflection $\iota_j(a, b, c) = (a, b, c - 2\Pi_j c)$ the same pairing argument applies to all of

$$W_j = \{X : S_J(X) \neq S_J(\iota_j X),\ S_J(X) \neq j,\ S_J(\iota_j X) \neq j\}, \qquad \mathrm{err}_J^* \geq \tfrac{1}{2}\mu_J(W_j),$$

and $W_j$ is defined by endpoint labels alone. This subsection estimates $\mu_J(W_j)$.

**What is and is not being estimated.** This is a Monte Carlo estimate of a measure that mathematically lower-bounds $\mathrm{err}_J^*$. It is not an estimate of $\mathrm{err}_J^*$, and a point estimate is not a proved bound; the interval below is a binomial confidence statement about $\mu_J(W_j)$, and the displayed floor column is half its endpoint. The factor of one half is not optional: each reflected pair contains two states and the contradiction forces an error at one of them, not at both. The confidence statement is conditional on the integrator's resolved labels, which are numerical, not certified endpoint labels.

**Protocol.** The analysis plan was written before this statistic was computed. It records a 400-draw pilot, used only to set the per-group sample sizes and discarded with its seed, and one amendment of

those sizes; the final sample sizes are those reported in Table 4. The timing of the plan is self-reported rather than independently verified. Draws are $\mu_J$ on $\mathcal{M}_1$ at width one under the ordinary joint metric $g_1^J$, a mismatched weight the negative direction covers. The integrator constants are those of Appendix H.1. Every nontrivial sector is used as $j$ and every one is reported, so no reflection is selected after the fact; the confidence bounds are one-sided Clopper–Pearson at level $0.05/k$ with $k$ the number of nontrivial sectors, which makes them simultaneous over $j$ and the largest of them valid at $95\%$. A draw counts as a witness only when both flows resolve and the event holds, while the denominator stays the full sample, so the estimate is conservative for $\mu_J(W_j)$. The reflection was checked numerically to preserve every sector mass, both input blocks and the output norm to machine precision on every draw.

Table 4: Reflected-witness measure under $\mu_J$ at width one, $g_1^J$. A witness for sector $j$ is a draw whose leading label changes when sector $j$'s output block is reflected, with neither label equal to $j$. $\widehat{\mu}$ is the witness fraction over the full sample; $L_j$ is a one-sided Clopper–Pearson lower bound on $\mu_J(W_j)$ at level $0.05/k$, simultaneous over the $k$ reflections within a group; the last column is the resulting lower bound on $\mathrm{err}_J^*$, simultaneous within the group and conditional on the numerical labels. Every draw resolved.

| group | reflected sector $j$ | $n$ | witnesses | $\widehat{\mu}(W_j)$ | $L_j$ | $\mathrm{err}_J^* \geq L_j/2$ |
|---|---|---|---|---|---|---|
| $C_2 \times C_2$ | $\chi_{+-}$ | 6000 | 2 | 0.00033 | 0.000032 | 0.000016 |
| | $\chi_{-+}$ | 6000 | 0 | 0.00000 | 0.000000 | 0.000000 |
| | $\chi_{--}$ | 6000 | 3 | 0.00050 | 0.000088 | **0.000044** |
| $D_4$ | $\chi_{+-}$ | 2000 | 2 | 0.00100 | 0.000083 | 0.000041 |
| | $\chi_{-+}$ | 2000 | 2 | 0.00100 | 0.000083 | 0.000041 |
| | $\chi_{--}$ | 2000 | 1 | 0.00050 | 0.000006 | 0.000003 |
| | $\rho_2$ (real) | 2000 | 3 | 0.00150 | 0.000237 | **0.000118** |
| $Q_8$ | $\chi_{+-}$ | 2000 | 4 | 0.00200 | 0.000440 | 0.000220 |
| | $\chi_{-+}$ | 2000 | 0 | 0.00000 | 0.000000 | 0.000000 |
| | $\chi_{--}$ | 2000 | 1 | 0.00050 | 0.000006 | 0.000003 |
| | $\rho_2$ (quaternionic) | 2000 | 7 | 0.00350 | 0.001219 | **0.000609** |

**Reading the table.** The event is rare: between zero and 3.5 draws in a thousand, against paired disagreement rates $q_{\mathrm{op}}$ near $0.15$, which bound $\mathrm{err}_J^*$ from above (Remark C.2). That is expected, since a witness needs the reflected sector to lose at both states while still tipping the contest between the other two. The best simultaneous bounds are $\mathrm{err}_J^* \geq 6.09 \times 10^{-4}$ on $Q_8$, $1.18 \times 10^{-4}$ on $D_4$ and $4.40 \times 10^{-5}$ on $C_2 \times C_2$, each at $95\%$ confidence simultaneous over that group's reflections and conditional on the numerically resolved labels; the corresponding point estimates of $\frac{1}{2}\mu_J(W_j)$ are $1.75 \times 10^{-3}$, $7.5 \times 10^{-4}$ and $2.5 \times 10^{-4}$. These lower bounds are small; they neither estimate $\mathrm{err}_J^*$ nor show that it is small. Each bound holds for each rule separately, the set of starts on which a rule errs depends on the rule, and no single start is misclassified by every rule. The zero counts give vacuous bounds at these sample sizes and are reported as they came out.

On both groups of order eight the largest measured witness fraction comes from reflecting the two-dimensional sector rather than a linear character. Nothing here explains that, and with these counts it is not established.

### H.3 A RAW-BLOCK RULE, AND THE OPTIMAL ERROR ON $C_2 \times C_2$

**Freezing the scalar coupling.** Fix $\omega > 0$ and let $X = (a, b, c)$ follow the leading $g_\omega^J$ ascent on $\mathcal{M}_1$, $\dot{a} = (G_a - 2fa)/\omega$, $\dot{b} = (G_b - 2fb)/\omega$, $\dot{c} = G_c - fc$, with $G$ the Euclidean partial gradients of $f$. Put $H(t) = \exp \int_0^t f(X)$, $A = H^{2/\omega}a$, $B = H^{2/\omega}b$, $C = Hc$ and $\mathrm{d}\xi/\,\mathrm{d}t = 1/(\omega H)$. Since the partial gradients are bilinear and $\|(A, B)\| = H^{2/\omega}$,

$$A' = \nabla_A f, \qquad B' = \nabla_B f, \qquad C' = u\,\nabla_C f, \qquad u = \omega\,\|(A, B)\|^{\omega-2}. \tag{122}$$

The gradient of $f$ is block diagonal over the real sectors, so the lifted sectors interact only through the scalar $u$, which reads the total input radius; at $\omega = 2$ it is the constant 2, and (122) is the clock system of Corollary C.4 up to a time scale. Replacing $u$ by its initial value $\omega$ leaves independent polynomial

systems, sector $j$ started from its raw initial blocks $(\Pi_j a, \Pi_j b, \Pi_j c)$, with maximal lifetimes $T_j$. The scores $(1+T_j)^{-1}$ read only sector $j$'s blocks and are Borel as in Appendix C, so at $\omega = 1$ the resulting rule $\sigma_{\mathrm{raw}}$ is sector-local and $\mathrm{err}_J^* \le \mathbb{P}_{\mu_J}[\sigma_{\mathrm{raw}} \neq S_J]$. Nothing here bounds how far $u$ drifts before the winner is decided; the rule's error is measured, not proved small.

On $C_2 \times C_2$ every sector is one real character, and its lifted system is, up to a time scale common to all sectors, $x' = yz$, $y' = xz$, $z' = xy$ from the sector's three coefficients. The three squares share one derivative, so with $U, V, W$ the squared initial coefficients the lifetime is $T = R_F(U, V, W)$ when $x_0 y_0 z_0 > 0$, with $R_F$ Carlson's symmetric elliptic integral. When $x_0 y_0 z_0 < 0$ and the ordered squares $m_1 \le m_2 \le m_3$ have $m_1 < m_2$, the smallest coefficient passes through zero first and $T = 2R_F(0, m_2 - m_1, m_3 - m_1) - R_F(U, V, W)$. On the remaining null set $T$ is still the maximal lifetime of the system, which these formulas need not give: with one zero coefficient the state enters $xyz > 0$ at once and $T = R_F(U, V, W)$; with two it is at rest, as at $(0, 0, 1)$, where the second formula reads $\infty - \infty$, and $T = +\infty$; and with $x_0 y_0 z_0 < 0$ and $m_1 = m_2$ the two smallest coefficients only tend to zero, and again $T = +\infty$. At $\omega = 2$ the ratio is exactly $u = 2$, and the exact clocks of Theorem 3.1 follow the same way: in the system $a_j' = b_j c_j/16$, $b_j' = a_j c_j/16$, $c_j' = a_j b_j/8$ of Section 3 the quantities $a_j^2$, $b_j^2$ and $c_j^2/2$ share one derivative, and with them as $U, V, W$ the lifetime is $8\sqrt{2}$ times the lifetime above in every case.

**Proposition H.1** (On $C_2 \times C_2$ the optimum is a finite optimization on each mass fiber)**.** *On $C_2 \times C_2$ at width one, under $\mu_J$ and ordinary joint normalization, let $\mathfrak{m}$ be the mass vector and $t = (t_1, t_2, t_3)$ the product signs $t_j = \mathrm{sign}(a_j b_j c_j)$ in the coordinates of Remark F.4. Then $S_J = S(\mathfrak{m}, t)$ almost everywhere, and*

$$\mathrm{err}_J^* = \mathbb{E}_{\mathfrak{m}}\big[r^*(\mathfrak{m})\big], \qquad r^*(\mathfrak{m}) = \min_{\pi} \frac{1}{8} \sum_{t \in \{\pm 1\}^3} \mathbf{1}\{p_\pi(t) \neq S(\mathfrak{m}, t)\},$$

*where $\pi$ runs over the 720 total orders of the six slots $(j, \pm)$ and $p_\pi(t)$ is the sector of the highest slot $(j, t_j)$.*

*Proof.* The sign flips of Remark F.4 that keep each product sign are coordinate reflections, hence isometries of $g_1^J$ as well as of $g_\lambda$, and they preserve $f$; the trivial coordinates do not enter the nontrivial field. So the argument given there for $S_\lambda$ shows $S_J = S(\mathfrak{m}, t)$ almost everywhere. Conditionally on $\mathfrak{m}$ the nine nontrivial coordinate signs are independent and uniform. Write $g_j$ for the signs of $a_j$ and $b_j$; with $t_j$ they determine sector $j$'s block. Conditionally on $\mathfrak{m}$ and all $g_j$, a sector-local rule gives sector $j$ one of two scores according to $t_j$, six numbers in all, and a total order of these six slots, with ties refined by the fixed priority, reproduces its decision at each of the eight sign states, since each state presents one slot per sector. Its conditional error is therefore at least $r^*(\mathfrak{m})$, and averaging over the gauges keeps the bound. Conversely, each of the 720 conditional losses is a measurable function of $\mathfrak{m}$, so the first minimizing order in a fixed enumeration is too, and ranking the slots by it is a sector-local rule, reading only the masses and its own product sign, that attains $r^*(\mathfrak{m})$ on every fiber. □

A rule may depend on the gauges as well; the proof shows only that doing so cannot beat $r^*$. The proposition is exact, but its expectation over masses is estimated below by Monte Carlo, not evaluated.

**Protocol and results.** Both computations followed an analysis plan fixed before these numerical results were generated. They are nevertheless post hoc, because $C_2 \times C_2$ was studied throughout rather than held out in a preregistered design. Target labels come from the ordinary joint leading flow in character coordinates, integrated by RK4 with step $0.01$ in $\tau$ and read at $\tau = 60$, $120$ and $180$. A label is resolved when the relative canonical-score gap is at least $0.999$, the selected product is positive and the selected coefficient is at least twice every rival's in each layer, a conservative invariant-cone check; an unresolved label would count as an error.

The raw-block rule was scored on exactly the $20{,}000$ $C_2 \times C_2$ draws behind $q_{\mathrm{op}}$, against target labels at step $0.005$. It disagrees with $S_J$ on $340$ of them, a rate of $0.0170$ with Wilson $95\%$ interval $[0.0152, 0.0189]$ rounded outward, against $3{,}686$ for the comparator of Remark C.2 on the same draws. No label was unresolved, halving the step changed none, and DOP853 at $\mathrm{rtol} = 2 \times 10^{-11}$ reproduced all $404$ labels it replayed, every disagreement of the rule included. The closed-form lifetimes agreed with adaptive integration of the lifted system on $48$ clocks to relative error at most

$1.1 \times 10^{-11}$, and at $\omega = 2$ the leading selector agreed with the constant-$u$ race at $u = 2$ on all 512 draws checked, as (122) requires.

For the optimum, 40,000 independent mass fibers were drawn under $\mu_J$ and all eight sign states of each were integrated, 320,000 trajectories, none unresolved. The minimal number of errors out of eight was 0 on 39,748 fibers, 1 on 247, 2 on 4 and 4 on one, so the estimate is $259/320{,}000 = 8.09 \times 10^{-4}$, with Monte Carlo standard error $0.52 \times 10^{-4}$ computed across fibers, since the eight states of one fiber are not independent draws. The fiber with four errors has label 1 when $t_3 = -1$ and 2 when $t_3 = +1$, whatever $t_1$ and $t_2$: the third sector, which never wins there, decides between the other two. The finite minimization was checked against an independent subset dynamic program on all $3^8$ possible label tables. Halving the step on 2,243 fibers, and replaying 284 fibers with DOP853, both sets containing every fiber with a positive minimum, changed no label and no fiber optimum. These controls are diagnostics, not an integration-error bound; the zero-error fibers were not all replayed; and the estimate concerns the leading selector on one group.

### H.4 Reproducibility

The disagreement measurements of Appendix H.1, the reflected-witness measurement of Appendix H.2 and the computations of Appendix H.3 are controlled float64 experiments, and they are the only experiments this paper reports. Group laws and representation homomorphisms were checked exhaustively, and every Fourier transform received a real spatial-domain input. The recorded runs used CPython 3.14.2, NumPy 2.4.1 and SciPy 1.17.0 on Windows 11. A seed does not fix a Monte Carlo stream across NumPy versions, so a rerun in another environment should be compared with the reported counts through their confidence intervals (Wilson for rates, Clopper–Pearson for the witness bounds) rather than bit for bit. The computations of Appendix H.3 also compile a short C# RK4 kernel with the .NET Framework compiler. The supplementary material contains the code, machine-readable outputs and reproduction commands. All mathematical claims, experimental settings, resolution rules, numerical checks and reported summaries used in the paper are stated in the manuscript.