ACTIVATION TAILS SELECT SUBOPTIMAL PROFILES IN EXPRESSIVE NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

ReLU homogenization removes Gaussian activation tails, but those tails can still determine the leading parameter profile of a positive-risk training trajectory. For any width $m \geq 4$ of the hidden layer and smoothing parameter $0 < \tau \leq 1/5$, we assume a bias-free one-hidden layer neural network with Gaussian-smoothed ReLU activations. All the weights and readouts are trained via population logistic gradient flow on the eight-point binary classification task. A nonempty open set of sufficiently large finite initializations gives rise to trajectories that misclassify two points forever. Both the smooth network and its width-m homogeneous ReLU version are able to classify all eight points. However, the loss for the trajectories of the smooth network tends to a positive non-optimal population loss. The parameter vector, normalized by $\sqrt{\log(t + t_0)}$, converges to one point on an explicitly constructed sphere of local constrained norm minimizers. Four constraints correspond to the margins obtained from the points that were correctly classified, while four constraints are related to the activation tails of the two points that were misclassified. The absence of the latter four constraints yields a different norm-minimizing parameter profile. Under the same Euclidean parameter metric and common layer learning rate, these Gaussian parameter profiles are incompatible with the neuronwise balance invariant of homogeneous ReLU gradient flow. Different smoothing parameters yield the same ReLU homogenization but different leading $\log(t + t_0)$-normalized logits at a specified off-support probe, and different limiting ratios to a reference training logit. The bad basin and profile convergence property are preserved under sufficiently small perturbations of each input coordinate independently and arbitrary positive probability weights.

1 QUESTION AND CONTRIBUTION

Homogenization ensures that the leading network output is stable under parameter rescaling. But does it also dictate the leading training profile for the case when training an expressive model fails to fit its data? We provide an explicit counterexample, where Gaussian activation tails disappear in the homogenized network but still place non-trivial restrictions on the leading parameters of the true positive-risk training trajectory. We prove that this failure persists indefinitely and that the network converges to a specific profile from a full-dimensional open set in a fully trained and bias-free network. Then we discuss why these internal tail restrictions are critical and generalize the basin argument to fixed-step gradient descent and SGD with independent and identically distributed (i.i.d.) samples. Finally, we prove that introduction of a trainable output bias term without changing the initial predictor will fix the failure.

Relation to previous work. Previous research revealed XOR saturation, infinitely large weight sub-optimal configurations, attractors at infinity, and smooth-network bad minima [1; 8; 9; 4; 2; 3; 5].[1] Other failure theorems investigate natural initialization regimes [6]. This theorem provides a dynamical description of profile selection by activation tails in an expressive Gaussian network. For softmax attention with a fixed prediction head, Theorem 3 of Tarzanagh et al. proves directional convergence of gradient descent to a token-selection max-margin solution from sufficiently large

[1] For the 1999 XOR study [4], only the author abstract was available; no detailed comparison with its full proof is claimed.

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.

initializations in a local cone, under their loss, token-score, and step-size assumptions [12]. The selected direction can be globally suboptimal; their joint-head results instead concern regularization paths. The analogy is directional selection through exponentially small softmax contributions. Here the constraints combine learned classification margins with Gaussian activation tails on missed examples. The joint-training concerns cover some special normalized updates and joint value-softmax flows [13; 14]. This paper incorporates the crucial mixed constraints, positive-risk attractivity of a rich Gaussian network from a full-dimensional open set, and persistence under samplewise training with positive activation derivatives. The homogeneous and non-homogeneous implicit bias results require successful-training entry criteria [10; 7]. The ReLU standard balance invariant [11] along with constant-step directed convergence and normal/tangential stochastic control [19; 20; 16] methods are applied. The cited studies about SGD trapping use flat activation intervals [17; 18]. In our case all the finite input Gaussian derivatives are positive, therefore we can control the vanishing gradients and energy dynamics of divergent positive-risk profile.

2 Model and result

Fix $m \geq 4$, $k = m - 2$ and $0 < \tau \leq 1/5$. The inputs are $x \in \{-1, 1\}^3$ with labels $y_x = -x_1 x_2$, probabilities $p_x > 0$ and $\sum_x p_x = 1$. Assign

$$\mathcal{B} = \{(-1, -1, 1), (1, 1, 1)\}, \qquad P_{\text{bad}} = \sum_{x \in \mathcal{B}} p_x, \qquad L_{\text{bad}} = 2P_{\text{bad}} \log 2.$$

The architecture, loss and physical-time flow are

$$f_\Theta(x) = \frac{1}{m} \sum_{j=1}^{m} a_j \phi_\tau(w_j^\top x), \qquad \phi_\tau(z) = z\Phi(z/\tau) + \tau\varphi(z/\tau), \tag{1}$$

$$L(\Theta) = \sum_x 2p_x \log(1 + e^{-y_x f_\Theta(x)}), \qquad \dot{\Theta} = -m\nabla L(\Theta). \tag{2}$$

where Φ, φ are standard normal CDF and density. The activation is Gaussian convolution of ReLU; its derivative is positive for each argument. All $4m$ parameters are trainable. There are no biases. Define the probability-weighted classification error by

$$\text{err}(\Theta) = \sum_x p_x \mathbf{1}\{y_x f_\Theta(x) \leq 0\},$$

counting zero margins as errors. We call the two points in $\mathcal{B}$ the missed examples and the six points in $\mathcal{G} = \{-1, 1\}^3 \setminus \mathcal{B}$ the learned examples (also called good examples in the proof). Along the bad trajectories below, these sets have strictly negative and strictly positive margins, respectively, at every finite time. The missed logits tend to zero from the wrong side; this finite-time sign statement is distinct from classification of limiting zero logits.

Employ the invertible coordinates

$$\begin{aligned} P_1 &= (a_1, r_1, q_1, z_1), & w_1 &= (r_1 + q_1, -r_1 + q_1, -z_1), \\ P_2 &= (a_2, r_2, q_2, z_2), & w_2 &= (-r_2 + q_2, r_2 + q_2, -z_2), \\ N_j &= (b_j, x_j, y_j, c_j), & w_{j+2} &= (x_j + y_j, -x_j + y_j, -c_j), \end{aligned}$$

whose readouts are $(a_1, a_2, -b_1, \ldots, -b_k)$. Neuron coordinates y_j are different from the labels y_x. Their concatenation v has inherited metric $D = \text{diag}(1, 2, 2, 1, \ldots, 1, 2, 2, 1)$ and $J(v) = v^T D v = \|\Theta\|^2$. Equality of neurons is not assumed in the flow. Define

$$S = \sqrt{2}\tau, \quad A^4 + SA^3 - mSA - 2m^2 = 0 \ (A > 0), \quad R = \frac{m}{A} + \frac{S}{2}, \quad T_* = \sqrt{m + 2AS}. \tag{3}$$

For any unit vector $u \in \mathbb{R}^k$, $w_*(u)$ has coordinates

$$P_1 = P_2 = (A, R, 0, S), \qquad B = C = T_* u, \qquad X = Y = 0, \tag{4}$$

where B, X, Y, C collect the negative-neuron coordinates.

Theorem 1 (Open bad basin and a definite profile)**.** *Given each fixed m, τ and positive population as defined above, there exist finite $t_0 > 1$ and a nonempty open $U \subset \mathbb{R}^{4m}$ so that any exact flow starting in U will exist for all $t \geq 0$, misclassify exactly $\mathcal{B}$ for each $t \geq 0$ and satisfies*

$$L(\Theta(t)) \longrightarrow L_{\mathrm{bad}} > 0, \qquad \mathrm{err}(\Theta(t)) = P_{\mathrm{bad}}, \qquad \frac{v(t)}{\sqrt{\log(t+t_0)}} \longrightarrow w_*(u_\infty). \tag{5}$$

The unit vector u_∞ has positive coordinates and can depend on the trajectory. The profiles (4) do not depend on the positive probability weights. For both the smooth and width-m homogeneous ReLU models, the population-loss infimum is zero, approached by scaling the separating readouts.

The open set has full parameter dimension. In the shortcut population family $p_x = (1 + \delta y_x x_3)/8$ with $-1 < \delta < 1$, let $\beta = (1-\delta)/2$. Then $P_{\mathrm{bad}} = \beta/2$ and $L_{\mathrm{bad}} = \beta \log 2$. The theorem still applies in the absence of correlation.

Corollary 1 (Open perturbations of the population)**.** *Given m, τ, there is an open neighborhood of the eight Boolean input configurations in $\mathbb{R}^{24}$ for which Theorem 1 applies, where the profile sphere (4) varies smoothly. The labels are attached to the perturbed inputs; the probabilities can be fixed arbitrary strictly positive weights. The same two specific input configurations are misclassified with loss limit $2P_{\mathrm{bad}} \log 2$. Both architectures remain realizable.*

3 Why the activation tails matter

Place $\rho = \log(t + t_0)$ and use upper-case letters for the coordinates of the scaled vector $w = v/\sqrt{\rho}$. The four learned-margin rates that are active are:

$$\begin{aligned}
g_1 &= \{A_1(2R_1 + Z_1) - B \cdot C - 2B \cdot X\}/m, \\
g_2 &= \{A_2(2R_2 + Z_2) - B \cdot C + 2B \cdot X\}/m, \\
g_3 &= \{B \cdot C + 2B \cdot Y - A_1(Z_1 + 2Q_1) - A_2(Z_2 + 2Q_2)\}/m, \\
g_4 &= \{B \cdot C - 2B \cdot Y - A_1(Z_1 - 2Q_1) - A_2(Z_2 - 2Q_2)\}/m.
\end{aligned} \tag{6}$$

The four activation-tail rates are

$$l_{i,s} = Z_i - 2sQ_i, \qquad k_{i,s} = l_{i,s}^2/(2\tau^2), \qquad i = 1, 2, \quad s = \pm 1. \tag{7}$$

The sample-to-rate correspondence is

i	1	2	3	4
x_i^G	$(1,-1,-1)$	$(-1,1,-1)$	$(1,1,-1)$	$(-1,-1,-1)$

where g_i is the leading signed-margin rate of x_i^G and $p_i^G = p_{x_i^G}$. The rate $k_{i,s}$ belongs to positive neuron i on missed sample $(s, s, 1)$, with $p_s^B = p_{(s,s,1)}$. Thus four tail constraints arise from two neurons on two samples. The remaining learned samples $(1, -1, 1)$ and $(-1, 1, 1)$ have rate two at the profile.

Define the ordered rate vector

$$h = (g_1, g_2, g_3, g_4, k_{1,+}, k_{1,-}, k_{2,+}, k_{2,-}).$$

The profiles (4) are strict local minima of $J/4$ subject to $h \geq 1$, modulo their $k-1$ rotation directions. For the tail interpretation, restrict to the tube around the positive profile cap used in Appendix B, where $Z_i > 2|Q_i|$, the negative-neuron gates satisfy $C_j > 2|X_j|$ and $C_j > 2|Y_j|$, and the strict tail gap (20) holds. In this chamber $l_{i,s} > 0$, so $k_{i,s} \geq 1$ is equivalent to $l_{i,s} \geq S$. Since $l_{i,s} = S > 0$ at the profiles, this restriction does not change the local positive-tail constraint branch. The full algebraic profile sphere is used for the constrained geometry; its positive cap is used for the trapping dynamics. Four learned-margin constraints and four activation-tail constraints are active, while the two remaining learned margins have rate two.

Their influence will be demonstrated without resorting to an asymptotic formula. Fix A, u and let $Z_i = z$, $R_i = m/A + z/2$, $B = C = \sqrt{m + 2Az}\, u$, where the transverse coordinates equal zero. The four learned rate parameters are still equal to one, the two learned rate parameters equal to two, while

$$\frac{J}{4} = \frac{A^2}{2} + \left(\frac{m}{A} + \frac{z}{2}\right)^2 + \frac{z^2}{2} + \frac{m + 2Az}{2}, \qquad \frac{d}{dz}\frac{J}{4} = R(z) + z + A > 0.$$

By reducing z below S the norm will decrease without violating any learned-margin requirement. This goes against (7). The chosen profile is therefore not a local norm minimum for six learned examples.

Capacity is clear. Make use of four hidden weights $\pm(r, -r, 0)$ and $\pm(r, r, 0)$ and readout signs $+, +$ and $-, -$ with a common magnitude $a > 0$ and $r > 0$; set any other readout weights to zero. All ReLU margins equal $2ar/m > 0$. All Gaussian margins equal $a[\phi_\tau(2r)+\phi_\tau(-2r)-2\phi_\tau(0)]/m > 0$, since the activation is strictly convex. Rescaling the readouts makes the loss go to zero. Hence, the bad basin survives with enough capacity in both models.

3.1 A PROFILE EXCLUDED BY HOMOGENEOUS GRADIENT FLOW

Define the bias-free ReLU comparator with the same width, population, and Euclidean parameter metric by

$$f_\Theta^{\mathrm{ReLU}}(x) = \frac{1}{m}\sum_{j=1}^{m} a_j[w_j^\top x]_+, \qquad L_{\mathrm{ReLU}}(\Theta) = \sum_x 2p_x \log(1 + e^{-y_x f_\Theta^{\mathrm{ReLU}}(x)}).$$

A ReLU gradient flow is an absolutely continuous solution of

$$\dot\Theta(t) \in -m\,\partial^C L_{\mathrm{ReLU}}(\Theta(t)) \quad \text{for almost every } t,$$

where ∂^C is the Clarke subdifferential. The common scalar multiplier m applies to both layers. The neuronwise balance invariant [11, Theorem 2.1] gives a dynamical counterpart to the norm comparison.

Corollary 2 (Homogeneous incompatibility)**.** *Given the same Euclidean parameter metric and common scalar learning rate for both layers, every ReLU gradient flow defined above with finite initialization for which the parameter norm blows up cannot asymptotically approach the normalized direction of any nonzero multiple of* (4)*. Specifically, its $v/\sqrt{\log(t+t_0)}$ cannot converge to a nonzero multiple of that profile.*

The proof is deferred to Appendix E.

On the other hand, all ϕ_τ have the same ReLU homogenization, because $\phi_\tau(cz)/c \to [z]_+$ as $c \to \infty$. But the following comparison reveals their profile directions as well as leading functions outside the training support, while different orbit angles alone do not: with positive u, the negative-neuron homogeneous component is identically $-T_*^2[-x_3]_+/m$.

Corollary 3 (Homogenization equality with different profiles and logits)**.** *With fixed m and distinct $\tau \in (0, 1/5]$, there are different directions of normalized parameters for the Gaussian profiles. Along each of them, on its bad basin, with $x_* = (1, -1, 0)$ and $x_{\mathrm{ref}} = (1, -1, 1)$,*

$$\frac{f_{\Theta(t)}(x_*)}{\log(t+t_0)} \to 2 + \frac{AS}{m}, \qquad \frac{f_{\Theta(t)}(x_{\mathrm{ref}})}{\log(t+t_0)} \to 2, \qquad \frac{f_{\Theta(t)}(x_*)}{f_{\Theta(t)}(x_{\mathrm{ref}})} \to 1 + \frac{AS}{2m}. \tag{8}$$

The last expression is strictly increasing as τ increases.

See Appendix E for the proof.

The point x_* does not belong to the eight-point dataset. The expression gives the exact leading logit difference for the same ReLU homogenization.

4 HOW THE FULL-SPACE CAPTURE ARGUMENT WORKS

Write $\mathcal{M}$ for the profile sphere (4) and $J_* = J(w_*(u))$, which does not depend on u. The proof applies in a fixed neighborhood of a positive cap of this sphere. Write L in chart coordinates and use the logarithmic clock $\sigma = \log\rho$. The exact moving energy

$$W(\rho, w) = me^\rho[L(\sqrt{\rho}\,w) - L_{\mathrm{bad}}] + J(w)/4 \tag{9}$$

obeying $w_\sigma = -D^{-1}\nabla W$ has in the fixed neighborhood the form of a positive mixed exponential penalty plus a remainder whose value, gradient, and explicit time derivative are exponentially small

in ρ. The penalty is invariant under simultaneous orthogonal rotations of B, C, X, Y; the Gaussian loss has this symmetry only up to the remainder.

A positive explicit time derivative of a penalty implies a large radial gradient. This dissipation overpowers that positive drift, resulting in

$$\frac{dW}{d\sigma} \le -\tfrac{1}{2}\,\|\nabla W\|^2_{D^{-1}} + \eta(\rho), \qquad \int_{\rho_0}^{\infty} \eta(\rho)\,\frac{d\rho}{\rho} < \infty. \tag{10}$$

A full-space reduced-Hessian computation provides a positive normal curvature. Combining an inverse-function estimate for the constrained KKT equations with that normal energy barrier yields

$$\operatorname{dist}_D(w, \mathcal{M}) \le C\{\|\nabla W\|_{D^{-1}} + \log \rho/\rho\}. \tag{11}$$

The normal energy barrier by itself would not provide control over the drift along $\mathcal{M}$. But for the nearest-orbit phase $u = (B + C)/\,\|B + C\|$, rotation invariance gives

$$\|u_\sigma\| \le C \operatorname{dist}_D(w, \mathcal{M})\,\|\nabla W\|_{D^{-1}} + C\eta(\rho). \tag{12}$$

Combined with a lower energy bound converging to $J_*/4$, these estimates guarantee arbitrarily small total angular motion from a nonempty open set of sufficiently large finite initial states. Normal and angular exit is prevented. Finite angular path length determines the limiting phase; annular crossing makes distance from the orbit converge to zero. Appendices A–D construct this open set and produce all the estimates, the first exit proof, and continuation argument.

5 Persistence under discrete training

The same geometry in the local neighborhood is responsible for gradient flow, fixed-step gradient descent, and SGD using i.i.d. samples.

5.1 The same failure under fixed-step gradient descent

There is an exact discrete-time analog of the Gaussian construction as well. The step size is fixed in physical parameter coordinates using the same Euclidean layer metric.

Theorem 2 (Fixed step Gaussian capture)**.** *For any fixed Boolean population, $m \ge 4$, $0 < \tau \le 1/5$, positive probabilities, and any fixed finite step size $h > 0$, there exists $t_0 > 1$ and a nonempty full-dimensional open set of finite starting points for which the full batch iteration*

$$\Theta_{n+1} = \Theta_n - hm\nabla L(\Theta_n), \qquad n \ge 0, \tag{13}$$

yields sequences that permanently misclassify precisely the examples in $\mathcal{B}$, have non-increasing loss converging to L_{bad}, and satisfy

$$\frac{v_n}{\sqrt{\log(t_0 + nh)}} \longrightarrow w_*(u_\infty)$$

for some unit vector u_∞ with all strictly positive coordinates.

See Appendix F for the interpolation of actual steps in physical time. The gradient and Hessian of the loss are exponentially small in the capture region, so the rescaling of the discretization error is bounded in total and in quadratic norms. These bounds extend the energy and phase argument to every fixed finite step size.

5.2 Constant-step sampled SGD can remain in the bad basin

The failure is also robust to unbiased sampling noise. Let us define the individual loss $\ell_x(\Theta) = 2\log(1 + e^{-y_x f_\Theta(x)})$, where $\mathbb{E}_{X\sim p}\ell_X = L$.

Theorem 3 (High-probability Gaussian SGD capture)**.** *Fix the Boolean population, $m \ge 4$, $0 < \tau \le 1/5$, its positive probabilities $p_x > 0$, any fixed finite $h > 0$, and $\delta \in (0, 1)$. For independent samples $X_n \sim p$, consider*

$$\Theta_{n+1} = \Theta_n - hm\nabla \ell_{X_n}(\Theta_n). \tag{14}$$

There exists $t_0 > 1$ and a deterministic nonempty full-dimensional open set U of finite starts so that for any $\Theta_0 \in U$, with probability at least $1 - \delta$, every iterate misclassifies exactly $\mathcal{B}$, $L(\Theta_n) \to L_{\mathrm{bad}}$, and, in the original chart,

$$\frac{v_n}{\sqrt{\log(t_0 + nh)}} \longrightarrow w_*(u_\infty)$$

for one random unit vector u_∞ with coordinates all strictly positive.

In other words, one single event controls all subsequent iterations, with a uniform per-start probability bound. Appendix G controls the stopped energy and phase martingales and proves normal convergence without assuming absolute summability of the sampling noise. As shown in Appendix I.1, there exists an exterior cone of bad flow starts such that the limit conditional SGD capture probability remains strictly positive and approaches unity as the radius increases indefinitely. Hence, isotropic Gaussian initialization is proven to exhibit a positive lower limiting failure probability as variance tends towards infinity.

6 Architectural scope: an ordinary intercept repairs the basin

With the same Gaussian activation, the initial predictor can be retained with a single additional trainable feature. In particular, letting

$$f^{\mathrm{bias}}_{\Theta,b}(x) = f_\Theta(x) + b, \qquad \mathcal{L}(\Theta, b) = \sum_x 2p_x \log(1 + e^{-y_x(f_\Theta(x)+b)}),$$

and training $\dot{\Theta} = -m\nabla_\Theta \mathcal{L}$ and $\dot{b} = -\eta\partial_b \mathcal{L}$ with a constant value of $\eta > 0$, taking $\eta = m$ results in the common multiplier originally specified.

Theorem 4 (Bias-Recovery)**.** *Under the Gaussian-model assumptions of Theorem 1, there exist a nonempty open ball $K \subset \mathbb{R}^{4m}$ and finite constants $s_0, T_{\mathrm{rec}} > 0$ such that, for every $s \geq s_0$ and $z \in K$, the biasless flow starting at $\Theta_0 = sz$ permanently misclassifies exactly $\mathcal{B}$. For every $b_0 \in (-1, 1)$, the biased flow starting at (sz, b_0) correctly classifies every sample by time T_{rec} and forever thereafter. All these flows are globally defined.*

At $b_0 = 0$, the biased and biasless flows have exactly the same initial predictor on every input and the same initial velocity on all common parameter coordinates.

For each fixed $s \geq s_0$, the biased starts $sK \times (-1, 1)$ form a full-dimensional open recovery set. Exact initial-predictor matching with the biasless start is asserted only at $b_0 = 0$.

A finite time capture proof, below the threshold value of $2 \min_x p_x \log 2$ loss, is included within Appendix H.

The two missed examples share the negative label, whereas all learned examples have large positive margins. As a consequence, the bias moves at order-one speed, whereas none of the initial parameters experience any appreciable alterations throughout this period.

7 Computational checks and limitations

The theoretical findings are analytical. While the inheritance of metric, rank eight Jacobian from constraint equations, and the reduction of Hessian can all be verified analytically through exact symbolic computations, these computations alone do not prove the energy, angular-control, or continuation bounds.

For widths four and eight, $\tau = 0.1, 0.2$, and $\rho = 2, 20, 100, 1000$, we compared the signed-log field at non-symmetric states on Boolean and perturbed supports with a separate original-coordinate implementation using 100-digit arithmetic. The comparisons covered all $4m$ field coordinates and eight margins. Eight automatic-differentiation tests used the same widths, temperatures, and support conditions at unscaled initial states. Four clock/checkpoint tests compared physical-time and log-time Radau/BDF integrations at widths four and eight on Boolean supports, with $\tau = 0.2, \delta = 0.5$ and

$\rho \in [2, 3]$. The field/autograd and clock/checkpoint seeds were 49018 and 49019, respectively. Table 1 reports these diagnostics.

These are author-run implementation diagnostics. Their code, raw states, original-coordinate references, configurations, outputs, environment logs, and sealed checkpoints are retained locally but are not supplied with this PDF. The numerical entries in Table 1 therefore cannot be independently reproduced from the submitted materials; the analytical proofs do not depend on these checks. Signed logarithms preserve classifications when binary64 values underflow; a displayed zero velocity or margin is never treated as an exact zero. These tests concern finite states and a moderate physical-time interval. No numerical experiment establishes indefinite capture, and no new GD, SGD, or bias trajectory campaign is claimed.

Implementation check	Cases	Maximum discrepancy
100-digit field, signs and log magnitudes	32	1.735×10^{-9}
100-digit margin, signs and log magnitudes	32	7.276×10^{-12}
Original-coordinate autograd, relative field norm	8	7.123×10^{-16}
Physical/log-clock endpoint, absolute coordinate error	4	8.415×10^{-10}
Sealed checkpoint reconstruction and continuation	4	bitwise equal

Table 1: Finite-state implementation checks, not evidence of all-time trapping. All sign comparisons passed. Checkpoint continuation used matching Python/NumPy/SciPy environments and compared t, y, h_{abs}, not every internal solver field. No trajectory enclosure or certified basin entry follows.

What the example does and does not explain. The Gaussian theorem identifies a specific failure of homogenization as a description of the leading training parameters. It does not characterize training outcomes from activation tails. The open basin, support perturbation radius, and discrete starting scales can be unfavorable. This does not give useful ordinary-scale initialization probabilities, practical step-size rules, or finite-precision trapping guarantees. The missed margins approach zero from below, leaving no uniform negative-margin gap; finite-state numerical checks alone do not certify indefinite capture. The off-support comparison gives an exact leading logit ratio, but no generalization guarantee. The incompatible ReLU direction does not prove successful training from the same start. The Output Bias Theorem demonstrates the fragility of this no-bias failure when adding an ordinary constant feature. It proves permanent classification on the specified starts, but neither zero limiting loss nor general biased-network convergence.

The Stochastic Gradient Descent Theorem assumes finite supports with independent and identically distributed samples and exact arithmetic; population loss need not be monotone.

REPRODUCIBILITY STATEMENT

The main text specifies the model, data distribution, training dynamics, and assumptions of the theoretical results. The appendices in this PDF provide complete proofs and supporting mathematical details. The computational-checks section reports the diagnostic settings, random seeds, comparison methods, and observed discrepancies. These author-run finite-state checks are reported as implementation diagnostics; their code and raw records are not supplied with this PDF, so the numerical entries cannot be independently reproduced from the submitted materials. They do not certify infinite-time trapping or a finite numerical basin.

AI DISCLOSURE

Checks on proof arguments, debugging errors, and computation implementation were done using OpenAI Codex. Fresh instances of OpenAI Codex provided adversarial feedback.

REFERENCES

[1] D. E. Rumelhart, G. E. Hinton and R. J. Williams. Learning Internal Representations by Error Propagation. In *Parallel Distributed Processing*, vol. 1, chapter 8, 1986. See pp. 332–333. https://web.stanfo

rd.edu/class/psych209a/ReadingsByDate/02_06/PDPVolIChapter8.pdf.

[2] K. L. Blackmore, R. C. Williamson and I. M. Y. Mareels. Local Minima and Attractors at Infinity for Gradient Descent Learning Algorithms. *Journal of Mathematical Systems, Estimation, and Control* 6(2), 1996. https://scholar.lib.vt.edu/ejournals/JMSEC/v6n2/85167.pdf.

[3] H. Petzka and C. Sminchisescu. Non-attracting Regions of Local Minima in Deep and Wide Neural Networks. *Journal of Machine Learning Research* 22(143):1–34, 2021. Theorem 4 and Appendix A.2. https://jmlr.org/papers/volume22/19-586/19-586.pdf.

[4] I. G. Sprinkhuizen-Kuyper and E. W. Boers. A local minimum for the 2-3-1 XOR network. *IEEE Transactions on Neural Networks* 10(4):968–971, 1999. https://doi.org/10.1109/72.774274.

[5] F. Martinelli, A. van Meegen, B. Şimşek, W. Gerstner and J. Brea. Flat Channels to Infinity in Neural Loss Landscapes. NeurIPS 38, 2025; preprint revised 2026. Version 4. https://arxiv.org/abs/2506.14951v4.

[6] E. Boursier and N. Flammarion. Early alignment in two-layer networks training is a two-edged sword. *Journal of Machine Learning Research* 26(183):1–75, 2025. https://jmlr.org/papers/v26/24-1523.html.

[7] Y. Cai, K. Zhou, J. Wu, S. Mei, M. Lindsey and P. Bartlett. Implicit Bias of Gradient Descent for Non-Homogeneous Deep Networks. ICML, PMLR 267:6426–6504, 2025. https://proceedings.mlr.press/v267/cai25m.html.

[8] I. G. Sprinkhuizen-Kuyper and E. J. W. Boers. Het 2-3-1 XOR-Netwerk heeft lokale Minima. Leiden technical report 96–19, 1996. https://liacs.leidenuniv.nl/assets/PDF/TechRep/tr96-19.pdf.

[9] I. G. Sprinkhuizen-Kuyper and E. J. W. Boers. The local minima of the error surface of the 2–2–1 XOR network. Leiden technical report 96–10, 1996. See the cross-entropy discussion, pp. 2–4. https://liacs.leidenuniv.nl/assets/PDF/TechRep/tr96-10.pdf.

[10] K. Lyu, Z. Li, R. Wang and S. Arora. Gradient Descent on Two-layer Nets: Margin Maximization and Simplicity Bias. NeurIPS 2021. Theorem 5.6. https://papers.nips.cc/paper_files/paper/2021/file/6c351da15b5e8a743a21ee96a86e25df-Paper.pdf.

[11] S. S. Du, W. Hu and J. D. Lee. Algorithmic Regularization in Learning Deep Homogeneous Models: Layers are Automatically Balanced. NeurIPS 2018. Theorem 2.1. https://proceedings.neurips.cc/paper_files/paper/2018/file/fe131d7f5a6b38b23cc967316c13dae2-Paper.pdf.

[12] D. Ataee Tarzanagh, Y. Li, X. Zhang and S. Oymak. Max-Margin Token Selection in Attention Mechanism. NeurIPS 36, 2023. https://papers.nips.cc/paper_files/paper/2023/file/970f59b22f4c72aec75174aae63c7459-Paper-Conference.pdf.

[13] B. Vasudeva, P. Deora and C. Thrampoulidis. Implicit Bias and Fast Convergence Rates for Self-attention. Preprint, 2024, revised 2025, version 2. https://arxiv.org/abs/2402.05738v2.

[14] A. Varre, M. Rofin and N. Flammarion. Gradient Flow Polarizes Softmax Outputs towards Low-Entropy Solutions. Preprint, 2026, version 1. https://arxiv.org/abs/2603.06248v1.

[15] L. Chizat, E. Oyallon and F. Bach. On Lazy Training in Differentiable Programming. NeurIPS 32, 2019. https://papers.neurips.cc/paper/8559-on-lazy-training-in-differentiable-programming.pdf.

[16] M. Benaïm. Dynamics of stochastic approximation algorithms. *Séminaire de probabilités* 33:1–68, 1999. Sections 4 and 7. https://www.numdam.org/item/SPS_1999__33__1_0.pdf.

[17] P. Cheridito, A. Jentzen and F. Rossmannek. Non-convergence of stochastic gradient descent in the training of deep neural networks. Preprint, version 2, 2021. Lemma 3.4 and Theorem 5.3. https://arxiv.org/abs/2006.07075v2.

[18] T. Do, A. Jentzen and A. Riekert. Non-convergence to the optimal risk for Adam and stochastic gradient descent optimization in the training of deep neural networks. Preprint, arXiv:2503.01660v1, 2025. https://arxiv.org/html/2503.01660v1.

[19] M. S. Nacson, N. Srebro and D. Soudry. Stochastic Gradient Descent on Separable Data: Exact Convergence with a Fixed Learning Rate. AISTATS, PMLR 89:3051–3059, 2019. https://proceedings.mlr.press/v89/nacson19a.html.

[20] B. Fehrman, B. Gess and A. Jentzen. Convergence Rates for the Stochastic Gradient Descent Method for Non-Convex Objective Functions. *Journal of Machine Learning Research* 21(136):1–48, 2020. https://jmlr.org/papers/volume21/19-636/19-636.pdf.

A CONSTRAINED GEOMETRY IN THE FULL PARAMETER SPACE

The quartic in (3) has exactly one positive root by Descartes' rule of signs. Its values at $\sqrt{m}$ and $2^{1/4}\sqrt{m}$ have opposite signs. Hence

$$\sqrt{m} < A < 2^{1/4}\sqrt{m}, \qquad 0 < S < 0.3. \tag{15}$$

Substitution gives $h(w_*(u)) = \mathbf{1}$ and $J_* = 2A^2 + 4R^2 + 2S^2 + 2T_*^2$. For the Lagrangian $J/4 - \sum_j \lambda_j(h_j - 1)$, the positive multipliers are

$$\lambda = \frac{mR}{2A} \quad (g_1, g_2), \qquad \mu = \lambda + \frac{m}{4} \quad (g_3, g_4), \qquad \nu = \frac{\tau^2}{2S}\left[\frac{S}{2} + \frac{(\lambda + m/2)A}{m}\right] \quad (k_{i,s}). \tag{16}$$

The stationarity equations reduce to the quartic.

Simultaneous orthogonal rotations of B, C, X, Y preserve J, every h_j, A_i, and $l_{i,s}$. Algebraic calculations may therefore use $u = e_1$; this point is not a physical trapping seed, since some readouts vanish there. An eight-column Jacobian minor has absolute value

$$\frac{4096A^2(2AS + m)}{m^4 S^4} > 0.$$

The eight gradient columns are independent in the full $4m$-dimensional space. At e_1 every constraint-tangent vector can be written as

$$\begin{aligned}
\Delta Z_i &= \Delta Q_i = 0, & \Delta Y_{\parallel} &= 0,\\
\Delta A_1 &= a + b, & \Delta A_2 &= a - b,\\
\Delta B_{\parallel} &= Sa/T_* + h_{\parallel}, & \Delta C_{\parallel} &= Sa/T_* - h_{\parallel},\\
\Delta X_{\parallel} &= j, & (\Delta R_1, \Delta R_2) &= (p + q, p - q),\\
p &= -ma/A^2, & q &= -db + ej,\\
d &= (2R + S)/(2A), & e &= T_*/A, \qquad K_0 = 2R/A,\\
\Delta B_{\perp} &= z + h_{\perp}, & \Delta C_{\perp} &= z - h_{\perp},\\
\Delta X_{\perp} &= x_{\perp}, & \Delta Y_{\perp} &= y_{\perp}.
\end{aligned}$$

Here $z, h_{\perp}, x_{\perp}, y_{\perp} \in \mathbb{R}^{k-1}$ are independent. Direct differentiation of the full Lagrangian gives its Hessian quadratic form

$$a^2 + b^2 + 2(p^2 + q^2) - 2K_0(ap + bq) + 2\|h\|^2 + j^2 + \|x_{\perp}\|^2 + \|y_{\perp}\|^2. \tag{17}$$

The a coefficient is $1 + 2m^2/A^4 + 2K_0 m/A^2 > 0$. The (b, j) block has

$$M_{11} = 1 + 2d^2 + 2K_0 d, \quad M_{22} = 1 + 2e^2, \quad M_{12} = -e(2d + K_0),$$

and determinant $1 + 2d^2 + 2K_0 d + e^2(2 - K_0^2)$. By (15), $d > 0.7$, $1.4 < K_0 < 2.15$, and $e^2 < 1.3$. When $2 - K_0^2 < 0$, the determinant is larger than

$$1 + 2(0.7)^2 + 2(1.4)(0.7) - 1.3[(2.15)^2 - 2] = \frac{2123}{4000} > \frac{1}{2};$$

when the coefficient is nonnegative it can be dropped. The only zero directions in (17) are the $k - 1$ directions z along the orbit. All other $3m - 5$ constraint-tangent directions have positive curvature. Each added perpendicular negative-neuron coordinate contributes the separate block $2h_j^2 + x_j^2 + y_j^2$ and one zero z_j, proving the statement at every finite width without a numerical extrapolation.

Choose a small closed spherical cap around $u_0 = (1, \ldots, 1)/\sqrt{k}$, whose slightly larger cap satisfies $u_j > 1/(2\sqrt{k})$ for every j. Near this cap the closest point of the full algebraic orbit $\mathcal{M}$ in the D metric is $w_*(u(w))$, where

$$u(w) = \frac{B + C}{\|B + C\|}, \qquad d_{\mathcal{M}}(w) = \|w - w_*(u(w))\|_D. \tag{18}$$

This follows by minimizing the orbit-dependent cross term $-2T_*(B + C) \cdot u$ in the squared distance.

Lemma 1 (Local geometry modulo rotations)**.** *In a sufficiently small tube around the cap, feasible points satisfy local quadratic growth of $J/4$ above $J_*/4$ in the normal distance $d_{\mathcal{M}}$. The equality-KKT system on a phase slice has an invertible derivative at the orbit and a uniform local residual error bound. Every nearby inequality-KKT point belongs to $\mathcal{M}$ and has all eight constraints active.*

Proof. Rotate the phase to a fixed reference vector and impose the slice $(B+C)_\perp = 0$. The rotation directions are transverse to this slice. The eight projected gradients remain independent, since the constraint gradients annihilate orbit directions, and (17) is positive on the constraint tangent in the slice. The usual slack-coordinate Taylor argument with positive multipliers gives the strict constrained minimum and quadratic growth in the slice. Rotations transfer the same bounds throughout the cap.

For explicitness, let E_s be an orthonormal column basis of the linear slice directions $(B+C)_\perp = 0$, and write $w = w_*(u_0) + E_s\xi$. For support X near the Boolean support X_0, define

$$\mathcal{K}(\xi, \lambda; \mathsf{X}) = \begin{pmatrix} E_s^T[\nabla J(w)/4 - G(w;\mathsf{X})\lambda] \\ h(w;\mathsf{X}) - \mathbf{1} \end{pmatrix}.$$

At the Boolean solution its derivative in (ξ, λ) is

$$\begin{pmatrix} E_s^T H_{\mathrm{Lag}} E_s & -E_s^T G \\ G^T E_s & 0 \end{pmatrix}.$$

A null vector has a primal component tangent to the constraints. Multiplication by that component and positive reduced curvature force it to vanish; gradient independence then forces the multiplier component to vanish. The inverse function theorem gives the residual bound. If G has the eight full gradient columns, the coefficients $G^\dagger \nabla(J/4)$ are continuous and strictly positive near $\mathcal{M}$ by (16). At an inequality-KKT point these are exactly its multipliers. Thus no constraint can be inactive, and the local uniqueness on the slice identifies the point with the orbit. □

Choose an outer tube of radius $2r$ with these properties and a smaller radius r. Compactness and strict constrained growth give $d_0 > 0$ such that

$$h \geq 1, \quad d_{\mathcal{M}} = r, \quad u \text{ in the cap} \quad \Longrightarrow \quad J/4 \geq J_*/4 + d_0. \tag{19}$$

No statement about a trajectory limit is used to construct this region. The order of choices is fixed: choose nested positive caps and a phase slice, then the geometric tubes; shrink the tubes for the gate gaps in the next section, and obtain d_0 for the final radius r. Fix $\overline{E} = J_*/4 + d_0$ before choosing the small initial excess e_0. Only after these choices take ρ_0 large enough for all uniform estimates, radial feasibility corrections, and tail budgets.

B Uniform loss expansion and dissipation

Shrink the outer tube until all $h \geq h_{\min} > 3/4$, all readout magnitudes are positive, and the relevant gates have strict slack. In particular $Z_i > 2|Q_i|$, $C_j > 2|X_j|$, $C_j > 2|Y_j|$, and

$$\min_j (C_j - 2|Y_j|) > \max_i (Z_i + 2|Q_i|). \tag{20}$$

At the balanced profile $C_j = T_*/\sqrt{k} > 1$; on the cap, $C_j > T_*/(2\sqrt{k}) > 1/2 > S$. Also $2R - S = 2m/A$ has a strict positive gap. These are the signed gate conditions used to interpret (7); in this tube $k_{i,s} \geq 1$ is equivalent to $l_{i,s} \geq S$. Thus the discarded negative-neuron and positive-input activation tails have rates larger than $1 + \epsilon_0$ after further shrinking. The other two learned margins have rates larger than $1 + \epsilon_0$. All constants in this proof may depend on the fixed m, τ, positive probabilities, and cap.

For the missed input $(s, s, 1)$ the exact margin is

$$V_s = \frac{1}{m}\left\{ \sum_j \sqrt{\rho} B_j \phi_\tau[-\sqrt{\rho}(C_j - 2sY_j)] - \sum_i \sqrt{\rho} A_i \phi_\tau[-\sqrt{\rho}(Z_i - 2sQ_i)] \right\}. \tag{21}$$

The negative-label input $(s, s, -1)$ has leading margin ρg_3 or ρg_4, plus V_{-s}. The positive-label margins give the other two active rates and the two redundant rates, with only discarded tails as corrections. Let p_i^G denote the four probabilities corresponding to g_i, and p_s^B those of the missed inputs.

Lemma 2 (Uniform expansion)**.** *For w in the fixed outer tube and cap, and sufficiently large ρ,*

$$L(\sqrt{\rho}w) - L_{\text{bad}} = \sum_{i=1}^{4} 2p_i^G e^{-\rho g_i} + \sum_{i=1}^{2} \sum_{s=\pm 1} \frac{p_s^B}{m} \sqrt{\rho} A_i \phi_\tau(-\sqrt{\rho} l_{i,s}) + E(\rho, w). \tag{22}$$

For $R_0 = me^{\rho}E$, some finite $C, N, \epsilon > 0$ give

$$|R_0| + \|\nabla_w R_0\|_{D^{-1}} + |\rho \partial_\rho R_0| \le C\rho^N e^{-\epsilon\rho} =: \eta(\rho). \tag{23}$$

Proof. For $z > 0$, $v = z/\tau$, the exact integral identity

$$\phi_\tau(-z) = \frac{\tau\varphi(v)}{v^2} \int_0^\infty q e^{-q} e^{-q^2/(2v^2)}\, dq$$

bounds this expression between $(1 - 3/v^2)\tau\varphi(v)/v^2$ and $\tau\varphi(v)/v^2$ for $v \ge \sqrt{6}$. Derivatives use $\Phi(-v)$ and introduce only polynomial factors at the same exponential rate. The missed logistic terms have expansion $2p\ell(V) = 2p\log 2 - pV + O(V^2)$, giving exactly the positive tail coefficient in (22). Negative-neuron tails in (21) have rate strictly above one. All retained good and harmful tail rates are at least $h_{\min}$. Their quadratic logistic remainders and products of a good loss with its small tail correction have rate at least $2h_{\min} > 1$. The discarded learned margins and activation tails have rates above one with uniform slack. Differentiation in w or ρ adds polynomial factors and preserves these gaps. This proves (23) uniformly, without assuming that a trajectory stays in the region. □

For the exact energy (9), differentiating the original flow gives

$$w_\sigma = -D^{-1}\nabla W, \qquad W_\sigma = -\|\nabla W\|_{D^{-1}}^2 + \rho\partial_\rho W. \tag{24}$$

Indeed $d\rho/dt = e^{-\rho}$, $v = \sqrt{\rho}w$, and the derivative of $J/4$ accounts for $-w/2$ in the scaled equation. By Lemma 2,

$$\begin{aligned} W &= W_0 + R_0, & W_0 &= J/4 + \sum_i H_i^G + \sum_{i,s} H_{i,s}^B, \\ H_i^G &= 2mp_i^G e^{\rho(1-g_i)}, & H_{i,s}^B &= p_s^B \sqrt{\rho} A_i e^{\rho} \phi_\tau(-\sqrt{\rho} l_{i,s}). \end{aligned} \tag{25}$$

Every H is positive. Every term of W_0 is rotation-invariant, including for unequal probabilities.

Lemma 3 (Eventual uniform dissipation)**.** *Throughout the temporary tube and cap, for sufficiently large ρ,*

$$W_\sigma \le -\tfrac{1}{2}\|\nabla W\|_{D^{-1}}^2 + \eta(\rho). \tag{26}$$

Proof. Put $X_0 = \rho \sum H$ and $\psi(z) = \Phi(-z/\tau)/\phi_\tau(-z)$. Mills bounds give $\psi(z) \ge z/\tau^2$. For the explicit derivatives $P = \rho\partial_\rho H$,

$$P_G = \rho(1-g)H_G, \qquad P_B = [\rho + 1/2 - z\psi(z)/2]H_B, \qquad z = \sqrt{\rho} l.$$

Their sum is at most $2X_0$. If any $P_G > 0$, its H_G has a uniform positive lower bound. If any $P_B > 0$, the corresponding tail rate satisfies $k_{i,s} < 1 + 1/(2\rho)$, and the Gaussian lower bound gives $H_B \ge c\rho^{-1/2}$. Thus whenever any explicit derivative is positive, $X_0 \ge c\sqrt{\rho}$. Homogeneity and the tail radial derivative $(1 - z\psi(z))H_B$ imply

$$w \cdot \nabla W_0 \le J_{\max}/2 - h_{\min} X_0$$

for large ρ. By Cauchy–Schwarz and (23), $\|\nabla W\|_{D^{-1}} \ge h_{\min} X_0/(2\sqrt{J_{\max}})$ in that case. Its square eventually exceeds twice the sum of the positive drift terms. If no P is positive, the explicit leading drift is nonpositive. Combining both cases with (24) and (23) gives (26). The remainder integrates in the correct clock as $\int_{\rho_0}^\infty \eta(\rho)d\rho/\rho$, tending to zero with ρ_0. □

C NORMAL ERROR, ANGULAR CONTROL, AND CONVERGENCE

Lemma 4 (Normal error bound)**.** *On a sufficiently small tube and for sufficiently large ρ,*

$$d_{\mathcal{M}}(w) \le C_1\{\|\nabla W\|_{D^{-1}} + \log\rho/\rho\}. \tag{27}$$

Proof. Let G contain the eight independent rate gradients. Differentiating (25) yields

$$\nabla W = \nabla(J/4) - G\lambda_\rho + B_0 + \nabla R_0. \tag{28}$$

For a good penalty $\lambda_\rho = \rho H$. For a tail it is $\rho H\,\tau^2\psi(\sqrt{\rho}l)/(\sqrt{\rho}l)$. Its extra factor lies between 1 and $(1-3\tau^2/(\rho l^2))^{-1}$, hence is $1+O(1/\rho)$ uniformly. The amplitude derivative satisfies $\|B_0\| \le C\,\|\lambda_\rho\|/\rho$. Apply a uniformly bounded left inverse $G^\dagger$ in (28) and absorb this last contribution to obtain $\|\lambda_\rho\| \le C(1+\|\nabla W\|)$.

When the gradient is below a sufficiently small fixed threshold, continuity of $G^\dagger\nabla(J/4)$, a small tube radius, and (16) force every component of λ_ρ between fixed positive constants close to its KKT value. Taking logarithms in the exact formulas for the penalties gives

$$|h_j - 1| \le C\log\rho/\rho. \tag{29}$$

For tails one uses $H^B_{i,s} = c_{i,s}(w)\rho^{-1/2}e^{\rho(1-k_{i,s})}(1+O(1/\rho))$ with $c_{i,s}(w)$ uniformly positive. The KKT stationarity residual in (28) is at most $C(\|\nabla W\| + 1/\rho)$. Rotate to the phase slice and apply Lemma 1's inverse-function residual estimate, together with (29). This proves (27) for small gradient. For larger gradient it follows by increasing C_1, since $d_{\mathcal{M}}$ is bounded by the tube radius. No global bound on the stiff Hessian of W is invoked. □

Lemma 5 (Lower energy and angular velocity)**.** *On an energy sublevel $W \le \overline{E}$ in the small tube, with fixed $\overline{E} > J_*/4$, one has*

$$W \ge J_*/4 - \zeta(\rho), \qquad 0 \le \zeta(\rho) \le C\log\rho/\rho + \eta(\rho), \tag{30}$$

$$\|u_\sigma\| \le C d_{\mathcal{M}}(w)\,\|\nabla W\|_{D^{-1}} + C\eta(\rho). \tag{31}$$

The bound on ζ admits a decreasing upper envelope eventually.

Proof. Positivity of all penalties bounds each H on the energy sublevel. The lower Gaussian estimate then forces $h_j \ge 1-b(\rho)$ with $b(\rho) = C\log\rho/\rho < 1/2$. Scale w to $\widehat{w} = (1-b)^{-1/2}w$. Homogeneity makes every constraint feasible; the phase is unchanged and $\widehat{w}$ stays inside the outer tube for large ρ. There $J(\widehat{w})/4 \ge J_*/4$ by Lemma 1. Since $W \ge J(w)/4 - \eta$, this gives (30).

For a unit $e \perp u$, set $\Omega = eu^T - ue^T$. Let K_Ω apply Ω to each of B, C, X, Y and zero to positive-neuron coordinates. It is skew-adjoint in the D metric, and $(K_\Omega w)\cdot\nabla W_0 = 0$. At $w_*(u)$ the D-gradient of the scalar differential $e\cdot du$ is $K_\Omega w_*(u)/(2T_*^2)$. Differentiating (18) shows that at nearby w its difference from $K_\Omega w/(2T_*^2)$ is $O(d_{\mathcal{M}})$, uniformly in e. Apply (24), the symmetry identity, and (23), then take the supremum over unit $e \perp u$, to prove (31). Rotation invariance alone does not assert a constant phase. □

We now close both possible exits. On $d_{\mathcal{M}} = r$ with phase in the cap, the energy satisfies, uniformly for large ρ,

$$W \ge J_*/4 + 3d_0/4. \tag{32}$$

Otherwise take a convergent subsequence of violating boundary points as $\rho\to\infty$. A limiting constraint below one makes its positive penalty diverge. If all limiting constraints are feasible, the limiting objective contradicts (19). The remainder vanishes in either case. This controls the normal boundary only.

Take initial phases in a smaller cap separated from the outer cap boundary by $a_0 > 0$. Choose small $e_0 > 0$ and then large finite ρ_0; impose

$$d_{\mathcal{M}}(w_0) < r/2, \quad u(w_0) \text{ in the smaller cap}, \quad W(\rho_0, w_0) < J_*/4 + e_0. \tag{33}$$

This set is nonempty: use $w_0 = (1+\log\rho_0/\rho_0)w_*(u_0)$. Each rate is the square of the radial factor; good penalties are $O(\rho_0^{-2})$ and tails are $O(\rho_0^{-5/2})$, while $J(w_0)\to J_*$. All conditions in (33) are strict and continuous in $4m$ independent coordinates.

Until a first normal or angular exit, Lemma 3 bounds energy by $J_*/4 + e_0 + E_{\text{tail}}$, where $E_{\text{tail}} = \int_{\rho_0}^{\infty} \eta(\rho) d\rho/\rho$. Choose this below $J_*/4 + d_0/2$, preventing normal exit and providing a fixed $\overline{E}$ for Lemma 5. Integrating dissipation and using (30) gives, for every such terminal time,

$$\int \|\nabla W\|_{D^{-1}}^2 \, d\sigma \le 2[e_0 + \zeta(\rho_0) + E_{\text{tail}}] =: d_{\text{small}}, \tag{34}$$

$$\int d_{\mathcal{M}}(w)^2 d\sigma \le 2C_1^2[d_{\text{small}} + F_{\text{tail}}], \qquad F_{\text{tail}} = \int_{\rho_0}^{\infty} (\log \rho/\rho)^2 \, d\rho/\rho \longrightarrow 0. \tag{35}$$

Cauchy–Schwarz in (31) bounds total phase path length by

$$C\sqrt{(d_{\text{small}} + F_{\text{tail}})d_{\text{small}}} + CE_{\text{tail}}. \tag{36}$$

Choose e_0 small and then ρ_0 large enough that this is less than $a_0/2$. Angular exit is impossible as well, closing the bootstrap. The exact flow is smooth, and $v = \sqrt{\rho} w$ stays bounded on every finite physical-time interval, so it cannot blow up at finite time. Set $t_0 = e^{\rho_0}$. The inverse coordinate map and scale $\sqrt{\rho_0}$ carry (33) to a nonempty open $U \subset \mathbb{R}^{4m}$.

Finite phase path length gives $u(\sigma) \to u_\infty$ in the interior of the positive cap. By (27), outside any fixed small normal tube the gradient is bounded below for sufficiently large ρ. Finite dissipation prevents staying there forever. Nor can there be infinitely many crossings from $d_{\mathcal{M}} = \varepsilon$ to $d_{\mathcal{M}} = 2\varepsilon$: distance to the fixed closed set $\mathcal{M}$ is 1-Lipschitz in D, speed equals $\|\nabla W\|_{D^{-1}}$, and each such crossing costs at least $c_\varepsilon \varepsilon$ in squared-speed dissipation. Thus $d_{\mathcal{M}} \to 0$. Together with phase convergence, this proves the exact profile limit.

The six learned margins diverge positively, because their rates remain above $h_{\min}$ and their corrections vanish. By (20), the positive contribution from the negative-readout neurons in (21) is exponentially smaller than the negative contribution from the positive-readout neurons, uniformly on the tube. Enlarging the finite ρ_0 ensures $V_s < 0$ from elapsed time zero, while $V_s \to 0$. Equation (22) gives $L \to L_{\text{bad}}$ and error P_{bad}. The capacity construction in the main text completes Theorem 1. No universal limiting angle, convergence rate, or bounded unscaled transverse coordinate is claimed.

For completeness, the direct contribution of either missed example to any negative neuron's physical parameter velocity has finite future integral. Its preactivation is at most $-\sqrt{\rho}\, l_{\min}$ with $l_{\min}^2/(2\tau^2) > 1$. Mills bounds and readouts of order $\sqrt{\rho}$ give $\|F_{j,x}(t)\| \le C(t + t_0)^{-\kappa}$ for some $\kappa > 1$, and therefore

$$\int_t^{\infty} \|F_{j,x}(s)\| \, ds \le \frac{C}{\kappa - 1}(t + t_0)^{1-\kappa}.$$

The loss residual on that example tends to $1/2$. This is a bound on its direct contribution; other examples also move the neuron, so total neuron motion is not asserted finite.

D Independent perturbations of the support

We prove Corollary 1. All gates in the temporary region have strict slack, so the same gate pattern persists under small independent changes to all 24 input coordinates. Each leading learned margin remains a homogeneous quadratic function of the parameters, now depending smoothly on the support. Retain the four active learned margins; the other two stay above one. The four harmful rates remain squares of positive linear forms on positive-neuron coordinates. All negative neurons share the same gate pattern, so their contribution to each active learned rate is a support-dependent linear combination of $B \cdot C, B \cdot X, B \cdot Y$. The leading energy and every constraint remain invariant under simultaneous $O(k)$ rotations.

Apply the equality-KKT inverse function theorem on the fixed phase slice from Lemma 1. There is a unique smooth branch of local slice solutions and multipliers. Rank eight, strictly positive multipliers, and positive reduced curvature transverse to the orbit persist. Slice stationarity implies full stationarity: invariance annihilates the orbit directions, which remain transverse to the slice.

At a continued KKT point, the negative-neuron part of the Lagrangian has the form

$$\tfrac{1}{4}\|B\|^2 + \tfrac{1}{2}\|X\|^2 + \tfrac{1}{2}\|Y\|^2 + \tfrac{1}{4}\|C\|^2 - B \cdot (\alpha X + \beta Y + \gamma C)$$

for scalars depending on support and multipliers. Its stationary equations force $X = \alpha B$, $Y = \beta B$, and $C = 2\gamma B$. Thus the four negative vectors remain collinear and the orbit remains a sphere of dimension $k-1$, not a higher-dimensional rotation orbit. Write it as fixed positive-neuron coordinates plus $(B, X, Y, C) = (b_* u, x_* u, y_* u, c_* u)$, with smoothly varying coefficients. Its closest phase is the normalization of

$$b_* B + 2x_* X + 2y_* Y + c_* C,$$

and its squared radius in the negative D metric is $b_*^2 + 2x_*^2 + 2y_*^2 + c_*^2 > 0$. The proof of (31) holds with this radius replacing $2T_*^2$. For clarity, the Boolean identity involving V_{-s} is not used for perturbed support. Let $\mathcal{W}_i(w)$ be the hidden weights reconstructed from the linear chart, let ξ_s^B be the perturbed missed inputs, and put $l_{i,s}(\mathsf{X}, w) = -\mathcal{W}_i(w) \cdot \xi_s^B > 0$ for $i = 1, 2$. Let $g_i(\mathsf{X}, w)$ be the four active homogeneous learned margins under the preserved gate pattern. Directly expanding the eight logistic terms gives

$$L_{\mathsf{X}}(\sqrt{\rho} w) - L_{\mathrm{bad}} = \sum_i 2p_i^G e^{-\rho g_i(\mathsf{X}, w)} + \sum_{i,s} \frac{p_s^B}{m} \sqrt{\rho} A_i \phi_\tau[-\sqrt{\rho} l_{i,s}(\mathsf{X}, w)] + E_{\mathsf{X}}(\rho, w).$$

All learned-margin corrections, even those no longer equal to a missed margin, have Gaussian rates above a fixed lower bound greater than $3/4$; multiplying by a learned loss gives a rate greater than one. Quadratic missed remainders likewise have rate greater than one. Discarded negative-neuron tails and redundant learned margins retain their strict gaps above one. On a sufficiently small closed support neighborhood and the continued tube, the same bound (23) holds for $me^{\rho} E_{\mathsf{X}}$ and its stated derivatives, uniformly in support for fixed positive probabilities. This is the separate perturbed-support expansion required by the argument. All other local KKT bounds and gate gaps persist after shrinking the joint neighborhood. Lemmas 2–5 and the same two-boundary bootstrap therefore give the open basin and convergence to one continued profile. Probability weights enter positive prefactors, so any fixed positive weights are admissible with their own finite constants and starting scale.

The fixed four-neuron separating construction has strict finite margins for both activations. These persist under sufficiently small support changes before readout scaling. Intersect its capacity neighborhood with the geometric neighborhood above. This proves the corollary without a quantitative radius, an assumption of exact XOR labels on perturbed coordinates, or a claim about random initialization.

E Proofs of the homogeneous and Gaussian-family comparisons

E.1 Proof of Corollary 2

Proof. For each ReLU neuron, homogeneity gives $a_j \partial_{a_j} L_{\mathrm{ReLU}} = w_j \cdot \partial_{w_j} L_{\mathrm{ReLU}}$. Hence

$$\frac{d}{dt}(a_j^2 - \|w_j\|^2) = -2m(a_j \partial_{a_j} L_{\mathrm{ReLU}} - w_j \cdot \partial_{w_j} L_{\mathrm{ReLU}}) = 0.$$

At a kink the identity holds for every limiting gradient, hence every Clarke subgradient, and therefore almost everywhere along an absolutely continuous subgradient flow. For either positive neuron in the Gaussian profile, the quartic instead gives

$$A^2 - 2R^2 - S^2 = -S(A + R + S) < 0. \tag{37}$$

Along the Gaussian flows this is the limit of the neuron imbalance divided by $\log(t + t_0)$. A finite ReLU imbalance is constant. Dividing it by any divergent squared parameter norm gives zero, whereas the Gaussian normalized direction has nonzero imbalance. This excludes the profile direction, including its nonzero rescalings or neuron permutations. It does not assert that ReLU training succeeds. □

E.2 Proof of Corollary 3

Proof. Implicit differentiation of the quartic, with $S = \sqrt{2}\tau$, gives

$$A'(S) = -\frac{A(A^2 - m)}{4A^3 + 3SA^2 - mS} < 0, \qquad (AS)' = \frac{2A^3(2A + S)}{4A^3 + 3SA^2 - mS} > 0.$$

The denominator is positive since $A^2 > m$. Thus S/A increases strictly, distinguishing directions even after overall normalization. Uniform convergence of $\phi_\tau(cz)/c$ to ReLU on bounded z, and

Theorem 1, imply $f_{\Theta(t)}(x)/\rho \to f^{\mathrm{ReLU}}_{w_*}(x)$ for every fixed x, including zero gates. At x_* this is $2AR/m = 2 + AS/m$; at the reference training input it is two. This proves (8) and strict monotonicity. □

F Fixed-step capture for the original Gaussian model

We prove Theorem 2. All derivatives and updates in this appendix use the original Euclidean parameter metric. Apply the fixed isometry from the chart with metric D to physical coordinates; write $\mathcal{M}$ for the resulting profile manifold. The concluding limit is then converted back to the original chart. This coordinate change does not alter the layer learning rates.

F.1 Pointwise energy estimates and physical derivative bounds

For a continuous interpolation parameter $t \geq 0$, put

$$\begin{aligned} \rho = \log(t_0 + t), \qquad \sigma = \log \rho, \qquad w = \Theta/\sqrt{\rho}, \\ W(\rho, w) = me^{\rho}[L(\sqrt{\rho}w) - L_{\mathrm{bad}}] + \|w\|^2/4, \qquad p = \nabla_w W. \end{aligned} \tag{38}$$

Choose nested compact positive phase caps and a sufficiently small fixed normal tube $\mathcal{K}$. Every gate has its prescribed strict sign there. All four retained learned-margin rates and four positive-neuron tail rates exceed $3/4$; the discarded learned and negative-neuron tail rates have their original strict gaps above one. These conditions hold on the tube before asserting capture of any discrete trajectory.

The exact Gaussian expansion and constrained geometry from Appendices A–C give the following pointwise estimates for sufficiently large ρ:

$$\begin{aligned} \rho\partial_\rho W \leq \tfrac{1}{2}\|p\|^2 + \psi(\rho), \qquad & d_{\mathcal{M}}(w) \leq C[\|p\| + r(\rho)], \\ W(\rho, w) \geq F_* - \zeta(\rho), \qquad & F_* = \|w_*\|^2/4. \end{aligned} \tag{39}$$

The lower bound is used on a fixed bounded-energy region. Here $\zeta \to 0$, $\psi \geq 0$, and the exact remainder bound η_R permits

$$r(\rho) = C[\log \rho/\rho + \eta_R(\rho)], \qquad \int^{\infty} \psi(\rho)\frac{d\rho}{\rho} < \infty, \qquad \int^{\infty} r(\rho)^2\frac{d\rho}{\rho} < \infty.$$

In particular (39) is an explicit-time drift inequality for the energy, not an inequality that presumes the interpolant follows its unperturbed gradient flow. The nearest-profile projection π is smooth with bounded derivative. The leading rotation symmetry also gives

$$\|D\pi(w)[-p]\| \leq Cd_{\mathcal{M}}(w)\|p\| + C\eta_R(\rho). \tag{40}$$

These follow from Lemmas 3, 4 and 5. Equation (32) gives a strict normal-boundary energy gap on bounded-energy sublevels at sufficiently large ρ.

There are fixed constants C, N such that, on $\mathcal{K}$,

$$\|\nabla_\Theta L(\sqrt{\rho}w)\| + \left\|\nabla^2_\Theta L(\sqrt{\rho}w)\right\| \leq C\rho^N e^{-a\rho}, \qquad a = 2/3, \tag{41}$$

for all sufficiently large ρ. To verify this directly, a learned margin is $\rho g_x(w)$ plus an exponentially small polynomially bounded correction, with $g_x > 3/4$. Its logistic residual and second loss derivative are bounded by $Ce^{-\rho g_x}$, while its first two physical logit derivatives grow at most polynomially in ρ. On a missed example every gate is negative. The Gaussian tail and each required activation derivative are a polynomial factor times $e^{-\rho l_j(w)^2/(2\tau^2)}$, whose smallest relevant rate exceeds $3/4$. The first two logit derivatives are thus exponentially small, even though the logistic residual is not. The chain rule and finite population sum prove (41). Compactness and fixed positive τ make the constants uniform; taking $a = 2/3$ leaves slack. This is a physical-space Hessian estimate. No uniform bound on the potentially stiff Hessian of W is asserted.

F.2 An exact interpolant with integrable forcing

Interpolate (13) linearly at $t_n = nh$. On $t_n \leq t < t_{n+1}$ its exact derivative is

$$\dot\Theta(t) = -m\nabla L(\Theta_n) = -m\nabla L(\Theta(t)) + e(t), \qquad e(t) = m[\nabla L(\Theta(t)) - \nabla L(\Theta_n)]. \tag{42}$$

First work only until this normalized interpolant exits the temporary normal tube or phase cap. Every segment between Θ_n and the current interpolated state is itself part of that stopped path; its Hessian bound therefore uses only states in the tube. We do not assume that an unchecked next endpoint stays there.

Take $t_0 \geq 2h$. Within one step, $0 \leq \rho(t) - \rho(t_n) \leq h/(t_0 + t_n) \leq 1/2$. The mean-value integral and (41), first for the grid-point gradient and then along the segment, imply

$$\|e(t)\| \leq Ch\rho(t)^{2N} e^{-2a\rho(t)}. \tag{43}$$

Constants include m and the bounded comparison of neighboring clocks. Changing variables in (42) yields the exact almost-everywhere equation

$$w_\sigma = -p + E(\sigma), \qquad E = e^\rho \sqrt{\rho}\, e(t), \qquad \|E\| \leq E_h(\rho) := Ch\rho^{2N+1/2} e^{-(2a-1)\rho}. \tag{44}$$

Indeed multiplication by $(t_0 + t)\sqrt{\rho}$ produces $w_\sigma + w/2$, and $w/2$ is the norm term in p. Since $2a - 1 > 0$, both tails

$$\int_{\rho_0}^{\infty} E_h(\rho) \frac{d\rho}{\rho}, \qquad \int_{\rho_0}^{\infty} E_h(\rho)^2 \frac{d\rho}{\rho} \tag{45}$$

are finite and tend to zero as $\rho_0 \to \infty$, for each fixed h. The pointwise tail supremum tends to zero as well. Both small physical derivatives are essential: using only a small gradient would not produce this integrable forcing estimate.

F.3 Energy, angular length, and exclusion of exit

The interpolation is locally absolutely continuous, so the energy chain rule holds almost everywhere, including across its grid partition. Equations (39) and (44) give

$$\frac{dW}{d\sigma} \leq -\tfrac{1}{2} \|p\|^2 + \psi + p^T E \leq -\tfrac{1}{4} \|p\|^2 + \psi + E_h^2. \tag{46}$$

Let $e_0 = W(\rho_0, w_0) - F_*$ and define

$$B_0 = e_0 + \sup_{\rho \geq \rho_0} \zeta(\rho) + \int_{\rho_0}^{\infty} [\psi(\rho) + E_h(\rho)^2] \frac{d\rho}{\rho}.$$

The supremum is needed; no monotonicity of ζ is assumed. Integration before a first exit yields

$$\int \|p\|^2 \, d\sigma \leq 4B_0, \qquad \int d_{\mathcal{M}}(w)^2 d\sigma \leq 2C^2 \left[4B_0 + \int_{\rho_0}^{\infty} r(\rho)^2 \frac{d\rho}{\rho} \right]. \tag{47}$$

The upper-energy inequality first fixes a bounded sublevel on which the lower-energy estimate is used. No a priori lower bound along an uncontrolled trajectory is needed.

The projection's bounded derivative applied to E modifies (40) only by

$$\left\| \frac{d\pi(w)}{d\sigma} \right\| \leq C d_{\mathcal{M}}(w) \|p\| + C\eta_R(\rho) + CE_h(\rho). \tag{48}$$

Cauchy–Schwarz and (47) bound the total angular length by a quantity tending to zero with e_0 and the tails of the displayed integrals. Choose an interior phase seed and make this length smaller than its gap to the larger cap. This excludes angular exit.

For clarity, the normal-boundary gap applies as follows. Bounded energy and positive retained penalties imply $g_i \geq 1 - C \log \rho/\rho$. Their positive homogeneous rates admit a radial correction of size $O(\log \rho/\rho)$ making every constraint feasible. Constrained quadratic growth then gives $W \geq F_* + d_0 - o(1)$ at the fixed normal boundary, for $d_0 > 0$. Choose ρ_0 so that this lower bound is at least $F_* + 3d_0/4$, while (46) keeps $W < F_* + d_0/2$. Normal exit is impossible.

The original strict radial seed

$$w_0 = (1 + K \log \rho_0/\rho_0) w_*, \qquad K > 1,$$

has energy excess tending to zero. The extra forcing budgets (45) also tend to zero for fixed h. Thus all conditions can be met simultaneously. They remain strict on a small open ball of full physical dimension after multiplication by $\sqrt{\rho_0}$. This establishes capture of the actual interpolated iterates; it does not presume shadowing of one specified ODE trajectory.

F.4 INDIVIDUAL-PROFILE CONVERGENCE

Equation (48) gives finite total phase length, so $\pi(w(\sigma))$ converges. Normal convergence requires an additional argument because no uniform rescaled Hessian bound was assumed. Fix a small $\delta > 0$. On every sufficiently late annulus $\delta/2 \le d_{\mathcal{M}}(w) \le \delta$, the normal estimate and $r \to 0$ give $\|p\| \ge c\delta$. Since $E_h \to 0$, (44) implies $\|w_\sigma\| \le 2\|p\|$ there. A crossing from normal distance $\delta/2$ to δ therefore costs

$$\delta/2 \le \int \|w_\sigma\| \, d\sigma \le \frac{2}{c\delta} \int \|p\|^2 \, d\sigma.$$

Every crossing consumes a fixed positive multiple of δ^2 from the finite gradient budget. Meanwhile finite $\int d_{\mathcal{M}}^2 d\sigma$ forces arbitrarily late returns below $\delta/2$. Infinitely many excursions are impossible, proving $d_{\mathcal{M}}(w) \to 0$. Together with phase convergence this gives an individual profile limit, including on the grid subsequence. Since $\log(t_0 + nh)/\log n \to 1$, the alternative normalization by $\sqrt{\log n}$ has the same limit.

F.5 SIGNS, RISK, AND MONOTONE DISCRETE LOSS

For a sufficiently large common starting clock, the captured tube has uniformly correct learned-margin signs and strict wrong-side tail dominance on the two missed examples. These hold throughout the interpolant and hence at every iterate. Learned margins tend to infinity and missed logits tend to zero from above for their negative labels. Thus loss tends to L_{bad} and error equals P_{bad}.

Let M_n bound the physical loss Hessian along the entire captured step segment. By (41), increasing ρ_0 ensures $hmM_n \le 1$ for all n. Taylor's integral remainder yields

$$\begin{aligned} L(\Theta_{n+1}) &\le L(\Theta_n) - hm(1 - hmM_n/2) \|\nabla L(\Theta_n)\|^2 \\ &\le L(\Theta_n) - (hm/2) \|\nabla L(\Theta_n)\|^2 . \end{aligned} \tag{49}$$

Every finite iterate is well-defined by the smooth update map; the tube bound also controls its norm at each finite index. No ODE continuation assertion is substituted for this discrete fact. This proves Theorem 2.

G CONSTANT-STEP SAMPLED SGD IN THE GAUSSIAN BAD BASIN

We prove Theorem 3. The samples are independent with the original population law p, and

$$\ell_x(\Theta) = 2\log(1 + e^{-y_x f_\Theta(x)}), \qquad L = \sum_x p_x \ell_x.$$

In particular the factor two in the stochastic loss agrees with the population-loss convention. The update trains every coordinate in the original Euclidean metric. Apply the fixed isometry from the original chart with metric D to physical coordinates, and write $\mathcal{M}$ for the isometric image of its profile manifold. We return to the original chart only in the final limit statement. This transformation changes neither the optimizer nor its layer learning rates.

G.1 POINTWISE ESTIMATES ON TWO FIXED TUBES

Choose nested compact positive phase caps and a sufficiently small outer normal tube $\mathcal{K}_{\mathrm{out}}$. Every retained learned-margin and harmful-tail rate equals one on $\mathcal{M}$, so all these rates can be made larger than $19/20$ on this tube. The discarded rates retain their strict gaps above one. Gate signs and the original wrong-side tail dominance are uniform. Choose a smaller capture region $\mathcal{K}_{\mathrm{in}}$ whose closure is inside this outer tube. Its normal radius is $r_* > 0$; its phase boundary is separated from the initial phase cap. Both boundaries have fixed room before the boundary of $\mathcal{K}_{\mathrm{out}}$.

For sufficiently large ρ, define

$$\begin{gathered} \sigma = \log \rho, \qquad W(\rho, w) = me^{\rho}[L(\sqrt{\rho} w) - L_{\mathrm{bad}}] + \|w\|^2/4, \qquad p = \nabla_w W, \\ d(w) = \operatorname{dist}(w, \mathcal{M}), \qquad F_* = \|w_*\|^2/4. \end{gathered} \tag{50}$$

The exact Gaussian expansion and geometry in Appendices A–C give pointwise estimates on the outer tube:

$$\partial_\sigma W \le \tfrac{1}{2}\|p\|^2 + \psi(\rho), \qquad d(w) \le C[\|p\| + b(\rho)], \tag{51}$$

$$\|D\pi(w)p\| \le Cd(w)\,\|p\| + C\eta_R(\rho), \qquad W(\rho, w) \ge F_* - \zeta(\rho). \tag{52}$$

Here π is the smooth nearest-profile projection, with bounded first and second derivatives on these tubes. The energy lower bound is used on a fixed bounded-energy sublevel. Its use below follows, rather than precedes, an upper-energy bound. Positive majorants may be chosen so that, for fixed $C, N, c > 0$,

$$\eta_R = C\rho^N e^{-c\rho}, \qquad b, \psi, \zeta \le C[\log\rho/\rho + \eta_R], \qquad \zeta \longrightarrow 0. \tag{53}$$

Thus ψ, b^2, η_R have integrable tails against $d\rho/\rho$. We use eventually decreasing explicit upper envelopes of the form (53) when comparing sampled sums to integrals; mere integrability of an arbitrary envelope would not suffice.

Constrained normal quadratic growth also supplies constants $g_*, \varepsilon_* > 0$ such that, for large ρ, on that bounded-energy region with phase in the outer cap,

$$r_* \le d(w) \le r_* + \varepsilon_* \quad \Longrightarrow \quad W(\rho, w) \ge F_* + g_*. \tag{54}$$

The band accommodates discrete overshoot. To obtain it, bounded energy and the positive retained penalties force their rates to be at least $1 - C\log\rho/\rho$. A radial correction of that vanishing size makes them feasible. The constrained objective has a uniformly positive excess at normal distance at least r_*, and the remainder vanishes uniformly. Choose ε_* small enough that this band and a slightly enlarged phase cap remain in the outer tube. These are pointwise facts; no stochastic capture has been assumed.

G.2 Physical and joint-clock derivatives

Fix $a = 9/10$. Uniformly on $\mathcal{K}_{\mathrm{out}}$, for each example and all sufficiently large ρ,

$$\begin{aligned} \|\nabla_\Theta \ell_x(\sqrt{\rho}w)\| + \left\|\nabla^2_\Theta \ell_x(\sqrt{\rho}w)\right\| &\le C\rho^N e^{-a\rho}, \\ |L(\sqrt{\rho}w) - L_{\mathrm{bad}}| &\le C\rho^N e^{-a\rho}. \end{aligned} \tag{55}$$

For a learned example, its signed margin has rate above $19/20$; its logistic residual is exponentially small and its first two physical logit derivatives are at most polynomially large. For a missed example, every gate is negative; its logit and first two physical derivatives are polynomial factors times Gaussian tails whose rates exceed $19/20$. The missed logistic residual need not vanish. The chain rule, compactness, fixed positive τ, and the slack from $19/20$ to a prove (55).

Put $T = e^\rho$. Differentiating the exact energy then gives

$$\begin{aligned} \|p\| &\le C\rho^N T^{1-a}, \\ \left\|\nabla^2_w W\right\| + \|\partial_\sigma \nabla_w W\| + |\partial^2_\sigma W| &\le C\rho^N T^{1-a}. \end{aligned} \tag{56}$$

For example $\nabla^2_w W = mT\rho\nabla^2_\Theta L + I/2$. The clock derivatives of $L(\sqrt{\rho}w)$ involve only $L - L_{\mathrm{bad}}$, its first two physical derivatives, and polynomial factors in ρ. This is a joint second-derivative bound in (σ, w); the rescaled Hessian is not asserted uniformly bounded. Constants may be enlarged, and the finite polynomial exponent N may increase between estimates.

G.3 Exact recursion, noise, and the exit time

For a deterministic $t_0 > 1$ to be chosen, set

$$\begin{aligned} T_n &= t_0 + nh, \quad \rho_n = \log T_n, \quad \sigma_n = \log\rho_n, \quad \Delta_n = \sigma_{n+1} - \sigma_n, \\ w_n &= \Theta_n/\sqrt{\rho_n}, \qquad p_n = p(\rho_n, w_n), \qquad \mathcal{F}_n = \sigma(X_0, \dots, X_{n-1}). \end{aligned} \tag{57}$$

For $t_0 \ge 2h$ and large ρ_0, neighboring clocks are uniformly comparable, $\Delta_n \asymp h/(T_n\rho_n)$, and $\sum_n \Delta_n = \infty$. Define

$$\xi_n = -\frac{hm}{\sqrt{\rho_{n+1}}}[\nabla\ell_{X_n}(\Theta_n) - \nabla L(\Theta_n)]. \tag{58}$$

Then $\mathbb{E}[\xi_n \mid \mathcal{F}_n] = 0$ exactly. Normalizing the actual update gives

$$w_{n+1} - w_n = -\Delta_n p_n + \xi_n + r_n, \tag{59}$$

$$\|r_n\| \le Ch^2\rho_n^N[T_n^{-2} + T_n^{-1-a}], \qquad \|\xi_n\| \le Ch\rho_n^N T_n^{-a}. \tag{60}$$

The remainder r_n is $\mathcal{F}_n$-measurable. For verification, put $c_n = \sqrt{\rho_n/\rho_{n+1}} = e^{-\Delta_n/2}$. The exact mean increment is $(c_n - 1)w_n - hm\nabla L(\Theta_n)/\sqrt{\rho_{n+1}}$. After adding $\Delta_n p_n$, its two errors are

$$(c_n - 1 + \Delta_n/2)w_n, \qquad m\left(\Delta_n T_n\sqrt{\rho_n} - \frac{h}{\sqrt{\rho_{n+1}}}\right)\nabla L(\Theta_n).$$

The first is $O(h^2T_n^{-2}\rho_n^{-2})$; the coefficient in parentheses is $O(h^2T_n^{-1}\rho_n^{-1/2})$. Equation (55) proves (60). No sampling error is included in the deterministic clock remainder.

Let $\kappa = \inf\{n : w_n \notin \mathcal{K}_{\text{in}}\}$, with $\kappa = \infty$ if this set is empty. This is a stopping time for $\mathcal{F}_n$, so $\mathbf{1}_{\{n<\kappa\}}$ is predictable. We first use every estimate only before exit and for its possible endpoint. Directly from the physical update,

$$\|w_{n+1} - w_n\| \le Ch\rho_n^N T_n^{-a}. \tag{61}$$

This tends to zero uniformly. Taking t_0 sufficiently large makes every possible one-step segment in joint (σ, w) coordinates from the inner region stay in $\mathcal{K}_{\text{out}}$. Any normal overshoot is less than ε_*. Taylor's theorem therefore uses only a checked outer tube even for the step causing first exit; it does not assume invariance of the inner region.

G.4 Energy Taylor expansion and deterministic tails

Write $W_n = W(\rho_n, w_n)$. Joint Taylor expansion, using (51), (56), and (59), yields for $n < \kappa$

$$W_{n+1} - W_n \le -\tfrac{1}{2}\Delta_n \|p_n\|^2 + \Delta_n\psi(\rho_n) + p_n^T\xi_n + e_n, \tag{62}$$

$$0 \le e_n \le Ch^2\rho_n^N[T_n^{1-3a} + T_n^{-1-a} + T_n^{-2a}].$$

The three powers respectively bound the spatial, time, and mixed second-order Taylor errors. Indeed the joint Hessian has size $\mathrm{poly}(\rho)T^{1-a}$, the state increment has size $h\,\mathrm{poly}(\rho)T^{-a}$, and the clock increment has size $O(h/(T\rho))$. The term $|p_n^T r_n|$ is covered by the latter two powers. These are deterministic upper bounds on the stopped domain, even when the actual Taylor remainder is random.

Since $a = 9/10$, all relevant tails below tend to zero as $t_0 \to \infty$, for each fixed h:

$$E_* = \sum_{n\ge 0} e_n, \qquad R_* = \sum_{n\ge 0} Ch^2\rho_n^N[T_n^{-2} + T_n^{-1-a}], \tag{63}$$

$$J_*^{\text{step}} = \sum_{n\ge 0} Ch^2\rho_n^N T_n^{-2a}, \qquad \Psi_* = \sum_{n\ge 0}\Delta_n\psi(\rho_n),$$

$$D_* = \sum_{n\ge 0}\Delta_n b(\rho_n)^2, \qquad A_* = \sum_{n\ge 0}\Delta_n\eta_R(\rho_n).$$

Here E_* denotes the sum of the displayed deterministic bounds for e_n. For any fixed $b_1 > 1$, a sum of $h^2(\log T_n)^N T_n^{-b_1}$ is bounded by a constant times its first term plus $h\int_{t_0}^{\infty}(\log t)^N t^{-b_1}\,dt$. This proves the first three assertions because $1 - 3a < -1$ and $-2a < -1$. The last three follow from the explicit eventually decreasing majorants in (53) and the clock partition. In particular J_*^{step} bounds the sum of squared step sizes and the corresponding projection Taylor errors.

G.5 Three stopped martingales

Define the scalar energy martingale and two vector martingales

$$M_N = \sum_{n<N}\mathbf{1}_{\{n<\kappa\}}p_n^T\xi_n, \tag{64}$$

$$Q_N = \sum_{n<N}\mathbf{1}_{\{n<\kappa\}}D\pi(w_n)\xi_n, \qquad Z_N = \sum_{n<N}\mathbf{1}_{\{n<\kappa\}}\xi_n.$$

Their increments are centered conditionally on $\mathcal{F}_n$. The whole-tube bounds imply deterministic bounds on their predictable quadratic variations:

$$\sum_n \mathbb{E}[(M_{n+1}-M_n)^2 \mid \mathcal{F}_n] \le V_E(t_0) := \sum_n Ch^2\rho_n^N T_n^{2-4a}, \tag{65}$$

$$\sum_n \mathbb{E}[\|Q_{n+1}-Q_n\|^2 \mid \mathcal{F}_n], \quad \sum_n \mathbb{E}[\|Z_{n+1}-Z_n\|^2 \mid \mathcal{F}_n] \le V_P(t_0) := \sum_n Ch^2\rho_n^N T_n^{-2a}.$$

Both right sides tend to zero with t_0. The energy estimate needs $2-4a<-1$, namely $a>3/4$; square summability of the raw sampling noise alone would not prove it. Orthogonality of martingale increments and the L^2 maximal inequality give, for $\varepsilon_E,\varepsilon_\pi>0$,

$$\mathbb{P}\Big(\sup_N |M_N|>\varepsilon_E\Big) \le \frac{4V_E(t_0)}{\varepsilon_E^2}, \qquad \mathbb{P}\Big(\sup_N \|Q_N\|>\varepsilon_\pi\Big) \le \frac{4V_P(t_0)}{\varepsilon_\pi^2}. \tag{66}$$

The vector bound follows by applying the scalar maximal theorem to its norm submartingale. All three martingales converge almost surely and in L^2, since their second moments are uniformly bounded. Let $\mathcal{E}$ be the event that these three converge and

$$\sup_N |M_N| \le \varepsilon_E, \qquad \sup_N \|Q_N\| \le \varepsilon_\pi.$$

For fixed positive $\varepsilon_E,\varepsilon_\pi$, taking t_0 large makes $\mathbb{P}(\mathcal{E})\ge 1-\delta$ by (66). No finite-variation property is asserted for any martingale. In particular the raw noise is not assumed absolutely summable; convergence of Z_N will instead control its sum over late normal-annulus crossings.

G.6 Energy and phase budgets exclude exit

Stop both clock and state in the energy: the endpoint is $W_{N\wedge\kappa}$, not the energy at an advancing clock with a frozen state. Summing (62) gives on $\mathcal{E}$

$$W_{N\wedge\kappa} + \tfrac{1}{2}\sum_{n<N\wedge\kappa} \Delta_n \|p_n\|^2 \le W_0 + \Psi_* + E_* + \varepsilon_E. \tag{67}$$

Assume $W_0 \le F_* + e_0$ with a prescribed small $e_0>0$. Choose the upper bound to be less than $F_* + g_*/2$ and within the fixed energy ceiling used for (52). This upper bound first verifies membership in that sublevel, so applying the energy floor is noncircular. Set

$$\zeta_* := \sup_{\rho\ge\rho_0} \zeta(\rho), \qquad B = e_0 + \Psi_* + E_* + \varepsilon_E + \zeta_*.$$

No monotonicity of ζ is assumed. Uniformly in stopped horizons,

$$\sum \Delta_n \|p_n\|^2 \le 2B, \qquad \sum \Delta_n d(w_n)^2 \le 2C^2(2B+D_*). \tag{68}$$

Apply spatial Taylor expansion to $\pi(w_{n+1})-\pi(w_n)$. Its predictable nonmartingale increment is $-\Delta_n D\pi(w_n)p_n + D\pi(w_n)r_n$. The quadratic Taylor error can depend on X_n; its norm is absolutely summable pathwise by (61). Together these nonmartingale terms have total variation at most

$$V_{\mathrm{phase}} \le C\sqrt{(2B)\,2C^2(2B+D_*)} + CA_* + CR_* + CJ_*^{\mathrm{step}}. \tag{69}$$

The first term follows by Cauchy–Schwarz from $\sum \Delta_n d(w_n)\,\|p_n\|$; the remaining martingale is exactly Q_N. Hence phase displacement from the initial phase at every stopped endpoint is at most $V_{\mathrm{phase}}+\varepsilon_\pi$. Choose it smaller than the gap to the phase boundary of $\mathcal{K}_{\mathrm{in}}$. No angular exit, including a first-exit overshoot, is then possible. A normal exit would lie in the band (54) and have energy at least F_*+g_*, contrary to (67). Thus $\kappa=\infty$ on $\mathcal{E}$.

The choices have a fixed order. First choose the caps, radii, normal energy gap, and phase gap. Next take small $e_0,\varepsilon_E,\varepsilon_\pi>0$ leaving strict slack in the energy and phase budgets. Then increase t_0 until all deterministic tails and ζ_* are small, the maximal probability bound is at most δ, and the one-step outer-tube condition holds. Finally use the original radial seed

$$w_0 = (1+K\log\rho_0/\rho_0)w_*, \qquad K>1, \tag{70}$$

at an interior phase. Its normal distance vanishes and its energy tends to F_* as $\rho_0\to\infty$. After further increasing t_0, all required initial inequalities hold strictly. They also hold uniformly in a small full-dimensional open ball about this seed. Multiplication by $\sqrt{\rho_0}$ produces the deterministic physical set U. All bounds above are uniform over U.

G.7 Convergence to one random profile

On $\mathcal{E}$, (68) holds for the full sequence. The phase drift has finite total variation, its adapted Taylor errors and predictable clock remainders are absolutely summable, and Q_N converges. Therefore $\pi(w_n)$ converges; the strict phase budget keeps its limit in an interior positive cap. This uses martingale convergence, not finite martingale variation.

Finite $\sum \Delta_n d(w_n)^2$ and divergent $\sum \Delta_n$ give $\liminf_n d(w_n) = 0$. To prove a full normal limit, fix a sufficiently small $\varepsilon > 0$. On the late annulus $\varepsilon/2 \le d(w_n) \le \varepsilon$, the normal bound in (51) and $b(\rho_n) \to 0$ imply $\|p_n\| \ge c\varepsilon$. Suppose there are infinitely many excursions to distance at least ε. Choose disjoint late crossings beginning at the last node below $\varepsilon/2$ before the first subsequent node at or above ε. Remove that first crossing step. Since (61) tends to zero, the remaining interval $[j, k)$ has all its increment-start nodes in the annulus and its endpoints differ by at least $\varepsilon/2 - o(1)$ in Euclidean norm. This follows from the 1-Lipschitz property of distance to $\mathcal{M}$.

Sum (59) over this interval. On capture, the noise sum equals $Z_k - Z_j$. Almost-sure convergence of Z_N makes this uniformly small over all sufficiently late intervals; the r_n tail is absolutely summable. Thus every sufficiently late crossing obeys

$$\sum_{j \le n < k} \Delta_n \|p_n\| \ge \varepsilon/4, \qquad \sum_{j \le n < k} \Delta_n \|p_n\|^2 \ge c\varepsilon^2/4.$$

Infinitely many disjoint crossings contradict the finite budget (68). The arbitrarily late returns below $\varepsilon/2$ therefore imply $d(w_n) \to 0$. This argument does not bound a stochastic step by its mean drift, nor assume a uniformly bounded stiff Hessian or uniform continuity in the rescaled clock. With phase convergence it yields $w_n \to w_*(u_\infty)$ in the isometric coordinates. Transforming back proves the stated limit in the original chart. Since $\log(t_0 + nh)/\log n \to 1$, using $\sqrt{\log n}$ instead gives the same asymptotic profile.

G.8 Signs, loss, quantifiers, and limitations

For a sufficiently large common starting clock, the separated tube has uniform strict learned-margin signs and wrong-side tail dominance on $\mathcal{B}$. Capture gives exactly these signs at every node, including the initial one. The six learned margins diverge positively, whereas the missed logits approach zero from the wrong side, so $L(\Theta_n) \to L_{\mathrm{bad}}$ and classification error equals P_{bad} at every iteration on $\mathcal{E}$. Population loss may fluctuate; no monotone-loss statement is used or implied.

Each finite update is well defined because the activation and loss are smooth at every finite state. On the event established above, the compact normalized tube also gives $\|\Theta_n\| \le C\sqrt{\log(t_0 + nh)}$. The same deterministic U and constants work for each of its initializations, with probability at least $1 - \delta$ separately for each start. We do not assert a single common event for all uncountably many initializations. The auxiliary t_0 selects an initial scale; it is not an unobserved training interval or a time-varying modification of the constant-step optimizer.

This is an exact-arithmetic statement about a specially prepared large-state region. It provides neither a useful onset bound nor a basin probability for standard random initialization, and it does not cover the exceptional event or finite-precision underflow. Its stochastic argument depends on stopped energy and phase martingales as well as raw-noise convergence: small individual gradients or square-summable raw noise alone would not establish permanent classification failure or a definite profile limit.

H Proof of output-bias recovery

We prove Theorem 4 for the original Gaussian activation, every fixed allowed width and temperature, arbitrary fixed positive population probabilities, and fixed $\eta > 0$. All body coordinates below are physical Euclidean coordinates.

A fixed normalized ball for the comparator. Let $\mathcal{I}$ be the fixed isometry from the chart with metric D to physical coordinates, and put $q = \mathcal{I} w_*(u_0)$ at an interior positive phase. Write $F_* = \|q\|^2/4$. The proof in Appendix C chooses its phase caps, normal radius, and a positive initial energy allowance

e_0 before choosing the starting clock. Its capture conditions are a fixed inner phase cap, a fixed inner normal tube, and

$$W(\rho_0, z) < F_* + e_0, \qquad W(\rho, z) = me^{\rho}[L(\sqrt{\rho}z) - L_{\mathrm{bad}}] + \|z\|^2/4.$$

These conditions are sufficient uniformly for every sufficiently large ρ_0. Indeed its squared-gradient budget is at most $2[e_0 + \sup_{\rho \geq \rho_0} \zeta(\rho) + \int_{\rho_0}^{\infty} \eta_R(\rho) d\rho/\rho]$. Here η_R denotes the integrable Gaussian remainder envelope from that proof, distinct from the bias learning rate η. The last two terms vanish; the normal-square budget and phase-length bound then have strictly small limits after e_0 is fixed sufficiently small. The energy upper bound remains below the fixed normal barrier. Thus neither boundary requires an initial energy excess tending to zero.

Choose a small fixed $\varepsilon > 0$. The point $c = (1+\varepsilon)q$ has the same phase, small normal distance, and $\|c\|^2/4 < F_* + e_0/4$. Every retained rate there equals $(1+\varepsilon)^2 > 1$. Continuity supplies a nonempty open ball K about c whose compact closure lies in the fixed inner region and satisfies, for a fixed $\delta > 0$,

$$\|z\|^2/4 < F_* + e_0/2, \qquad h_i(z) \geq 1 + \delta \quad (z \in \overline{K}).$$

Here the homogeneous rate functions are transported by $\mathcal{I}$; the Gaussian activation itself is not asserted homogeneous. The exact positive Gaussian penalties and Mills-integral bound give

$$\sup_{z \in \overline{K}} |W(\rho, z) - \|z\|^2/4| \leq C(1 + \rho^{-1/2})e^{-\delta\rho} + \eta_R(\rho) \longrightarrow 0.$$

Consequently the same K satisfies the comparator's capture conditions for every sufficiently large $\rho_0 = s^2$. The bias-free flow therefore has the claimed permanent error set from every sz, $z \in K$. This uniformization of the existing basin does not assert a useful radius or an ordinary initialization regime.

Full-space estimates for the augmented flow. After shrinking K if needed, its six learned homogeneous margins are at least $\gamma > 0$, and for every missed example every normalized hidden preactivation is at most $-\lambda < 0$. These strict conditions hold on a slightly enlarged compact neighborhood. Fix $\Theta_0 = sz$, $z \in K$, and consider the physical ball $\|\Theta - \Theta_0\| \leq 1$. For large s, every missed preactivation is at most $-\lambda s/2$, and every parameter is bounded by $C(1+s)$. Since

$$0 < \phi_\tau(-v) \leq \tau\varphi(v/\tau), \qquad \phi_\tau'(-v) = \Phi(-v/\tau) \leq Ce^{-v^2/(2\tau^2)} \quad (v \gg 1),$$

the readout and incoming-weight derivatives give

$$\max_{x \in \mathcal{B}} \big[|f_\Theta(x)| + \|\nabla_\Theta f_\Theta(x)\|\big] \leq C(1+s)e^{-\kappa s^2}. \tag{71}$$

The factors a_j in the incoming-weight derivatives are included in the factor $1+s$; no symmetry of the perturbed body is imposed. For learned points, the global bound $|\phi_\tau(v) - [v]_+| \leq \tau/\sqrt{2\pi}$ and the local Lipschitz bound for the homogeneous predictor give

$$y_x f_\Theta(x) \geq \gamma s^2 - C(1+s), \qquad \|\nabla_\Theta f_\Theta(x)\| \leq C(1+s), \quad x \notin \mathcal{B}. \tag{72}$$

Every constant is uniform in $z \in K$ and large s.

A fixed bias interval and its first exit. Put $p_{\min} = \min_x p_x > 0$ and choose a fixed $D_* > 2$ such that

$$2P_{\mathrm{bad}} \log(1 + e^{-(D_*-1)}) < p_{\min} \log 2.$$

Set $\nu = \eta P_{\mathrm{bad}}/(1 + e^{D_*+1})$ and $T_{\mathrm{rec}} = (D_* + 2)/\nu$. Temporarily restrict to

$$\|\Theta - \Theta_0\| \leq 1, \qquad -D_* \leq b \leq 2, \qquad 0 \leq t \leq T_{\mathrm{rec}}.$$

For all sufficiently large s, (72) implies $y_x(f_\Theta(x) + b) \geq \gamma s^2/2$ on learned points. Their loss residuals are exponentially small. On missed points the residual may be order one, but (71) bounds its body Jacobian. Therefore throughout this stopped region,

$$\left\|\dot{\Theta}\right\| \leq C(1+s)e^{-\kappa_1 s^2} =: e_s.$$

The exact bias equation is

$$\dot{b} = 2\eta \sum_x \frac{p_x y_x}{1 + e^{y_x(f_\Theta(x)+b)}}.$$

Both missed labels are -1. Enlarging the scale so their logits have absolute value at most one, their combined contribution is at most $-2\eta P_{\mathrm{bad}}/(1+e^{D_*+1})$. The learned contribution has absolute value at most $2\eta e^{-\gamma s^2/2}$. Thus $\dot{b} \le -\nu$ throughout the stopped region for sufficiently large s.

Now take s large enough that $T_{\mathrm{rec}} e_s < 1/2$, and start at any $b_0 \in (-1, 1)$. Before hitting $b = -D_*$, an upper bias exit is impossible, while a body-radius exit by time T_{rec} would contradict the displacement bound $T_{\mathrm{rec}} e_s < 1/2$. If no hit occurred by T_{rec}, integration would give $b(T_{\mathrm{rec}}) \le 1 - \nu T_{\mathrm{rec}} = -D_* - 1$, a contradiction. Hence a hit occurs at some $t_* < T_{\mathrm{rec}}$ while the body remains in the verified physical neighborhood. This is a bootstrap for the actual joint flow, not an assumed frozen-body trajectory.

Permanent signs and continuation. At the hit, the two missed margins are at least $D_* - 1$, so their total loss is below $p_{\min} \log 2$. For sufficiently large s, the learned loss is at most $2e^{-\gamma s^2/2} < p_{\min} \log 2$. Consequently

$$\mathcal{L}(\Theta(t_*), b(t_*)) < 2p_{\min} \log 2.$$

Along the exact augmented flow,

$$\frac{d\mathcal{L}}{dt} = -m \left\|\nabla_\Theta \mathcal{L}\right\|^2 - \eta |\partial_b \mathcal{L}|^2 \le 0.$$

Any later nonpositive margin would contribute at least $2p_{\min} \log 2$ by itself. All subsequent margins are therefore strictly positive, without any continued tube assumption. The bounded logistic residuals, bounded ϕ'_τ, and linear growth of ϕ_τ imply globally $\left\|\dot{\Theta}\right\| \le C(1 + \|\Theta\|)$ and $|\dot{b}| \le 2\eta$. Gronwall's inequality and local smoothness exclude finite-time blowup. These arguments prove the theorem uniformly over K and the fixed initial-bias interval.

At $b_0 = 0$, the initial predictor and its derivatives with respect to every shared coordinate agree with the bias-free model. Thus the initial shared-coordinate velocities agree exactly. The new coordinate contributes the constant-feature term $\eta \mathbf{1}\mathbf{1}^T$ to the weighted tangent kernel, which is not claimed equal between models. At nonzero initial bias, recovery remains proved but exact initial-function matching does not.

The mechanism is ordinary optimization of a useful linear feature during a finite interval of negligible body motion, followed by a loss barrier. Related finite-time linearized comparisons are established tools [15]. No new general recovery mechanism, zero limiting loss, or result for mixed-label missed cohorts is claimed. This control exposes a limitation of the specific bias-free failure construction rather than extending that failure to biased networks.

I A RESTRICTED INITIALIZATION CONSEQUENCE

Corollary 4 (An exterior bad cone)**.** *Fix the original Gaussian model and an SGD step size $h > 0$. There are a nonempty open cap $\Omega \subset \mathbb{S}^{4m-1}$ and $r_0 < \infty$ such that every bias-free flow initialized with $\Theta_0/\|\Theta_0\| \in \Omega$ and $\|\Theta_0\| \ge r_0$ has the permanent two-error and profile conclusions of Theorem 1. For sampled SGD, conditional on each such start, the corresponding conclusions hold with probability at least $1 - \varepsilon_h(\|\Theta_0\|)$, where $\varepsilon_h(r) \to 0$. Consequently, for independent Gaussian initialization $\Theta_0 = sG$, $G \sim N(0, I_{4m})$, both failure probabilities satisfy*

$$\liminf_{s\to\infty} \mathbb{P}(\textit{failure}) \ge \mathrm{Unif}(\mathbb{S}^{4m-1})(\Omega) > 0.$$

Appendix I.1 formalizes this consequence of the fixed normalized basin. It gives neither a useful numerical probability nor access at ordinary initialization scales. The cap may be extremely narrow; no width-uniform bound or common sample-stream event is claimed.

I.1 PROOF OF THE EXTERIOR-CONE CONSEQUENCE

We prove Corollary 4; all norms here are physical Euclidean norms. The assertion concerns gradient flow and the iid sampled update of Theorem 3, not an additional deterministic-GD result.

One normalized ball and fixed budgets. Use the fixed-ball construction in Appendix H. For the stochastic comparison, first choose the two tubes of Appendix G, shrinking them so retained rates exceed $19/20$. Choose its energy allowance e_0 and martingale thresholds $u, v > 0$ before the starting clock, with strict slack in its energy and phase inequalities. They may also satisfy the original flow's budgets. A small fixed radial inflation of an interior profile then has objective excess below $e_0/4$ and every retained rate strictly above one. A fixed ball $K = B(c, a)$ with compact closure preserves the inner phase/normal conditions, objective excess below $e_0/2$, and a common strictly positive rate gap. The positive Gaussian penalties vanish uniformly on this ball as the clock grows. Hence the same K satisfies the initial energy condition for every sufficiently large starting clock.

The original flow proof therefore captures every point of this ball at all such clocks. For SGD, the deterministic error, projection, and clock tails in (63), the future supremum of the energy-floor error, and the one-step overshoot all vanish uniformly with the starting clock. Thus the fixed strict budgets remain valid on K. The whole-tube variance bounds in (65) do not depend on the selected start. For sufficiently large L, set

$$q_h(L) = \sup_{\rho_0 \geq L} \left\{ 4V_E(e^{\rho_0})/u^2 + 4V_P(e^{\rho_0})/v^2 \right\} \longrightarrow 0.$$

The stopped maximal inequalities give capture probability at least $1 - q_h(L)$ for each deterministic start in K at any $\rho_0 \geq L$. Martingale convergence has probability one; the original normal-annulus and phase arguments then give a single limiting profile and permanent signs. This is uniform per-start confidence, not a common event over uncountably many starts.

The analysis clock supplies the cone. Put $c_* = \|c\| > 0$, shrink $a < c_*$ if necessary, and define

$$\Omega = \{\omega \in \mathbb{S}^{4m-1} : \|c_*\omega - c\| < a/2\}.$$

This open cap has positive spherical measure. For any physical initialization with direction $\omega \in \Omega$ and norm r, choose only the analytical starting clock $\rho_0 = (r/c_*)^2$, $t_0 = e^{\rho_0}$. Its normalized state is $c_*\omega \in K$. Thus sufficiently large r gives the original flow capture, and the stochastic conditional probability is at least $1 - q_h((r/c_*)^2)$. This proves the first assertions, with $\varepsilon_h(r) = q_h((r/c_*)^2)$. The clock does not alter either optimizer or require a hidden training interval. Each finite per-start shift gives the same asymptotic normalization by $\sqrt{\log t}$ for flow and $\sqrt{\log j}$ for SGD. The exterior cone itself is not claimed to be an invariant set.

Gaussian probability and the order of limits. Let A_Ω be the cap's spherical measure. Gaussian radius and direction are independent. For any fixed sufficiently large L, initialization $\Theta_0 = sG$ therefore gives

$$\mathbb{P}(\text{SGD failure}) \geq [1 - q_h(L)] A_\Omega \mathbb{P}\{\|G\| \geq c_*\sqrt{L}/s\}.$$

First let $s \to \infty$ with L fixed; the radial probability tends to one. Then let $L \to \infty$. The claimed lower bound follows. The flow argument is identical without q_h once its fixed clock threshold is met. Initialization and sampling are independent.

This formalizes a large-variance consequence of the existing confinement estimates. The cap and onset remain model-dependent and potentially unfavorable. It is not an ordinary-scale entrance theorem, a numerical probability estimate, or a new optimization mechanism.