A THRESHOLDED RELU UNIT WITH FIXED OUTPUT WEIGHT GOES DORMANT UNDER TARGET DRIFT: PHASE DIAGRAM AND ENTRY TIME

Anonymous authors
Paper under double-blind review

ABSTRACT

Networks tracking moving targets lose ReLU units to dormancy: units active on almost no input get almost no gradient. Lacking a target-speed parameter, existing dormancy theory explains why such units stay dormant, not when they become dormant. We study one ReLU unit with a learned threshold and fixed output weight, trained online on a teacher whose weight vector random-walks on the sphere at fixed speed. In the large-dimension limit, its weight length, teacher overlap and threshold obey three differential equations with closed-form resting states. We map the boundary between tracking and dormancy over drift speed and learning rate: at slow drift the tracking state collides with an unstable state and both vanish; at fast drift it survives but is dormant. With nothing fitted, the equations predict where and when dormancy sets in, and simulations match once finite-dimension fluctuations are kept. These equations give the mechanism: drift keeps an error alive, at rates causing dormancy each update's squared length inflates the weight vector, so the threshold must outgrow the weight length. Since this inflation vanishes with the learning rate, we prove an isolated unit in the large-dimension equations needs a rate bounded away from zero for dormancy. Yet one of two units sharing one output error goes dormant where the isolated unit never does. In both settings a trained output weight settles at a small scale and prevents dormancy everywhere tested.

1 INTRODUCTION

A network trained by reinforcement learning tracks a target that moves as its own policy improves, and a network trained on a stream of tasks tracks a target that changes with every task (Dohare et al., 2024). A ReLU unit outputs the positive part of a weighted sum of its inputs minus a threshold, so on the inputs below the threshold it outputs zero and receives no gradient. A unit is dormant when the fraction of inputs on which it is active collapses to nearly zero. Many units going dormant this way cost a network its plasticity, its ability to keep learning. Dormant units are one of the best-documented forms of this loss (Sokar et al., 2023; Dohare et al., 2024).

Existing theory explains why a dormant unit stays dormant and what drives units toward that state (Joudaki et al., 2026). Liu & Mou (2026) show through the geometry of loss basins that plasticity is preserved when the task changes slowly enough, but others carry no parameter for the speed of the target (Wu et al., 2026; Chen et al., 2023). None of this work derives how the dormancy of a unit depends on that speed and on the learning rate. Theory therefore leaves a practitioner who sees dormant units to guess whether a smaller learning rate or a slower task schedule would have prevented them.

We answer that question in the smallest setting in which a unit can go dormant under drift, and Section 5.4 marks how far the mechanism reaches. We train a single ReLU unit with a learned threshold by online stochastic gradient descent, one input per step, to match a teacher unit whose weight vector random-walks at a fixed speed set by the environment rather than by the unit. Three numbers then describe the unit: the length of its weight vector, its overlap with the teacher, and its threshold. As the input dimension grows they obey three deterministic differential equations.

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

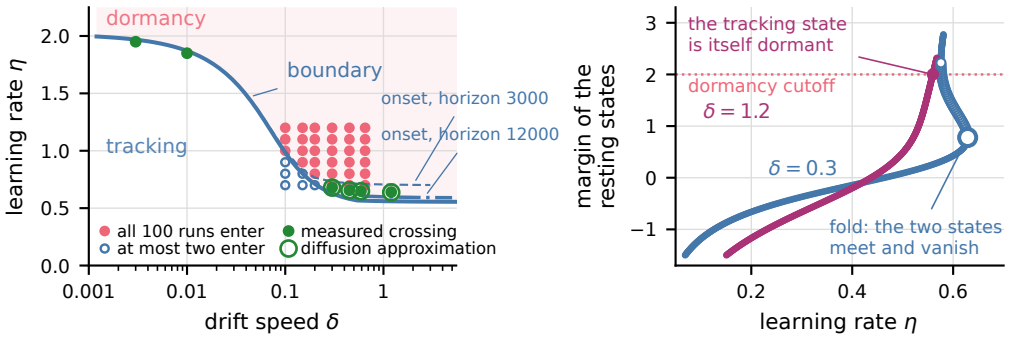


Figure 1: In the equations a unit converged on its teacher goes dormant above the boundary once the teacher drifts, and keeps tracking below it. Left: the boundary, cells of 100 runs at $d = 500$, horizon 3000 (12000 at the two slowest speeds). Green dots are the measured crossing, where half the runs went dormant; green rings the diffusion approximation, the equations with fluctuations kept. The two labelled onset curves are where the equations alone reach the cutoff within their horizon. Right: margins of resting states at two drift speeds, all stable but the arm above the ringed fold at $\delta = 0.3$.

The equations also say why a moving target empties a unit instead of merely making it track worse. Because drift keeps the unit misaligned with the teacher, its error never settles to zero. Each stochastic step adds its own squared length to the squared weight length, and at the rates where the unit goes dormant, a lasting error makes the weight vector grow. A longer weight vector makes the unit overshoot the teacher on the inputs where it is active, and its only answer is a higher threshold. Balance needs the threshold to grow faster than the weight vector, so the threshold measured in weight lengths keeps rising. We call that ratio the margin, and it fixes the fraction of inputs the unit is active on. The unit is dormant once its margin leaves it active on 2.3% of its inputs. Two things can end tracking: the resting state the unit would settle into can disappear, or it can remain but be dormant itself.

Question 1: at which combinations of drift speed and learning rate does a unit that tracks its target go dormant, and what shape does that boundary have? Question 2: on the dormant side of the boundary, how long does entry take? Question 3: does dormancy under drift need a learning rate bounded away from zero, and does the account extend from one unit to a layer of many?

The contributions of this paper are as follows. (1) We chart the boundary between tracking and dormancy from a closed form for the resting states of the three equations, locating its fold numerically (Figure 1). At slow drift it is a fold, whose learning rate rises as the drift slows; above $\delta = 0.546$ it is the curve on which the tracking state is itself dormant, and the boundary we computed never falls below $\eta = 0.554$ (Section 4.4). Simulated units cross into dormancy within 0.003 in η of the diffusion approximation (7), the same equations with their finite-dimension fluctuations kept (Section 5.2). (2) We derive the entry time by integrating those equations from the trained state, with no parameter fitted to the measurements. On a grid whose pass criteria were fixed before the first run, the predicted times order the cells almost as the measurements do, and keeping the fluctuations cuts their median error from 56% to 1.3% (Section 5.3). (3) We prove that dormancy in the isolated unit needs a learning rate bounded away from zero, since the term that inflates the weight vector is the one that vanishes with the learning rate (Proposition A.14, proved for the order-parameter equations as $\eta \to 0$, with the reduced trajectory enclosed by interval arithmetic). Two units sharing one output error enter where the isolated unit never does, while a trained output weight settles at a small output scale and prevents entry in both at every cell tested. A layer of 128 units loses more units at larger rates within its training budget, but its dormant units shrink their weight vectors instead of inflating them, so the diagram mispredicts which of its settings go dormant (Section 5.4). Section 4 sets up the unit and derives the boundary and the entry time, and Section 5 tests them.

2 Related Work

Dormant units and the loss of plasticity, the phenomenon behind all three questions, are documented empirically. Questions 1 and 2 belong to the framework of online learning of a drifting rule, which has never contained a unit that can go dormant. Question 3 is closest to the mathematics of dying ReLU units under stochastic gradient descent, which has so far treated fixed targets only.

Plasticity loss and dormant units. This line documents the phenomenon across deep reinforcement learning and continual supervised learning, and explains why a dormant unit stays dormant. Dormant-unit fractions grow under non-stationary targets and recede under fixed ones (Sokar et al., 2023), and they persist across architectures and most remedies (Dohare et al., 2024). Lyle et al. (2025) decompose plasticity loss into several independent mechanisms and find that intervening on more than one at once is what preserves plasticity. On the theory side, Joudaki et al. (2026) prove that persistently saturated configurations form invariant sets of gradient descent, sets that gradient descent never leaves. Persistence is therefore their premise, and they trace the emergence of such units to a drive toward high rank and flat loss. Wu et al. (2026) derive how gradients decay with the number of episodes. Others tie the loss to a rank decrease of the neural tangent kernel that accompanies churn (Tang et al., 2025), to a collapse of the Hessian spectrum at new-task initialization (Prakash et al., 2026), or to abrupt shifts of the loss landscape at task transitions (Liu & Mou, 2026). None of these yields the time at which a unit becomes dormant as a function of drift speed and learning rate; our second contribution does, for the isolated unit.

Statistical physics of online learning under drift. This line, including Biehl & Schwarze (1993) and Kinouchi & Caticha (1993), has analyzed a drifting rule in the framework of Section 3.1. After the two-layer ReLU network of Straat et al. (2022) came a Hermite-expansion analysis of drift for arbitrary activations (Citton et al., 2025b) and the fixed-teacher ReLU network at any learning rate (Citton et al., 2025a). Because the hidden units in this line have no thresholds, their pre-activations are centered by construction and dormancy is structurally impossible, a simplification those authors state (Straat et al., 2022; Straat, 2022). Straat et al. (2022) do find a drift strength beyond which their threshold-free units stop specializing. Our fold is instead where the tracking state disappears, after which a unit with a threshold stops responding. Continual-learning analyses with switching fixed teachers (Lee et al., 2021; Mori et al., 2025; Hiratani, 2024) and studies of nonstationarity in wide or practical networks (Graldi et al., 2025; Lyle et al., 2026; Wang et al., 2026) ask when a network forgets or adapts, not when a single unit goes dormant. Thresholds enter this framework only for a fixed rule, either carried along with the weights at their own learning rate (West et al., 1997) or equilibrated at every step as a fast bias (Endres & Riegler, 1999). We use the first with a moving teacher. West et al. (1997) find that thresholds initialized beyond a critical value trap the student in suboptimal fixed points (a transition sharper at larger learning rates) and that an output gain shared by teacher and student rescales the learning rates by its square. Our first contribution is a drifting framework in which a unit can go dormant at all, with the dormancy boundary charted from closed-form resting states.

Attracting subnetworks and dying ReLU units. This line asks when gradient noise attracts stochastic gradient descent toward simpler subnetworks it never leaves again, and the line has so far kept the target fixed. Chen et al. (2023) give explicit learning-rate and data-alignment thresholds for the attraction toward sets of weights on which a neuron outputs zero, but leave the ReLU case at issue here to future work. Ziyin (2024) shows that under a rescaling symmetry a large enough weight decay or gradient noise makes the solution with dead parameters attractive on a stationary task, and attributes the loss of plasticity to that attraction. Here, however, the η^2 term of Section 4.3 produces no entry under a fixed teacher at the rates where drift produces it; under a fixed teacher with label noise the unit enters only near $\eta = 2$ (Section 5.2). In a simplified two-layer model, Ahn et al. (2023) prove a sharp step-size transition below which gradient descent fails to learn a threshold, through edge-of-stability oscillations at a fixed target. Here the threshold grows instead through the η^2 term while the target moves. A separate line analyzes dying ReLU at initialization through probability bounds in depth and width (Lu et al., 2020), and during training through target-scale and momentum effects in a single unit with a fixed target (Arnekvist et al., 2020). We locate a transition produced by drift together with the η^2 term, and because that term vanishes with the learning rate, our third contribution asks whether a finite rate is necessary.

3 Preliminaries

3.1 Online learning of a drifting teacher

In the teacher-student setting of online learning, a student unit is trained on one input per step, an input $\xi \in \mathbb{R}^d$ with independent standard Gaussian components. Its target is the output of a teacher unit with weight vector B on the same input (Saad & Solla, 1995; Biehl & Schwarze, 1995). The unit

with weight vector w is described by two summary statistics called order parameters: the squared length $Q = \|w\|^2$ and the overlap $R = w \cdot B$. They are the few averaged quantities whose rates of change depend on themselves alone in the large-d limit. The pre-activations $x = w \cdot \xi$ and $y = B \cdot \xi$ are jointly Gaussian with variances Q and $\|B\|^2$ and covariance R. Each step changes Q and R by an amount of order $1/d$, so for large d their random parts average out over the d steps that make one unit of time. They therefore obey deterministic differential equations whose right-hand sides are averages of the student error over the Gaussian inputs (Saad & Solla, 1995). For ReLU units these averages reduce to the standard Gaussian density, its distribution function, and Owen's T function (Owen, 1956).

A drifting teacher is one whose weight vector performs a random walk on the sphere. After every step it rotates by a fixed angle toward a random direction perpendicular to itself, giving consecutive teacher vectors a fixed inner product (Biehl & Schwarze, 1993; Straat et al., 2022). In the tracking regime the unit keeps moving with the teacher while its order parameters approach fixed values. Drift enters the equation of the overlap as a decay term proportional to the drift speed (Biehl & Schwarze, 1993).

3.2 Dormant units, folds, and first-passage times

A ReLU unit is dormant when its activity, the fraction of inputs on which it produces a nonzero output, falls below a small cutoff. On inactive inputs it receives no gradient, so a unit with zero activation on every input cannot recover by gradient descent alone (Sokar et al., 2023). The dormancy score of Sokar et al. (2023) normalizes the mean post-activation magnitude of a unit by the layer mean, and the appendix reports it beside the layer activity.

Two standard facts about differential equations tell us where such a state begins and how long it takes to arrive. In a saddle-node bifurcation a stable fixed point and an unstable one approach each other as a parameter changes, collide, and disappear. At the stable point the leading Jacobian eigenvalue $\lambda_{\max}$ vanishes as the square root of the distance to it, the test we use (Strogatz, 2018). The fold is another name for the same event. We call a fixed point of such equations a resting state. A first-passage time is the first time at which a trajectory crosses a prescribed threshold. Where the deterministic flow has no path across a threshold, small noise can still push a trajectory across (Freidlin & Wentzell, 2012).

4 Methodology

4.1 A thresholded ReLU unit tracking a drifting teacher

Teacher and unit. The target is a single ReLU teacher unit $g(B \cdot \xi)$ with $g(z) = \max(0, z)$ and $\|B\| = 1$ on the Gaussian inputs of Section 3.1. The unit is a single ReLU unit with a learned threshold, $g(w \cdot \xi - \theta)$, and its output weight v is fixed at the teacher value 1 unless we say otherwise. We train it by online stochastic gradient descent with batch size one on the squared loss.

Drift. The teacher performs the isotropic random walk of Section 3.1. After every step B moves a fixed distance so that consecutive teachers overlap by $B^{(\mu+1)} \cdot B^{(\mu)} = 1 - \delta/d$, where μ is the step index and δ is the drift speed. Because the drift speed is held fixed as $d \to \infty$ and set by the environment alone, speed and learning rate are separate inputs here. The convention of Citton et al. (2025b) instead couples the displacement to the learning rate.

Update rule and rate conventions. We update weights and threshold with the same per-parameter learning rate, as when biases and weights share one optimizer step:

$$w \leftarrow w + \tfrac{\eta}{d}\,\chi\,\xi, \qquad \theta \leftarrow \theta - \tfrac{\eta}{d}\,\chi, \qquad \chi := \big(g(B \cdot \xi) - g(w \cdot \xi - \theta)\big)\,\mathbb{1}\{w \cdot \xi - \theta > 0\}, \quad (1)$$

where χ is the per-example error signal, teacher output minus unit output on the inputs where the unit is active. Figures 1 and 2 use the learning rate $\eta = \eta_{\text{raw}}\, d$, the raw step multiplied by the input dimension, so that a raw step of 10^{-3} at $d = 4000$ is $\eta = 4$. Rescaling w and θ by v shows that a unit with output weight $v > 0$ behaves exactly like one at learning rate $v^2\eta$. A threshold stepped at a fixed multiple of the rate of the weights leaves the resting states and the fold unchanged while it changes the entry times (Lemma A.15).

These two rescalings place the layer of Section 5.4 ($d = 32$) on the same axes, at the converted rate $\eta = \eta_{\text{raw}} d v^2$ that Section 5.4 tests. The layer's four raw steps convert to between 0.12 and 0.96 at $v = 1$, and training its own output weights moves it lower still; the layer panels of Figure 3 keep the raw step.

4.2 Problem formulation

Definition 1 (margin, activity, dormancy)**.** Write $m = \theta/\sqrt{Q}$ for the margin of the unit, the threshold in units of the length of its weight vector. The activity of the unit is $a = \Pr(x > \theta) = \bar{\Phi}(m)$, with $\bar{\Phi}$ the standard Gaussian upper tail. Dormancy means $a \le a_\star$, equivalently m at or above the dormancy cutoff $m_\star = \bar{\Phi}^{-1}(a_\star)$, and we use $m_\star = 2$ ($a_\star \approx 2.3\%$) throughout. Entry is the first learning time at which m reaches $m_\star$.

This is the activity criterion of Section 3.2, read for one unit rather than normalized across a layer. Our trained state is $(Q, R, \theta) = (1, 1, 0)$, a unit that has converged on a fixed teacher which then starts to move. Every trajectory starts there unless stated.

The three questions of Section 1 become the following problems. Question 1 takes a point (δ, η) of the plane of drift speed and learning rate and asks whether the unit started from the trained state keeps tracking (settling at a resting state with a margin below the cutoff) or enters dormancy; its answer over the plane is the boundary between the two. Question 2 takes a point above that boundary and asks for the entry time $T(\delta, \eta)$ of Definition 1, which we predict from the equations of Section 4.3 and compare with simulated runs at finite d by rank correlation and median relative error. Question 3 takes a scan of the learning rate at a fixed drift speed and asks whether the unit goes dormant at small rates at all. It also asks whether the mechanism and the diagram extend to the layer of Section 5.4.

Two control conditions accompany every scan, and together they separate accumulated drift from mere noise in the target. The first, drift off, holds the teacher fixed. The second, no accumulation, displaces the teacher at every step from one fixed reference by the per-step displacement of drift, so that consecutive teachers differ by about twice the squared drift step while the target never wanders away.

4.3 The order-parameter equations and why a moving teacher empties a unit

Our order parameters are Q and R of Section 3.1 and the threshold θ itself, unrescaled. Time is the learning time $\alpha = \mu/d$, the number of examples per input dimension, with μ the number of steps. Because the threshold shares the η/d learning rate, its per-step fluctuation has variance $O(d^{-2})$. Its fluctuations therefore average out like those of Q and R, and it is a third deterministic order parameter. In the limit $d \to \infty$ the triple (Q, R, θ) obeys the order-parameter equations

$$\frac{dR}{d\alpha} = \eta\, \mathbb{E}[\chi y] \;-\; \delta R, \tag{2}$$

$$\frac{dQ}{d\alpha} = 2\eta\, \mathbb{E}[\chi x] \;+\; \eta^2\, \mathbb{E}[\chi^2], \tag{3}$$

$$\frac{d\theta}{d\alpha} = -\eta\, \mathbb{E}[\chi], \tag{4}$$

with all expectations over the Gaussian pair (x, y) at the current state, and the random walk of the teacher entering only through $-\delta R$ (Lemma A.3). The η^2 term in (3) is the squared length of each step, which inflates Q. After a rescaling of time by $1/\eta$ at fixed δ/η, this term alone still carries a factor of η, which gives the diagram its learning-rate axis.

Four links, each one a term of (2)–(4), explain why a moving teacher empties a unit rather than merely making it track badly. Drift keeps the unit misaligned, because $-\delta R$ pulls the overlap down while learning pushes it up. The alignment settles below one, so the error never vanishes. At the rates where the unit enters, a persistent error inflates the weight vector through the η^2 term, the only term of (3) that cannot change sign. A longer weight vector then forces a higher threshold, because on the inputs where the unit is active it now outputs more than the teacher and the threshold update of (1) answers by raising θ until the error averages to zero again. The threshold has to rise faster than the weight vector, since a zero mean error fixes the weight length as a function of the margin

and the alignment. Because that function rises with the margin at every alignment with the teacher on our numerical grid, only a larger margin can balance a longer weight vector. Past the cutoff the gradients of the unit decay at a Gaussian rate in the margin at bounded norm (Proposition A.4), and the integrated trajectories satisfy that norm bound up to entry.

Whether the climb ends in dormancy depends on the resting states: the unit tracks if it has a tracking state whose margin is below the cutoff, and otherwise the margin runs past the cutoff. Both ways of losing such a state occur here: the resting state can collide with a second, unstable state and vanish, or it can survive with a margin that is itself past the cutoff.

Lemma 2 (closed forms). *Write $\rho = R/\sqrt{Q} \in [-1, 1]$ for the alignment and $s = \sqrt{1-\rho^2}$. The four moments $\mathbb{E}[\chi]$, $\mathbb{E}[\chi x]$, $\mathbb{E}[\chi y]$, $\mathbb{E}[\chi^2]$ and the generalization error $\epsilon_g = \frac{1}{2}\mathbb{E}[(g(w\cdot\xi-\theta) - g(B\cdot\xi))^2]$ are closed forms in (m, ρ) and the scale Q, built from the standard Gaussian density φ, distribution function Φ, its complement $\bar{\Phi}$, and Owen's T function (explicit expressions and proof in Appendix A.1).*

Every closed form of Lemma 2, and the order-parameter equations built from them, agree with quadrature, Monte Carlo, and direct simulation of (1) at $d = 4000$ within statistical error.

4.4 Where the unit goes dormant

The resting states the order-parameter equations admit are easiest to list backwards. Instead of fixing a drift speed and a learning rate and solving for the state, we fix a state and ask which pair (δ, η) makes it stationary. That direction has a closed form.

Proposition 3 (explicit parameterization). *Fix a margin m and an alignment $\rho \in (0, 1)$, and take the closed-form moments of Lemma 2 at (m, ρ). Define $M(m) = \varphi(m) - m\bar{\Phi}(m)$, the mean output of the unit over all inputs at $Q = 1$. Define $G(m, \rho) = \rho\varphi(m)\Phi(\rho m/s) + \varphi(0)\bar{\Phi}(m/s)$, the mean teacher output over all inputs, counting zero where the unit is inactive. Define Q^*, η^*, δ^* by*

$$\sqrt{Q^*} = \frac{G(m,\rho)}{M(m)}, \qquad \eta^* = -\frac{2\,\mathbb{E}[\chi x]}{\mathbb{E}[\chi^2]}, \qquad \delta^* = \frac{\eta^*\,\mathbb{E}[\chi y]}{\rho\sqrt{Q^*}}, \tag{5}$$

where the moments in the second and third expressions are evaluated at $Q = Q^*(m, \rho)$. Wherever all three are positive and finite, the state $(Q, R, \theta) = (Q^*, \rho\sqrt{Q^*}, m\sqrt{Q^*})$ is a fixed point of (2)–(4) at parameters (δ^*, η^*). Conversely, every fixed point with $\rho \in (0, 1)$ and $\eta > 0$ arises this way. We write $\Psi : (m, \rho) \mapsto (\delta^*, \eta^*)$ for the resulting map.

The first expression is the threshold balance $\mathbb{E}[\chi] = 0$ solved for Q^*, the one used in the fourth link above. The norm balance is where the finite learning rate enters, since the inflation $\eta^2\mathbb{E}[\chi^2]$ cancels the contraction $2\eta\mathbb{E}[\chi x]$ at only one learning rate per pair (m, ρ) (proof in Appendix A.4). A resting state exists at (δ, η) exactly when that pair lies in the image of Ψ. We therefore seek the edge of the part of that image with margin below the cutoff, which bounds the tracking region.

The dormancy boundary. Below $\delta = 0.546$ this edge is a fold: there the tracking state and an unstable state meet and vanish, with the saddle-node signature $\lambda_{\max}^2 \propto (\eta_c(\delta) - \eta)$ of Section 3.2 (fitted exponent 1.05 ± 0.01). We locate it by a numerical root solve of $\det D\Psi = 0$, keeping the branch that bounds the region of the trained state. Above $\delta = 0.546$ the edge is the cutoff curve $\Psi(m_\star, \cdot)$, on which the margin of the resting state equals $m_\star$; where the fold persists there, up to $\delta = 0.594$, the unit just past it still settles below the cutoff. The two pieces join into one continuous curve, the dormancy boundary of Figure 1, and we write it $\eta = \eta_c(\delta)$. Where the boundary is the fold it does not depend on the cutoff at all, since two resting states meeting is a property of the equations alone; where it is the cutoff curve it moves with the cutoff. Both pieces describe the equations at infinite dimension and without a time limit. Section 4.5 shows what a finite dimension and a finite horizon do to the crossing. Starts far from the trained state can behave differently: a start at larger Q enters below the boundary.

The boundary from slow to fast drift. Appendix A.5 proves two limits of the resting states as the margin grows. The boundary itself is computed numerically. At $\delta = 0$ the resting states other than the trained state form a branch along which, as the margin grows at interior alignment, the learning rate tends to 2. At full misalignment the resting-state learning rate tends to $2/\pi \approx 0.637$ as the margin grows, a regime the dynamics approach as $\delta \to \infty$ because (2) forces the overlap to

vanish like $1/\delta$. As $\delta \to 0^+$ the boundary rises numerically to $\eta_0 \approx 2.02$, the maximum of the learning rate along the branch at $\delta = 0$, and is monotone there (Remark A.11). As the drift grows the boundary falls and flattens at $\eta^*(m_\star, 0) = 0.554$, the learning rate at which a fully misaligned unit rests exactly at the cutoff. No state we computed at full misalignment or at fast drift rests above $2/\pi$, so this larger constant bounds those resting states rather than the boundary. The fold itself stops before the boundary flattens, at the critical end point $(\delta_{\mathrm{cp}}, \eta_{\mathrm{cp}}) = (0.5939, 0.5623)$, the fastest drift at which two states still collide. The boundary has left the fold for the cutoff curve before that, at $\delta = 0.546$; the end point is therefore a landmark of the fold rather than a point of the boundary. Past it dormancy sets in through a continuous crossover.

4.5 How long entry takes, and what a finite unit measures

Above the boundary the trained state flows deterministically to dormancy, and the entry time is the first learning time at which the margin reaches the cutoff,

$$T(\delta, \eta) = \inf\{\alpha : m(\alpha) \geq m_\star\}, \tag{6}$$

computed by integrating (2)–(4) from the trained state with nothing fitted. It diverges at the boundary.

A simulation cannot measure that boundary directly, because it runs at a finite dimension and for a finite time, and the two limits pull the crossing in opposite directions. Stopping a run early hides the slow entries just above the boundary and pushes the measured crossing up, as far as the deterministic onset, the least rate at which the equations themselves cross the cutoff within that horizon. Finite dimension works the other way, since fluctuations of order $1/\sqrt{d}$ push runs across where the deterministic flow would not. Section 5.2 finds the measured crossing between the boundary and the deterministic onset at every cell compared.

To predict where in between, we keep the fluctuations. At finite d the order parameters $X = (Q, R, \theta)$ follow the diffusion approximation

$$dX = \Big[F(X) + \tfrac{\eta^2}{d}\,\mathbb{E}\big[\chi^2(y^2 + z^2 - 2)\big]\, e_Q\Big]\, d\alpha + \sqrt{\Sigma(X)/d}\; dW, \tag{7}$$

with F the right-hand side of (2)–(4), Σ the covariance of the per-example increment of (Q, R, θ) under the same Gaussian closure, including what the walk of the teacher adds to the variance of the overlap, e_Q the unit vector of the norm coordinate, z the standard Gaussian independent of y with $x = \sqrt{Q}\,(\rho y + \sqrt{1-\rho^2}\, z)$, and W a three-dimensional Wiener process. The bracket keeps an $O(1/d)$ correction to F in the norm that the Gaussian closure omits; dropping it gives the diffusion approximation at leading order, which we report where the two differ. We also drop the $O(1/d)$ correction to the drift of the overlap, which changes the entry fraction by less than 0.001 (components of Σ: Appendix D.2). Below the boundary the same fluctuations drive a rare escape across the unstable state, or across the cutoff with no barrier near the end point. At $(0.3, 0.62)$ the entry fraction falls exponentially in d, with a measured slope of -0.055 ± 0.002 per unit of dimension against -0.051 ± 0.001 from the diffusion approximation.

Small learning rates break the second link, norm inflation, and with it the whole chain. Misalignment alone still raises the threshold, since at the trained norm and threshold $d\theta/d\alpha = \eta\varphi(0)(1-\rho)/2 > 0$ once drift has taken ρ below 1. What decides whether that rise continues is the η^2 term, which inflates the norm over the scanned range and is absent as $\eta \to 0$. Without it the rise is transient: the unit settles at an active resting state with a negative threshold, one that answers almost every input with close to the mean teacher output. For the order-parameter equations Proposition A.14 proves that no entry occurs as $\eta \to 0$, with the reduced trajectory enclosed by interval arithmetic. Below $\eta \approx 0.3$ active resting states are computed numerically at five drift speeds (Appendix D.3).

5 Experiments

5.1 Protocol and criteria

The three subsections answer the questions of Section 4.2 in order, on a main grid crossing $\delta \in \{0.1, 0.15, 0.2, 0.3, 0.45, 0.65\}$ with η from 0.7 to 1.2 in steps of 0.1 ($d = 500$, 100 runs per cell,

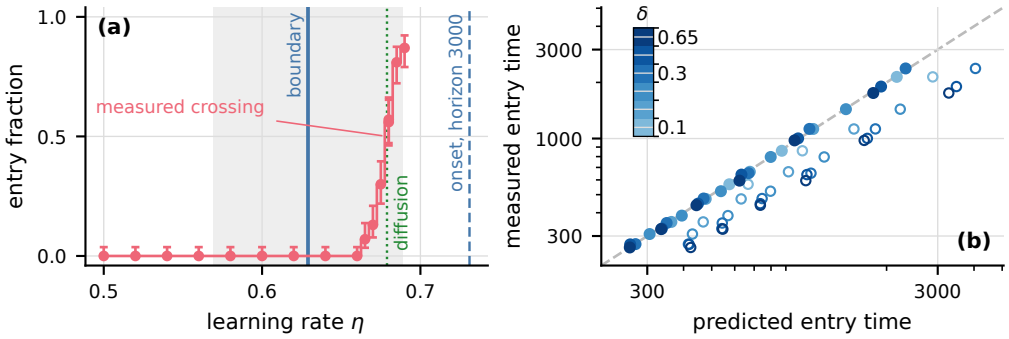


Figure 2: The measured crossing is within 0.003 in η of the diffusion approximation. (a) Entry fraction against η at $\delta = 0.3$, with the boundary, the onset, the diffusion approximation, and in grey the window fixed in advance. (b) Measured against predicted mean entry time for the thirty entering cells by δ; dashed equality, open the equations, filled the diffusion approximation.

one cell per speed-rate pair, trained state). All but one of the 93 criteria were written down before the runs they judge: 46 are met, 26 are not, and 21 have a qualified outcome. The three criteria that judge the main grid all hold: the entry times agree up to one fitted factor, the measured crossing at $\delta = 0.3$ falls inside the window fixed around the boundary, and neither control produces an entry. We keep the record in the appendices: Appendix A the proofs and the statement-scope table; Appendix B the protocols; Appendix C the estimators and the derivation tests; and Appendices D.1 to D.5 the experimental record for the three questions, the protocol audits and the criteria.

5.2 The measured crossing lands where the diffusion approximation puts it

At $\delta = 0.3$ and $d = 500$ the entry fraction climbs from zero at $\eta \leq 0.66$ to 0.87 at $\eta = 0.69$, crossing one half at $\eta = 0.679$ (Figure 2a). That crossing lies between the two curves of Section 4.5: the boundary at 0.629, which a run of unlimited length at infinite dimension would find, and the deterministic onset at 0.731, the least rate at which the equations reach the cutoff within the horizon 3000 of these runs. Because the diffusion approximation keeps the fluctuations that place it in between, it predicts 0.679 as well, against a measured interval of $[0.677, 0.681]$. At every speed, dimension, and horizon compared, on both sides of the end point, the diffusion approximation is within 0.003 of the crossing, again between boundary and onset (Appendix D.1). Raising the dimension moves the crossing up toward the onset, and a longer horizon moves it down toward the boundary. Two readings fixed in advance failed: the measured crossing has no interior minimum through the end point, and the escape-scaling slope below the boundary stayed unassessable at $(0.3, 0.60)$, which gives zero entries; Section 4.5 reports it at $(0.3, 0.62)$, nearer the boundary.

A merely noisy target is not enough at these rates, while accumulated drift is. We see no entry under either control in 1800 runs, and none under a fixed teacher with matched label noise in 2300 runs at rates up to 1.2. A target independent of the input produces entry above $\eta = 0.65$, since like drift it leaves no fixed direction to align with. Near $\eta = 2$ label noise or a random start also produces entry.

5.3 The entry time comes out right once the fluctuations are kept

Thirty of the thirty-six grid cells enter in all 100 runs, and the timing comparison uses them. Integrating the order-parameter equations orders them almost as the measurements do, at Spearman 0.976 against 0.85 for a ranking by learning rate alone. Yet it overestimates every time by a median of 55.8% (Figure 2b, open markers). Keeping the fluctuations removes the excess with nothing fitted: the diffusion approximation has median error 1.3% and Spearman 0.999 (filled markers). The overestimate is a finite-dimension effect: at $(0.3, 0.8)$ the measured mean climbs with dimension toward the $d \to \infty$ value 1830, so their ratio in Figure 3a falls to 1.26 at $d = 1920$. The diffusion approximation tracks that climb, undershooting by 21% at $d = 30$ and by 5% at 120, and agreeing within 0.2% at the two largest dimensions; the leading-order scheme agrees within 6.4% throughout.

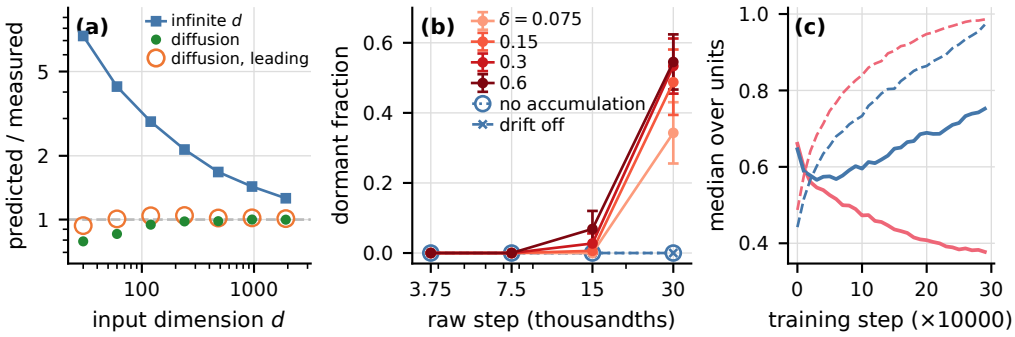


Figure 3: The prediction improves with dimension; the layer shows the rate dependence but not the mechanism. (a) Predicted over measured mean entry time at $(\delta, \eta) = (0.3, 0.8)$. (b) Dormant fraction of the layer against the raw step; controls on zero. (c) Median weight norm (solid) and threshold (dashed) of its units at $(\delta, \eta_{\text{raw}}) = (0.6, 0.03)$: red dormant, blue surviving, not the speeds of (b).

5.4 The isolated unit needs a finite rate, and the layer enters another way

For the isolated unit, dormancy stops at a positive learning rate rather than fading away, which Proposition A.14 proves for its order-parameter equations as $\eta \to 0$. Below $\eta \approx 0.3$ the equations settle at an active state with a negative threshold at every scanned speed. No run enters at $\eta \le 0.66$ at $\delta = 0.3$ (Figure 2a). Runs at batch 500 enter at none of six cells up to $\eta = 2$, since averaging over 500 inputs divides the η^2 term by 500 and moves the fold to η near 10^3 (Lemma A.16).

A second unit changes this, and the same treatment says why. Two units sharing an output error have a symmetric state in which they are interchangeable, and its transverse eigenvalue is positive at all 25 cells scanned (Appendix A.8). Below the single-unit fold the instability always ends with exactly one margin past the cutoff, as 48 of 50 runs do at $(\delta, \eta) = (0.3, 0.1)$, where the isolated unit never enters. The measured growth rate of the threshold difference matches that eigenvalue at three learning rates, and the η^2 term speeds this route rather than causing it.

With a trained output weight no run enters, for the isolated unit or for two units, at any of the eight cells tested. The output weight settles near 0.15; holding it fixed there gives the isolated unit the same outcome. That unit then behaves like one at a fortieth of its nominal rate, below the boundary.

A layer of 128 units at input dimension 32, trained at batch one for 3×10^5 steps, shows the rate dependence and the signature of the other route. At the end of training its dormant fraction is zero at the two smallest raw steps and reaches 34 to 55% of its units at the largest, at every drift speed, with both controls at zero (Figure 3b). Per unit, however, the dormant ones do not track a teacher direction: their alignment stays at chance, and their median norm falls from 0.66 to 0.38 (Figure 3c), as the losing unit of a pair does and the isolated unit does not. They enter about 4.5 times later than predicted, and only removing the shared output error nearly restores which units go dormant. The converted rate of Section 4.1 fails both ways on reruns, needing a raw step four to eight times higher at sixteen times the input dimension, and predicting none where batch 32 gives 15 to 16% dormancy.

6 Conclusion

A single ReLU unit with a learned threshold and a fixed output weight goes dormant under target drift by the mechanism of Section 4.3. The diffusion approximation, which keeps finite-dimension fluctuations, says where and when to 0.003 in η and a median 1.3% in the timing (Section 5). In the large-dimension limit and under plain stochastic gradient descent without weight decay, an isolated unit needs a rate bounded away from zero: none we computed goes dormant below $\eta = 0.554$. Two units sharing one output error enter by an instability the isolated unit lacks, and the loser's weight vector shrinks into the cutoff, as the layer's dormant units do. A trained output weight settles at a small output scale and prevents entry in the one- and two-unit settings everywhere tested, so dormancy under it has some other cause. Whether the same analysis reaches a whole layer is the next question.

AI USE STATEMENT

The authors chose the solvable thresholded unit as the route to dormancy under drift, decided which extensions to pursue and where the claims stop, and kept the criteria register. The authors used generative AI tools for the propositions and the first versions of their proofs, for the hypotheses, for the design, code and execution of the simulations, for the reading of their results, and for the figures and the text. The tools produced no observations: every data point comes from the simulations described. Appendix C.3 covers the derivations with numerical tests, and Appendix D.5 lists the failed and qualified criteria, with every outcome in the supplement. The authors read every sentence and are responsible for all of it.

REPRODUCIBILITY STATEMENT

Appendix B gives every grid, run count, horizon, initial state and seed convention. Appendix C the estimators and the derivation tests. The supplement holds the code (closed forms; integrator; the validated enclosure of the reduced trajectory and the interval evaluations of the constants it needs; every simulation program and analysis script), the per-run result files with their parameters and seeds, the boundary and entry-time data, and the criteria register.

REFERENCES

Kwangjun Ahn, Sébastien Bubeck, Sinho Chewi, Yin Tat Lee, Felipe Suarez, and Yi Zhang. Learning threshold neurons via edge of stability. In *Advances in Neural Information Processing Systems 36 (NeurIPS 2023)*, 2023.

Isac Arnekvist, Joao Frederico Carvalho, Danica Kragic, and Johannes A. Stork. The effect of target normalization and momentum on dying ReLU, 2020. arXiv:2005.06195.

Michael Biehl and Holm Schwarze. Learning drifting concepts with neural networks. *Journal of Physics A: Mathematical and General*, 26(11):2651–2665, 1993. doi: 10.1088/0305-4470/26/11/014.

Michael Biehl and Holm Schwarze. Learning by on-line gradient descent. *Journal of Physics A: Mathematical and General*, 28(3):643–656, 1995.

Feng Chen, Daniel Kunin, Atsushi Yamamura, and Surya Ganguli. Stochastic collapse: How gradient noise attracts SGD dynamics towards simpler subnetworks. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Otavio Citton, Frederieke Richert, and Michael Biehl. The role of the learning rate in layered neural networks with ReLU activation function. In *Proceedings of the European Symposium on Artificial Neural Networks (ESANN)*, pp. 437–442, 2025a. doi: 10.14428/esann/2025.ES2025-94.

Otavio Citton, Frederieke Richert, Michael Biehl, and Michiel Straat. Hermite polynomials facilitating on-line learning analysis of layered neural networks with arbitrary activation function. *Neurocomputing*, 655:131328, 2025b. doi: 10.1016/j.neucom.2025.131328.

Shibhansh Dohare, J. Fernando Hernandez-Garcia, Qingfeng Lan, Parash Rahman, A. Rupam Mahmood, and Richard S. Sutton. Loss of plasticity in deep continual learning. *Nature*, 632:768–774, 2024.

Dominik Endres and Peter Riegler. Learning dynamics on different timescales. *Journal of Physics A: Mathematical and General*, 32(49):8655–8663, 1999. doi: 10.1088/0305-4470/32/49/306.

Mark I. Freidlin and Alexander D. Wentzell. *Random Perturbations of Dynamical Systems*, volume 260 of *Grundlehren der mathematischen Wissenschaften*. Springer, 3rd edition, 2012. doi: 10.1007/978-3-642-25847-3.

Jacopo Graldi, Alessandro Breccia, Giulia Lanzillotta, Thomas Hofmann, and Lorenzo Noci. The importance of being lazy: Scaling limits of continual learning. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 20326–20367, 2025.

Naoki Hiratani. Disentangling and mitigating the impact of task similarity for continual learning. In *Advances in Neural Information Processing Systems*, volume 37, pp. 3243–3274, 2024.

Amir Joudaki, Giulia Lanzillotta, Mohammad Samragh Razlighi, Iman Mirzadeh, Keivan Alizadeh, Thomas Hofmann, Mehrdad Farajtabar, and Fartash Faghri. Barriers for learning in an evolving world: Mathematical understanding of loss of plasticity. In *International Conference on Learning Representations*, 2026.

Osame Kinouchi and Nestor Caticha. Lower bounds on generalization errors for drifting rules. *Journal of Physics A: Mathematical and General*, 26(22):6161–6171, 1993. doi: 10.1088/0305-4470/26/22/017.

Sebastian Lee, Sebastian Goldt, and Andrew Saxe. Continual learning in the teacher-student setup: Impact of task similarity. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6109–6119, 2021.

Tianhui Liu and Lili Mou. Do neural networks lose plasticity in a gradually changing world?, 2026. arXiv:2602.09234.

Lu Lu, Yeonjong Shin, Yanhui Su, and George Em Karniadakis. Dying ReLU and initialization: Theory and numerical examples. *Communications in Computational Physics*, 28(5):1671–1706, 2020. doi: 10.4208/cicp.OA-2020-0165.

Clare Lyle, Zeyu Zheng, Khimya Khetarpal, Hado van Hasselt, Razvan Pascanu, James Martens, and Will Dabney. Disentangling the causes of plasticity loss in neural networks. In *Proceedings of the 3rd Conference on Lifelong Learning Agents (CoLLAs)*, volume 274 of *Proceedings of Machine Learning Research*, pp. 750–783, 2025.

Clare Lyle, Ghada Sokar, András György, and Razvan Pascanu. What can grokking teach us about learning under nonstationarity? In *Proceedings of the 4th Conference on Lifelong Learning Agents (CoLLAs)*, volume 330 of *Proceedings of Machine Learning Research*, pp. 635–656, 2026.

Francesco Mori, Stefano Sarao Mannelli, and Francesca Mignacco. Optimal protocols for continual learning via statistical physics and control theory. *Journal of Statistical Mechanics: Theory and Experiment*, 2025(8):084004, 2025. doi: 10.1088/1742-5468/adf296.

Donald B. Owen. Tables for computing bivariate normal probabilities. *The Annals of Mathematical Statistics*, 27(4):1075–1090, 1956. doi: 10.1214/aoms/1177728074.

Arjun Prakash, Naicheng He, Kaicheng Guo, Saket Tiwari, Tyrone Serapio, Ruo Yu Tao, Amy Greenwald, and George Konidaris. Spectral collapse drives loss of plasticity in deep continual learning. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

David Saad and Sara A. Solla. Exact solution for on-line learning in multilayer neural networks. *Physical Review Letters*, 74(21):4337–4340, 1995. doi: 10.1103/PhysRevLett.74.4337.

Ghada Sokar, Rishabh Agarwal, Pablo Samuel Castro, and Utku Evci. The dormant neuron phenomenon in deep reinforcement learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32145–32168, 2023.

Michiel Straat. *Machine learning: statistical physics based theory and smart industry applications*. PhD thesis, University of Groningen, 2022.

Michiel Straat, Fthi Abadi, Zhuoyun Kan, Christina Göpfert, Barbara Hammer, and Michael Biehl. Supervised learning in the presence of concept drift: a modelling framework. *Neural Computing and Applications*, 34(1):101–118, 2022. doi: 10.1007/s00521-021-06035-1.

Steven H. Strogatz. *Nonlinear Dynamics and Chaos*. CRC Press, Boca Raton, 2nd edition, 2018. doi: 10.1201/9780429492563.

Hongyao Tang, Johan Obando-Ceron, Pablo Samuel Castro, Aaron Courville, and Glen Berseth. Mitigating plasticity loss in continual reinforcement learning by reducing churn. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025.

Jiuqi Wang, Jayanth Srinivasa, Claire Chen, Shuze Daniel Liu, Ali Payani, and Shangtong Zhang. Predicting plasticity in deep continual learning: A theoretical perspective. *arXiv preprint arXiv:2605.09044*, 2026.

Ansgar H. L. West, David Saad, and Ian T. Nabney. The learning dynamics of a universal approximator. In Michael C. Mozer, Michael I. Jordan, and Thomas Petsche (eds.), *Advances in Neural Information Processing Systems 9*, pp. 288–294. MIT Press, 1997.

Zihao Wu, Hongyao Tang, Yi Ma, Jiashun Liu, Yan Zheng, and Jianye Hao. The rank and gradient lost in non-stationarity: Sample weight decay for mitigating plasticity loss in reinforcement learning. In *International Conference on Learning Representations*, 2026.

Liu Ziyin. Symmetry induces structure and constraint of learning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 62847–62866, 2024.

A PROOFS

Table 1 gives the scope and status of each formal statement of Section 4, restated and proven below.

Table 1: Statements of the model with their scope and status; P, proposition.

statement	value or regime	status
fixed-point map Ψ (P3)	$\rho \in (0, 1)$	proven; residual 10^{-14}
fold locus and its saddle-node character	scanned δ range	condition $\det D\Psi = 0$ explicit; root and branch located numerically; exponent 1.05 ± 0.01
slow-drift anchor η_0 (PA.10)	$\delta \to 0^+$: 2.0224	branch and its $m \to \infty$ limit proven given interior alignment; the $\delta \to 0^+$ convergence and the value numerical
fast-drift envelope $2/\pi \approx 0.6366$ (PA.8)	$\rho = 0$, $m \to \infty$	asymptote proven at fixed alignment; the branch's approach numerical
critical end point $(\delta_{\mathrm{cp}}, \eta_{\mathrm{cp}})$	$(0.5939, 0.5623)$, attained by Ψ at $(m, \rho) \approx (1.352, 0.0459)$	numerical root; the boundary leaves the fold at $\delta \approx 0.55$
entry-time law (6)	above the boundary, first passage of $m_\star = 2$ within the horizon	first-passage functional, no closed form; tested in Section 5.3
no dormancy below $\eta \approx 0.3$	$\delta \in \{0.1, 0.3, 0.65\}$ to 10^5 units; fixed points at $\delta \in \{1.5, 3\}$	numerical: active fixed point with negative threshold at every rate up to 0.3 (App. D.3); proven at every δ for η small enough, with the reduced trajectory enclosed by interval arithmetic (PA.14)

Appendix A.1 derives the truncated Gaussian atoms and the four moments of Lemma 2, Appendix A.2 the drift term of the order-parameter equations, and Appendix A.3 the asymptotically absorbing structure of deep dormancy at bounded norm. Lemma A.16 derives the learning terms at general batch size, and (2)–(4) is the case of batch one. Appendix A.4 proves the explicit parameterization of Proposition 3, Appendix A.5 the two asymptotic landmarks of Propositions A.8 and A.10, Appendix A.6 the small-rate limit of Proposition A.14, Appendix A.7 what changes under a different threshold rate, minibatches and weight decay, and Appendix A.8 the same treatment for two units sharing one output error.

A.1 SETTING, ATOMS, AND THE CLOSED FORMS

Write $u = x/\sqrt{Q}$, which makes (u, y) a standard bivariate Gaussian pair with correlation $\rho = R/\sqrt{Q}$, and write $s = \sqrt{1 - \rho^2}$. Here the student output is $\sigma = \sqrt{Q}\,(u - m)\,\mathbb{1}_a$ with $\mathbb{1}_a = \mathbb{1}\{u >$

$m\}$, the teacher output is $\tau = y\,\mathbb{1}\{y > 0\}$, and the error signal is $\chi = (\tau - \sigma)\mathbb{1}_a$. Every moment of χ therefore reduces to truncated Gaussian moments over the events $\mathbb{1}_a$ and $\mathbb{1}_b = \mathbb{1}\{u > m,\, y > 0\}$. Throughout, φ and Φ are the standard Gaussian density and distribution function, and $\bar{\Phi} = 1 - \Phi$. Owen's T function is $T_{\mathrm{O}}(h, a)$. Its definition in Owen (1956) gives the identity

$$\int_h^\infty \varphi(t)\,\Phi(at)\,dt = \tfrac{1}{2}\,\bar{\Phi}(h) + T_{\mathrm{O}}(h, a). \tag{8}$$

Lemma A.1 (truncated Gaussian atoms)**.** *With $L(m;\rho) := \mathbb{E}[\mathbb{1}_b]$, $M(m) := \varphi(m) - m\,\bar{\Phi}(m)$, and $G(m,\rho) := \rho\,\varphi(m)\,\Phi(\rho m/s) + \varphi(0)\,\bar{\Phi}(m/s)$,*

$$L(m;\rho) = \tfrac{1}{2}\bar{\Phi}(m) + T_{\mathrm{O}}(m, \rho/s), \tag{9}$$

$$\mathbb{E}[(u-m)\mathbb{1}_a] = M(m), \tag{10}$$

$$\mathbb{E}[(u-m)^2\mathbb{1}_a] = (1+m^2)\,\bar{\Phi}(m) - m\,\varphi(m), \tag{11}$$

$$\mathbb{E}[u(u-m)\mathbb{1}_a] = \bar{\Phi}(m), \tag{12}$$

$$\mathbb{E}[y\,\mathbb{1}_b] = G(m,\rho), \tag{13}$$

$$\mathbb{E}[uy\,\mathbb{1}_b] = \rho L + \rho m\,\varphi(m)\,\Phi(\rho m/s) + s\,\varphi(0)\,\varphi(m/s), \tag{14}$$

$$\mathbb{E}[y^2\mathbb{1}_b] = L + \rho^2 m\,\varphi(m)\,\Phi(\rho m/s) + \rho s\,\varphi(0)\,\varphi(m/s). \tag{15}$$

At the edge $s \to 0$ ($\rho \to \pm 1$) the limit is removable: $\Phi(\rho m/s) \to \mathbb{1}\{\pm m > 0\}$ and $\varphi(m/s)/s \to 0$ for $m \neq 0$.

Proof. One-dimensional atoms follow from $\int_m^\infty u\,\varphi(u)\,du = \varphi(m)$ and $\int_m^\infty u^2\varphi(u)\,du = \bar{\Phi}(m) + m\,\varphi(m)$:

$$\begin{aligned}
\mathbb{E}[(u-m)\mathbb{1}_a] &= \int_m^\infty (u-m)\,\varphi(u)\,du = \varphi(m) - m\,\bar{\Phi}(m),\\
\mathbb{E}[(u-m)^2\mathbb{1}_a] &= \int_m^\infty (u^2 - 2mu + m^2)\,\varphi(u)\,du\\
&= \bar{\Phi}(m) + m\,\varphi(m) - 2m\,\varphi(m) + m^2\,\bar{\Phi}(m)\\
&= (1+m^2)\,\bar{\Phi}(m) - m\,\varphi(m),\\
\mathbb{E}[u(u-m)\mathbb{1}_a] &= \bar{\Phi}(m) + m\,\varphi(m) - m\,\varphi(m) = \bar{\Phi}(m).
\end{aligned}$$

For the bivariate atoms condition on u: from $y \mid u \sim \mathcal{N}(\rho u,\, s^2)$ and the first two moments of a Gaussian truncated to the positive half-line,

$$\Pr(y > 0 \mid u) = \Phi(\rho u/s),$$

$$\mathbb{E}[y\mathbb{1}\{y>0\} \mid u] = \rho u\,\Phi(\rho u/s) + s\,\varphi(\rho u/s), \tag{16}$$

$$\mathbb{E}[y^2\mathbb{1}\{y>0\} \mid u] = (\rho^2 u^2 + s^2)\,\Phi(\rho u/s) + \rho u s\,\varphi(\rho u/s). \tag{17}$$

Three integrals over $u \in (m, \infty)$ then give the remaining atoms. Completing the square with $s^2 + \rho^2 = 1$ gives $\varphi(u)\,\varphi(\rho u/s) = \varphi(0)\,\varphi(u/s)$, so

$$\int_m^\infty \varphi(u)\,\varphi(\rho u/s)\,du = \varphi(0)\int_m^\infty \varphi(u/s)\,du = s\,\varphi(0)\,\bar{\Phi}(m/s), \tag{18}$$

$$\int_m^\infty u\,\varphi(u)\,\varphi(\rho u/s)\,du = \varphi(0)\int_m^\infty u\,\varphi(u/s)\,du = s^2\,\varphi(0)\,\varphi(m/s). \tag{19}$$

Integrating by parts with $u\,\varphi(u) = -\varphi'(u)$ and using (18),

$$\begin{aligned}
\int_m^\infty u\,\varphi(u)\,\Phi(\rho u/s)\,du &= \Big[-\varphi(u)\,\Phi(\rho u/s)\Big]_m^\infty + \frac{\rho}{s}\int_m^\infty \varphi(u)\,\varphi(\rho u/s)\,du\\
&= \varphi(m)\,\Phi(\rho m/s) + \rho\,\varphi(0)\,\bar{\Phi}(m/s).
\end{aligned} \tag{20}$$

Integrating by parts with $u^2\varphi(u) = \varphi(u) - (u\,\varphi(u))'$ and using (19),

$$\begin{aligned}
\int_m^\infty u^2\varphi(u)\,\Phi(\rho u/s)\,du &= \int_m^\infty \varphi(u)\,\Phi(\rho u/s)\,du - \Big[u\,\varphi(u)\,\Phi(\rho u/s)\Big]_m^\infty\\
&\quad + \frac{\rho}{s}\int_m^\infty u\,\varphi(u)\,\varphi(\rho u/s)\,du\\
&= L + m\,\varphi(m)\,\Phi(\rho m/s) + \rho s\,\varphi(0)\,\varphi(m/s),
\end{aligned} \tag{21}$$

where $L = \int_m^\infty \varphi(u)\,\Phi(\rho u/s)\,du$ is (9) by the identity (8) with $a = \rho/s$. Integrating (16) against $\varphi(u)$ over (m, ∞) and using (20) and (18),

$$\begin{aligned}\mathbb{E}[y\,\mathbb{1}_b] &= \rho \int_m^\infty u\,\varphi(u)\,\Phi(\rho u/s)\,du + s\int_m^\infty \varphi(u)\,\varphi(\rho u/s)\,du \\ &= \rho\,\varphi(m)\,\Phi(\rho m/s) + \rho^2\,\varphi(0)\,\bar\Phi(m/s) + s^2\,\varphi(0)\,\bar\Phi(m/s) \\ &= \rho\,\varphi(m)\,\Phi(\rho m/s) + \varphi(0)\,\bar\Phi(m/s) = G(m,\rho).\end{aligned}$$

Multiplying (16) by u before integrating and using (21) and (19),

$$\begin{aligned}\mathbb{E}[uy\,\mathbb{1}_b] &= \rho \int_m^\infty u^2\varphi(u)\,\Phi(\rho u/s)\,du + s\int_m^\infty u\,\varphi(u)\,\varphi(\rho u/s)\,du \\ &= \rho L + \rho m\,\varphi(m)\,\Phi(\rho m/s) + \rho^2 s\,\varphi(0)\,\varphi(m/s) + s^3\,\varphi(0)\,\varphi(m/s) \\ &= \rho L + \rho m\,\varphi(m)\,\Phi(\rho m/s) + s\,\varphi(0)\,\varphi(m/s).\end{aligned}$$

Integrating (17) against $\varphi(u)$ and using (21), (9), and (19),

$$\begin{aligned}\mathbb{E}[y^2\mathbb{1}_b] &= \rho^2 \int_m^\infty u^2\varphi(u)\,\Phi(\rho u/s)\,du + s^2 L + \rho s\int_m^\infty u\,\varphi(u)\,\varphi(\rho u/s)\,du \\ &= \rho^2 L + \rho^2 m\,\varphi(m)\,\Phi(\rho m/s) + \rho^3 s\,\varphi(0)\,\varphi(m/s) + s^2 L + \rho s^3\,\varphi(0)\,\varphi(m/s) \\ &= L + \rho^2 m\,\varphi(m)\,\Phi(\rho m/s) + \rho s\,\varphi(0)\,\varphi(m/s),\end{aligned}$$

which completes the list. □

Restatement (Lemma 2, closed forms)**.** *Four moments of the error signal are*

$$\mathbb{E}[\chi] = G(m,\rho) - \sqrt{Q}\,M(m), \tag{22}$$

$$\mathbb{E}[\chi y] = \mathbb{E}[y^2\mathbb{1}_b] - R\,\bar\Phi(m), \tag{23}$$

$$\mathbb{E}[\chi x] = \sqrt{Q}\,\mathbb{E}[uy\,\mathbb{1}_b] - Q\,\bar\Phi(m), \tag{24}$$

$$\mathbb{E}[\chi^2] = \mathbb{E}[y^2\mathbb{1}_b] - 2\sqrt{Q}\left(\mathbb{E}[uy\,\mathbb{1}_b] - m\,\mathbb{E}[y\,\mathbb{1}_b]\right) + Q\left[(1+m^2)\bar\Phi(m) - m\varphi(m)\right], \tag{25}$$

with the atoms of Lemma A.1, and the generalization error is $\epsilon_g = \frac{1}{2}\mathbb{E}[(\sigma-\tau)^2] = \frac{1}{4} - \sqrt{Q}\left(\mathbb{E}[uy\,\mathbb{1}_b] - m\,\mathbb{E}[y\,\mathbb{1}_b]\right) + \frac{1}{2}Q\left[(1+m^2)\bar\Phi(m) - m\varphi(m)\right]$.

Proof. Expand $\chi = (\tau-\sigma)\mathbb{1}_a = y\,\mathbb{1}_b - \sqrt{Q}\,(u-m)\mathbb{1}_a$ and use $\mathbb{E}[y\,\mathbb{1}_a] = \rho\,\mathbb{E}[u\,\mathbb{1}_a] = \rho\,\varphi(m)$ and $\mathbb{E}[uy\,\mathbb{1}_a] = \rho\,\mathbb{E}[u^2\mathbb{1}_a] = \rho\big(\bar\Phi(m) + m\,\varphi(m)\big)$, both from $\mathbb{E}[y \mid u] = \rho u$. Then

$$\begin{aligned}\mathbb{E}[\chi] &= \mathbb{E}[y\,\mathbb{1}_b] - \sqrt{Q}\,\mathbb{E}[(u-m)\mathbb{1}_a] = G(m,\rho) - \sqrt{Q}\,M(m), \\ \mathbb{E}[\chi y] &= \mathbb{E}[y^2\mathbb{1}_b] - \sqrt{Q}\,\mathbb{E}[y(u-m)\mathbb{1}_a] \\ &= \mathbb{E}[y^2\mathbb{1}_b] - \sqrt{Q}\left(\mathbb{E}[uy\,\mathbb{1}_a] - m\,\mathbb{E}[y\,\mathbb{1}_a]\right) \\ &= \mathbb{E}[y^2\mathbb{1}_b] - \sqrt{Q}\,\rho\big(\bar\Phi(m) + m\,\varphi(m) - m\,\varphi(m)\big) \\ &= \mathbb{E}[y^2\mathbb{1}_b] - R\,\bar\Phi(m), \\ \mathbb{E}[\chi x] &= \sqrt{Q}\,\mathbb{E}[\chi u] = \sqrt{Q}\,\mathbb{E}[uy\,\mathbb{1}_b] - Q\,\mathbb{E}[u(u-m)\mathbb{1}_a] = \sqrt{Q}\,\mathbb{E}[uy\,\mathbb{1}_b] - Q\,\bar\Phi(m), \\ \mathbb{E}[\chi^2] &= \mathbb{E}[y^2\mathbb{1}_b] - 2\sqrt{Q}\,\mathbb{E}[y(u-m)\mathbb{1}_b] + Q\,\mathbb{E}[(u-m)^2\mathbb{1}_a] \\ &= \mathbb{E}[y^2\mathbb{1}_b] - 2\sqrt{Q}\left(\mathbb{E}[uy\,\mathbb{1}_b] - m\,\mathbb{E}[y\,\mathbb{1}_b]\right) + Q\left[(1+m^2)\bar\Phi(m) - m\varphi(m)\right],\end{aligned}$$

where the second line of the last chain uses $\mathbb{1}_b\mathbb{1}_a = \mathbb{1}_b$ and (11). For the generalization error, $\mathbb{E}[\tau^2] = \mathbb{E}[y^2\mathbb{1}\{y>0\}] = \frac{1}{2}$, $\mathbb{E}[\sigma\tau] = \sqrt{Q}\,\mathbb{E}[(u-m)y\,\mathbb{1}_b]$, and $\mathbb{E}[\sigma^2] = Q\,\mathbb{E}[(u-m)^2\mathbb{1}_a]$. Then $\epsilon_g = \frac{1}{2}\big(\mathbb{E}[\tau^2] - 2\mathbb{E}[\sigma\tau] + \mathbb{E}[\sigma^2]\big)$ is the stated expression. □

Remark A.2 (numerical anchor). At $m = 0$, $Q = 1$, $\rho \to 1$ all four moments vanish. Our implementation checks this vanishing as a limit at $\rho = 1 - 10^{-6}$, with residuals at most 2×10^{-4}, consistent with the removable-limit rate.

A.2 THE DRIFT TERM

Lemma A.3 (the drift term of the order-parameter equations)**.** *Let one step of the walk of Section 4.1 replace B by $B' = \cos(\omega)B + \sin(\omega)z$ with $z \perp B$, $\|z\| = 1$, z isotropic in the tangent space at B and independent of w, and $\cos(\omega) = 1 - \delta/d$. Then*

$$\mathbb{E}\big[w\cdot B' - w\cdot B \;\big|\; w, B\big] = -\frac{\delta}{d}\, w\cdot B, \qquad \mathrm{Var}\left[w\cdot B' - w\cdot B \;\middle|\; w, B\right] = \frac{\sin^2(\omega)}{d-1}\left(Q - R^2\right), \tag{26}$$

so over one unit of learning time (d steps) the walk contributes $-\delta R$ to $dR/d\alpha$ with fluctuations of variance $2\delta(Q - R^2)/d + O(d^{-2})$, and nothing to $dQ/d\alpha$ or $d\theta/d\alpha$.

Proof. Taking the inner product of B' with w,

$$\begin{aligned} w\cdot B' - w\cdot B &= (\cos\omega - 1)\, w\cdot B + \sin\omega\; w\cdot z \\ &= -\frac{\delta}{d}\, w\cdot B + \sin\omega\; w\cdot z. \end{aligned}$$

$\mathbb{E}[w\cdot z \mid w, B] = 0$ because z and $-z$ are equally likely under the isotropic law on the tangent space. This gives the mean. Write $w_\perp = w - (w\cdot B)B$ for the component of w orthogonal to B. Then $w\cdot z = w_\perp \cdot z$. For a unit vector z isotropic in the $(d-1)$-dimensional tangent space, $\mathbb{E}[(w_\perp\cdot z)^2] = \|w_\perp\|^2/(d-1)$. With $\|w_\perp\|^2 = Q - R^2$ this is the variance. Over d steps the mean increment accumulates to $-\delta R$ per unit of α, while the variance accumulates to

$$d\cdot\frac{\sin^2(\omega)}{d-1}\,(Q-R^2) = \frac{d}{d-1}\Big(\frac{2\delta}{d} - \frac{\delta^2}{d^2}\Big)(Q-R^2) = \frac{2\delta(Q-R^2)}{d} + O(d^{-2}),$$

the same $O(d^{-1})$ order as the fluctuations of the learning terms. This step changes neither w nor θ, so Q and θ receive no contribution from it. □

A.3 ASYMPTOTICALLY ABSORBING STRUCTURE AT BOUNDED NORM

Proposition A.4 (asymptotically absorbing structure at bounded norm)**.** *For $m \geq 0$ and every $\rho \in [-1, 1]$, each of $|\mathbb{E}[\chi]|$, $|\mathbb{E}[\chi x]|$, $|\mathbb{E}[\chi y]|$, and $\mathbb{E}[\chi^2]$ is at most*

$$C(Q)\,(1+m)^2\,\big(\varphi(m) + \bar{\Phi}(m)\big), \qquad C(Q) = 6\,(1+Q), \tag{27}$$

so every right-hand side of (2)–(4) except the drift term $-\delta R$ vanishes at a Gaussian rate as $m \to \infty$ at bounded Q, while $dR/d\alpha \to -\delta R$. In this limit the unit stops learning, and its overlap decays at rate δ up to a term of the same Gaussian order.

Proof. For $m \geq 0$ and $s \leq 1$ one has $m/s \geq m$; hence $\bar{\Phi}(m/s) \leq \bar{\Phi}(m)$ and $\varphi(m/s) \leq \varphi(m)$. Also $\Phi \leq 1$, $|\rho| \leq 1$, $s \leq 1$, $\varphi(0) \leq \frac{1}{2}$, and $L \leq \bar{\Phi}(m)$. Abbreviate $E := \varphi(m) + \bar{\Phi}(m)$. From Lemma A.1,

$$\begin{aligned} |G(m,\rho)| &\leq \varphi(m) + \tfrac{1}{2}\bar{\Phi}(m) \leq E, \\ 0 < M(m) &= \int_m^\infty (u-m)\,\varphi(u)\,du \leq \varphi(m) \leq E, \\ |\mathbb{E}[uy\,\mathbb{1}_b]| &\leq \bar{\Phi}(m) + m\,\varphi(m) + \tfrac{1}{2}\varphi(m) \leq (2+m)\,E, \\ |\mathbb{E}[y^2\mathbb{1}_b]| &\leq \bar{\Phi}(m) + m\,\varphi(m) + \tfrac{1}{2}\varphi(m) \leq (2+m)\,E, \\ (1+m^2)\bar{\Phi}(m) - m\varphi(m) &\leq (1+m^2)\,E. \end{aligned}$$

Inserting these into (22)–(25) with $|R| \leq \sqrt{Q}$,

$$\begin{aligned} |\mathbb{E}[\chi]| &\leq (1+\sqrt{Q})\,E, \\ |\mathbb{E}[\chi y]| &\leq (2+m+\sqrt{Q})\,E, \\ |\mathbb{E}[\chi x]| &\leq \big(\sqrt{Q}\,(2+m) + Q\big)\,E, \\ \mathbb{E}[\chi^2] &\leq \big((2+m) + 2\sqrt{Q}\,(2+2m) + Q\,(1+m^2)\big)\,E. \end{aligned}$$

Each bracket is at most $(1+m)^2\,(2+6\sqrt{Q}+Q)$, because $2+m \le 2(1+m)^2$, $2+2m \le 2(1+m)^2$, and $1+m^2 \le (1+m)^2$. Also $2+6\sqrt{Q}+Q \le 6(1+Q)$ since $6\sqrt{Q} \le 3(1+Q)$. This is (27). For $m \ge 1$ we have $\bar{\Phi}(m) \le \varphi(m)/m \le \varphi(m)$. This bounds the envelope by $2C(Q)(1+m)^2\varphi(m)$, a Gaussian rate. In (2) the only term without this envelope is $-\delta R$, so $dR/d\alpha + \delta R \to 0$. □

At the cutoff itself the moments are small but not negligible. Take the state where the trajectory from the trained state at $(\delta, \eta) = (0.3, 0.8)$ first reaches $m = 2$, namely $(Q, R, \theta) = (1.44, 0.03, 2.40)$. There the activity is 0.023 and $\mathbb{E}[\chi^2] = 0.012$, against 0.34 and 0.19 at the tracking state of $(0.3, 0.6)$. At $m = 3$ (reached at $(0.65, 1.2)$ after 1.4×10^4 units) they are 0.0014 and 0.0007. Lemma A.1 gives these values in closed form. Between the cutoff and the tracking state the ratio of $\mathbb{E}[\chi^2]$ is 16, and between $m = 3$ and the tracking state 280.

Remark A.5 (where the proposition applies). For this bound Q must be bounded along the trajectory. On the four integrated flows below, Q stays below 5 up to first passage of the cutoff.

The maximum of Q up to first passage of the cutoff is 1.44, 1.76, and 4.66 at $(\delta, \eta) = (0.3, 0.8)$, $(0.3, 1.2)$, and $(0.1, 2.1)$, and 1.05 on the tracking flow at $(1.2, 0.62)$ over 3000 units. Past first passage the norm keeps growing, to 1.99 at $(0.3, 1.2)$ by 3000 units in Figure 6.

The stationary dormant branch of Appendix A.5, on which Q^* diverges with the margin, is by contrast a family of fixed points and not a trajectory of the flow. Appendix D.2 reports the finite-d recrossings.

A.4 The explicit parameterization

Restatement (Proposition 3, explicit parameterization)**.** *For $m \in \mathbb{R}$ and $\rho \in (0, 1)$ define Q^*, η^*, δ^* by (5). Wherever all three are positive and finite, $(Q, R, \theta) = (Q^*, \rho\sqrt{Q^*}, m\sqrt{Q^*})$ is a fixed point of (2)–(4) at (δ^*, η^*), and every fixed point with $\rho \in (0, 1)$ and $\eta > 0$ arises this way.*

Proof. We solve the stationarity conditions of (2)–(4) in the order threshold, norm, alignment. By (22) the threshold equation is linear in $\sqrt{Q}$:

$$0 = \mathbb{E}[\chi] = G(m, \rho) - \sqrt{Q}\, M(m) \quad \Longleftrightarrow \quad \sqrt{Q^*} = \frac{G(m, \rho)}{M(m)},$$

and this equation never degenerates because $M(m) = \int_m^\infty (u - m)\,\varphi(u)\,du > 0$ for every m. With $Q = Q^*$ fixed, the norm equation factors:

$$0 = 2\eta\,\mathbb{E}[\chi x] + \eta^2\,\mathbb{E}[\chi^2] = \eta\big(2\,\mathbb{E}[\chi x] + \eta\,\mathbb{E}[\chi^2]\big) \quad \Longleftrightarrow \quad \eta = 0 \ \text{ or } \ \eta^* = -\frac{2\,\mathbb{E}[\chi x]}{\mathbb{E}[\chi^2]},$$

where $\mathbb{E}[\chi^2] > 0$ unless $\chi \equiv 0$. A positive learning rate exists exactly where $\mathbb{E}[\chi x] < 0$, that is, where the deterministic drive contracts the norm so that the η^2 term can balance it. With $R^* = \rho\sqrt{Q^*}$ the alignment equation gives

$$0 = \eta^*\,\mathbb{E}[\chi y] - \delta R^* \quad \Longleftrightarrow \quad \delta^* = \frac{\eta^*\,\mathbb{E}[\chi y]}{\rho\sqrt{Q^*}}.$$

Conversely, a fixed point (Q, R, θ) with $\rho = R/\sqrt{Q} \in (0, 1)$ and $\eta > 0$ determines (m, ρ). The three stationarity conditions then force $Q = Q^*(m, \rho)$, $\eta = \eta^*(m, \rho)$, and $\delta = \delta^*(m, \rho)$ in turn. Hence Ψ is onto the fixed-point manifold. □

Appendix C.3 tests the parameterization numerically.

Remark A.6 (the stationary threshold scales with the norm). At fixed (m, ρ) the stationary threshold is $\theta^* = m\sqrt{Q^*}$, so along the threshold balance it grows with $\sqrt{Q^*}$.

Here the margin $m = \theta^*/\sqrt{Q^*}$ is the quantity the η^2 term of Section 4.5 moves. Numerically the ratio $G(m, \rho)/M(m)$ of the threshold balance increases with the margin at every alignment: on a grid of ρ from 0.01 to 0.999 and m from -3 to 8 every difference quotient is positive. This is the statement Section 4.3 uses when it says that a longer weight vector can be balanced only at a larger margin. The integrated trajectories of Section 5 show the margin rise above the fold.

A.5 THE ANCHOR AND THE ENVELOPE

At the edge $\rho = 0$ the fixed-point equations extend by continuity: $s = 1$, and by Lemma A.1 the atoms are elementary,

$$G(m,0) = \varphi(0)\,\bar{\Phi}(m), \qquad \mathbb{E}[uy\,\mathbb{1}_b] = \varphi(0)\,\varphi(m), \qquad \mathbb{E}[y^2\mathbb{1}_b] = L(m;0) = \tfrac{1}{2}\bar{\Phi}(m), \quad (28)$$

since $T_{\mathrm{O}}(m,0) = 0$. The solutions of the extended balances at $\rho = 0$ are not fixed points of the order-parameter equations at any finite δ, since the alignment balance $\eta\mathbb{E}[\chi y] = \delta R$ has $\mathbb{E}[\chi y] = \frac{1}{2}\bar{\Phi}(m) > 0$ at $R = 0$. What extends is the pair of threshold and norm balances. Their $\delta^* = \eta^*\mathbb{E}[\chi y]/(\rho\sqrt{Q^*})$ tends to infinity as $\rho \to 0$.

At the other edge $\rho = 1$ the map Ψ gives $\delta^* < 0$ for every $m > 0$ and no positive learning rate for $m < 0$, so no fixed point with $\delta, \eta > 0$ lies there; at $m = 0$ the point $(m, \rho) = (0, 1)$ is the trained state, a fixed point only at $\delta = 0$.

Write $r(m) := \varphi(m)/\bar{\Phi}(m)$ for the Gaussian hazard function.

Lemma A.7 (the hazard gap)**.** *For every m, $r(m) - m = M(m)/\bar{\Phi}(m) > 0$, the function $r - m$ is strictly decreasing, and as $m \to \infty$*

$$r(m) - m = \frac{1}{m} - \frac{2}{m^3} + O(m^{-5}), \qquad r(m)\big(r(m) - m\big) = 1 - \frac{1}{m^2} + O(m^{-4}). \quad (29)$$

Proof. Positivity follows from $M > 0$, shown in the proof of Proposition 3. Differentiating with $\varphi' = -m\varphi$ and $\bar{\Phi}' = -\varphi$,

$$r'(m) = \frac{-m\,\varphi(m)\,\bar{\Phi}(m) + \varphi(m)^2}{\bar{\Phi}(m)^2} = r(m)\big(r(m) - m\big),$$

$$\frac{d}{dm}\big(r(m) - m\big) = r(m)\big(r(m) - m\big) - 1.$$

For a standard Gaussian U truncated to (m, ∞) one has $\mathbb{E}[U \mid U > m] = r(m)$ and $\mathbb{E}[U^2 \mid U > m] = 1 + m\,r(m)$, so

$$\mathrm{Var}[U \mid U > m] = 1 + m\,r(m) - r(m)^2 = 1 - r(m)\big(r(m) - m\big) > 0,$$

the variance of a non-degenerate distribution. Hence $r(r - m) < 1$ and $r - m$ is strictly decreasing. For the expansion, the tail series $\bar{\Phi}(m) = \varphi(m)\big(m^{-1} - m^{-3} + 3m^{-5} - O(m^{-7})\big)$ gives

$$\begin{aligned} r(m) &= \frac{m}{1 - m^{-2} + 3m^{-4} + O(m^{-6})} \\ &= m\Big(1 + \big(m^{-2} - 3m^{-4}\big) + \big(m^{-2} - 3m^{-4}\big)^2 + O(m^{-6})\Big) \\ &= m + m^{-1} - 2m^{-3} + O(m^{-5}), \end{aligned}$$

which is the first expansion, and multiplying,

$$\begin{aligned} r(m)\big(r(m) - m\big) &= \big(m + m^{-1} - 2m^{-3}\big)\big(m^{-1} - 2m^{-3}\big) + O(m^{-4}) \\ &= 1 - 2m^{-2} + m^{-2} + O(m^{-4}) = 1 - m^{-2} + O(m^{-4}). \end{aligned}$$

□

Proposition A.8 (fast drift: the learning rate at full misalignment)**.** *At $\rho = 0$ the threshold balance has the unique solution*

$$\sqrt{Q^*}(m) = \frac{\varphi(0)\,\bar{\Phi}(m)}{M(m)} = \frac{\varphi(0)}{r(m) - m} \quad (30)$$

at every margin, and the fixed-point learning rate $\eta^*(m, 0)$ tends to $2/\pi = 4\varphi(0)^2$ as $m \to \infty$.

Proof. By (28) and (22), $\mathbb{E}[\chi] = \varphi(0)\bar{\Phi}(m) - \sqrt{Q}\,M(m)$. Setting $\mathbb{E}[\chi] = 0$ gives (30), with $M(m) = \bar{\Phi}(m)\big(r(m) - m\big)$. By Lemma A.7 the balance exists at every margin. Abbreviate

$r = r(m)$ and $\bar{\Phi} = \bar{\Phi}(m)$, and insert $\sqrt{Q^*} = \varphi(0)/(r-m)$ into (24) and (25) with the atoms (28):

$$\begin{aligned}
\mathbb{E}[\chi x] &= \sqrt{Q^*}\,\varphi(0)\,\varphi(m) - Q^*\,\bar{\Phi} \\
&= \varphi(0)^2\,\bar{\Phi}\left[\frac{r}{r-m} - \frac{1}{(r-m)^2}\right] \\
&= \varphi(0)^2\,\bar{\Phi}\,\frac{r(r-m)-1}{(r-m)^2}, \qquad (31)\\
\mathbb{E}[\chi^2] &= \tfrac{1}{2}\bar{\Phi} - 2\sqrt{Q^*}\,\varphi(0)\big(\varphi(m) - m\,\bar{\Phi}\big) + Q^*\big[(1+m^2)\bar{\Phi} - m\,\varphi(m)\big] \\
&= \bar{\Phi}\left[\tfrac{1}{2} - 2\sqrt{Q^*}\,\varphi(0)\,(r-m) + Q^*\,(1+m^2-m\,r)\right] \\
&= \bar{\Phi}\left[\tfrac{1}{2} - 2\varphi(0)^2 + \varphi(0)^2\,\frac{1+m^2-m\,r}{(r-m)^2}\right]. \qquad (32)
\end{aligned}$$

Since $r(r-m) < 1$ by Lemma A.7, $\mathbb{E}[\chi x] < 0$ and the learning rate of (5) is positive:

$$\eta^*(m,0) = -\frac{2\,\mathbb{E}[\chi x]}{\mathbb{E}[\chi^2]} = \frac{2\varphi(0)^2\,\big[1 - r(r-m)\big]}{\big(\frac{1}{2} - 2\varphi(0)^2\big)(r-m)^2 + \varphi(0)^2\,(1+m^2-m\,r)}. \qquad (33)$$

By (29),

$$\begin{aligned}
1 - r(r-m) &= m^{-2} + O(m^{-4}),\\
(r-m)^2 &= m^{-2} - 4m^{-4} + O(m^{-6}),\\
1 + m^2 - m\,r &= 1 + m^2 - m\big(m + m^{-1} - 2m^{-3} + O(m^{-5})\big) = 2m^{-2} + O(m^{-4}),
\end{aligned}$$

and multiplying numerator and denominator of (33) by m^2,

$$\begin{aligned}
\eta^*(m,0) &= \frac{2\varphi(0)^2\,\big[1 + O(m^{-2})\big]}{\big(\frac{1}{2} - 2\varphi(0)^2\big)\big[1 + O(m^{-2})\big] + 2\varphi(0)^2\big[1 + O(m^{-2})\big]}\\
&= \frac{2\varphi(0)^2}{\frac{1}{2}}\,\big[1 + O(m^{-2})\big] = 4\varphi(0)^2\,\big[1 + O(m^{-2})\big] = \frac{2}{\pi}\,\big[1 + O(m^{-2})\big],
\end{aligned}$$

since $\varphi(0)^2 = 1/(2\pi)$. □

Remark A.9 (reading of the envelope). The transition at $\rho = 0$ is continuous, given the monotonicity of $\eta^*(m,0)$ observed on the scanned range. Numerically $\eta^*(m,0)$ increases from 0.504 at $m = 1$ to 0.6357 at $m = 30$, and the value at $m = 30$ is within 2×10^{-3} of $2/\pi$ with a monotone approach. For $\eta < 2/\pi$ the balance is then at a finite margin $m^*(\eta)$, and $m^*(\eta) \to \infty$ continuously as $\eta \uparrow 2/\pi$. For $\eta > 2/\pi$ no balance exists, and in the integrated trajectories $d\theta/d\alpha > 0$ persists: the threshold runaway observed numerically. This continuity is consistent with the fold ending before $\delta = \infty$. The continuation locates that end numerically (Section 4.4).

Proposition A.10 (slow drift: the stationary branch and its asymptote)**.** *At $\delta = 0$ the fixed points other than the trained state form the one-parameter branch $C_0 = \{(m,\rho) : \mathbb{E}[\chi y] = 0\}$ at $Q = Q^*(m,\rho)$. Along C_0, as $m \to \infty$ with the alignment held in a compact interior interval, $\rho \to 1/\sqrt{2}$ and $\eta^* \to 2$.*

Proof. At $\delta = 0$ the alignment equation $\eta\,\mathbb{E}[\chi y] = \delta R$ degenerates to $\mathbb{E}[\chi y] = 0$ for $\eta > 0$. With the threshold and norm balances of Proposition 3 this defines C_0. Fix $\rho \in (0,1)$, with $s < 1$, and let $m \to \infty$. The tails $\varphi(m/s)$ and $\bar{\Phi}(m/s)$ are superpolynomially small relative to $\varphi(m)$, because $\varphi(m/s)/\varphi(m) = \exp\big(-\frac{1}{2}m^2(s^{-2}-1)\big)$ with $s < 1$. The tail $\bar{\Phi}(\rho m/s) \le \exp(-\frac{1}{2}\rho^2 m^2/s^2)$ is superpolynomially small in m and enters below only as a relative error on terms already proportional to $\varphi(m)$. We write ε for any quantity superpolynomially small in m, giving $\Phi(\rho m/s) = 1 - \varepsilon$ and $0 \le \bar{\Phi}(m) - L \le \bar{\Phi}(\rho m/s)\,\bar{\Phi}(m) = \varepsilon\,\bar{\Phi}(m)$. With the tail series $\bar{\Phi}(m) = \varphi(m)\big(m^{-1} - m^{-3} +$

$3m^{-5} + O(m^{-7})\big)$ the atoms of Lemma A.1 reduce to

$$\begin{aligned} G(m,\rho) &= \rho\,\varphi(m)\,(1+\varepsilon),\\ M(m) &= \varphi(m) - m\,\bar{\Phi}(m) = \varphi(m)\,\big(m^{-2} - 3m^{-4} + O(m^{-6})\big),\\ \mathbb{E}[uy\,\mathbb{1}_b] &= \rho L + \rho m\,\varphi(m)(1+\varepsilon) + s\,\varphi(0)\,\varphi(m/s)\\ &= \rho\,\varphi(m)\,\big(m + m^{-1} - m^{-3} + O(m^{-5})\big),\\ \mathbb{E}[y^2\mathbb{1}_b] &= L + \rho^2 m\,\varphi(m)(1+\varepsilon) + \rho s\,\varphi(0)\,\varphi(m/s)\\ &= \varphi(m)\,\big(\rho^2 m + m^{-1} - m^{-3} + O(m^{-5})\big). \end{aligned}$$

Every $O(\cdot)$ and every ε here and below is uniform in ρ on any compact interior interval, because $s(\rho)$ stays below 1 there and the coefficients are continuous in ρ. Threshold balance then forces a norm of order m^4,

$$\begin{aligned} \sqrt{Q^*} &= \frac{G(m,\rho)}{M(m)} = \frac{\rho\,(1+\varepsilon)}{m^{-2} - 3m^{-4} + O(m^{-6})} = \rho m^2\,\big(1 + 3m^{-2} + O(m^{-4})\big),\\ Q^* &= \rho^2 m^4\,\big(1 + 6m^{-2} + O(m^{-4})\big). \end{aligned}$$

For the alignment moment (23), with $R^* = \rho\sqrt{Q^*}$,

$$\begin{aligned} R^*\,\bar{\Phi}(m) &= \rho^2 m^2\big(1 + 3m^{-2} + O(m^{-4})\big)\,\varphi(m)\,\big(m^{-1} - m^{-3} + O(m^{-5})\big)\\ &= \rho^2\,\varphi(m)\,\big(m + 2m^{-1} + O(m^{-3})\big),\\ \mathbb{E}[\chi y] &= \mathbb{E}[y^2\mathbb{1}_b] - R^*\,\bar{\Phi}(m)\\ &= \varphi(m)\,\big(\rho^2 m + m^{-1} - \rho^2 m - 2\rho^2 m^{-1}\big) + O\big(\varphi(m) m^{-3}\big)\\ &= \frac{\varphi(m)}{m}\,\big(1 - 2\rho^2\big) + O\big(\varphi(m)\,m^{-3}\big). \end{aligned} \tag{34}$$

Membership of C_0 requires $\mathbb{E}[\chi y] = 0$; hence $1 - 2\rho^2 = O(m^{-2})$ and $\rho \to 1/\sqrt{2}$. For the learning rate, by (24),

$$\begin{aligned} \sqrt{Q^*}\,\mathbb{E}[uy\,\mathbb{1}_b] &= \rho^2\,\varphi(m)\,m^2\big(1 + 3m^{-2} + O(m^{-4})\big)\big(m + m^{-1} + O(m^{-3})\big)\\ &= \rho^2\,\varphi(m)\,\big(m^3 + 4m + O(m^{-1})\big),\\ Q^*\,\bar{\Phi}(m) &= \rho^2\,\varphi(m)\,m^4\big(1 + 6m^{-2} + O(m^{-4})\big)\big(m^{-1} - m^{-3} + O(m^{-5})\big)\\ &= \rho^2\,\varphi(m)\,\big(m^3 + 5m + O(m^{-1})\big),\\ \mathbb{E}[\chi x] &= \sqrt{Q^*}\,\mathbb{E}[uy\,\mathbb{1}_b] - Q^*\,\bar{\Phi}(m) = -\rho^2 m\,\varphi(m)\,\big(1 + O(m^{-2})\big), \end{aligned} \tag{35}$$

and by (25), using $\mathbb{E}[uy\,\mathbb{1}_b] - m\,G(m,\rho) = \rho\,\varphi(m)\big(m^{-1} + O(m^{-3})\big)$ and

$$\begin{aligned} (1+m^2)\bar{\Phi}(m) - m\,\varphi(m) &= \varphi(m)\Big[(1+m^2)\big(m^{-1} - m^{-3} + 3m^{-5} + O(m^{-7})\big) - m\Big]\\ &= \varphi(m)\,\big(2m^{-3} + O(m^{-5})\big), \end{aligned}$$

the three parts of $\mathbb{E}[\chi^2]$ are

$$\begin{aligned} \mathbb{E}[y^2\mathbb{1}_b] &= \varphi(m)\,\big(\rho^2 m + O(m^{-1})\big),\\ -2\sqrt{Q^*}\big(\mathbb{E}[uy\,\mathbb{1}_b] - m\,G\big) &= -2\rho^2 m^2\big(1 + O(m^{-2})\big)\,\varphi(m)\big(m^{-1} + O(m^{-3})\big)\\ &= -2\rho^2 m\,\varphi(m)\,\big(1 + O(m^{-2})\big),\\ Q^*\big[(1+m^2)\bar{\Phi}(m) - m\,\varphi(m)\big] &= \rho^2 m^4\big(1 + O(m^{-2})\big)\,\varphi(m)\,\big(2m^{-3} + O(m^{-5})\big)\\ &= 2\rho^2 m\,\varphi(m)\,\big(1 + O(m^{-2})\big), \end{aligned}$$

which sum to

$$\mathbb{E}[\chi^2] = \rho^2 m\,\varphi(m)\,\big(1 - 2 + 2\big)\big(1 + O(m^{-2})\big) = \rho^2 m\,\varphi(m)\,\big(1 + O(m^{-2})\big). \tag{36}$$

Dividing (35) by (36),

$$\eta^* = -\frac{2\,\mathbb{E}[\chi x]}{\mathbb{E}[\chi^2]} = 2\,\big(1 + O(m^{-2})\big) \;\longrightarrow\; 2 \qquad \text{for every fixed } \rho \in (0,1),$$

uniformly on any compact interior interval of ρ as noted above, hence along C_0 wherever its alignment stays in one. □

Remark A.11 (numerical checks of the two limits and the anchor). A numerical check gives the next order of the alignment limit, with $\rho_0(m)$ the alignment of C_0 at margin m: $(\rho_0(m) - 1/\sqrt{2})\, m^2 \to 0.700 \approx 1/\sqrt{2}$. At every checked $m \in \{5, 8, 12, 17, 25\}$ the residual $|\eta^* - 2|$ on C_0 stays below 10^{-8}; it is 7×10^{-10} at $m = 5$ and reaches 3×10^{-13} by $m = 8$, consistent with a superpolynomial approach.

Our anchor η_0 is the interior maximum of η^* along C_0, located by solving $\{\mathbb{E}[\chi y] = 0,\ \frac{d}{dm}\eta^*|_{C_0} = 0\}$: $\eta_0 = 2.0223669083$ at $(m^\dagger, \rho^\dagger) = (0.10381538,\ 0.97365347)$, stable to 10^{-6} under refinement.

Numerically the fold converges to this maximum as $\delta \to 0$: the fold of Ψ merges into C_0 with $\eta \to \eta_0$, and the per-δ boundary values rise monotonically through 1.99909 ($\delta = 10^{-3}$), 2.01883 (10^{-4}), 2.02225 (3×10^{-6}).

A.6 The small-learning-rate limit at fixed drift

As $\eta \to 0$ at fixed $\delta > 0$ the order-parameter equations reduce to a two-dimensional slow system; with the reduced trajectory enclosed by interval arithmetic, the unit never enters dormancy. Section 4.5 reports that below $\eta \approx 0.3$ the order-parameter equations from the trained state settle into an active state at every scanned drift speed. In the rescaled time $\tilde{\alpha} = \eta\alpha$ the order-parameter equations (2)–(4) become

$$\frac{dR}{d\tilde{\alpha}} = \mathbb{E}[\chi y] - \frac{\delta}{\eta}\, R, \qquad \frac{dQ}{d\tilde{\alpha}} = 2\,\mathbb{E}[\chi x] + \eta\,\mathbb{E}[\chi^2], \qquad \frac{d\theta}{d\tilde{\alpha}} = -\mathbb{E}[\chi], \tag{37}$$

so as $\eta \to 0$ the overlap is a fast variable that relaxes at rate δ/η to $R = (\eta/\delta)\,\mathbb{E}[\chi y] + O(\eta^2)$, and the pair (Q, θ) is slow. At $R = 0$ the moments are the elementary atoms (28), and the slow pair obeys the reduced system

$$\begin{aligned} \frac{dQ}{d\tilde{\alpha}} &= 2\Big[\sqrt{Q}\,\varphi(0)\,\varphi(m) - Q\,\bar{\Phi}(m)\Big], \\ \frac{d\theta}{d\tilde{\alpha}} &= -\Big[\varphi(0)\,\bar{\Phi}(m) - \sqrt{Q}\,\big(\varphi(m) - m\,\bar{\Phi}(m)\big)\Big], \qquad m = \theta/\sqrt{Q}, \end{aligned} \tag{38}$$

which contains neither η nor δ.

Lemma A.12 (the reduced trajectory on a finite horizon)**.** *The solution of* (38) *from the trained state* $(Q, \theta) = (1, 0)$ *obeys*

$$m \le 0.35, \qquad 10^{-5} \le Q \le 1, \qquad |\theta| \le 0.42 \qquad \text{on } [0, 15],$$

and at $\tilde{\alpha} = 15$

$$10^{-4} \le Q \le 10^{-3}, \qquad -0.42 \le \theta \le -0.30.$$

Proof. One of the six is analytic. Since $\mathbb{E}[\tau \mathbb{1}] \le \mathbb{E}[\tau] = \varphi(0)$ and $(x - \theta)^+ \ge x - \theta$ with $\mathbb{E}[x] = 0$, the threshold drift obeys $\mathbb{E}[\chi] \le \varphi(0) + \theta$. Wherever $\theta < -\varphi(0)$ the threshold therefore rises, and the solution never goes below $-\varphi(0) = -0.399$.

The rest come from a validated enclosure of the trajectory, carried in the variable $\ln\sqrt{Q}$, which holds the four decades of Q at one relative accuracy. Each step finds a box the solution cannot leave over the step, by checking that one Picard step maps the box into itself; the Lipschitz constant this needs is finite because the box stays in $\{Q > 0\}$. It encloses the flow of the step's center in a Taylor polynomial whose remainder is the p-th coefficient evaluated over that box. And it carries the enclosure of the initial data forward by $\exp(h\hat{J})$, where $\hat{J}$ takes the diagonal entries of the Jacobian above and the off-diagonal ones in modulus over the box. The comparison inequality behind the last step is elementary, and we state it in the form the next proof also needs. Let e be the difference of two solutions with data in the box. Componentwise the mean value theorem gives $e' = J(\xi)\, e$ with ξ in the box, which is convex, so

$$D^+|e_i| \le \hat{J}_{ii}\,|e_i| + \sum_{j \ne i} \hat{J}_{ij}\,|e_j|,$$

the diagonal entry taken signed because it multiplies $|e_i|$ itself and the others in modulus. If u solves $u' = \hat{J}u + \kappa\mathbf{1}$ from above $|e(0)|$, a first time at which some $|e_i|$ meets u_i contradicts the strict inequality between the two derivatives there. So $|e| \le u$, and $\kappa \downarrow 0$ gives $|e(h)| \le \exp(h\hat{J})\,|e(0)|$ componentwise. The same argument with a forcing term p and $e(0) = 0$ gives $|e(\tilde{\alpha})| \le \int_0^{\tilde{\alpha}} \Pi(\tilde{\alpha}, \sigma)\,|p(\sigma)|\,d\sigma$, where Π is the transition matrix of $u' = \hat{J}u$ and has nonnegative entries. Because every quantity is a ball with a rigorous radius, the output is an enclosure rather than an estimate.

At step 2×10^{-3} and order ten the enclosure gives $\sup_{[0,15]} m \le 0.31412$ and $\theta \in [-0.38957, 0.17070]$, verifies $dQ/d\tilde{\alpha} < 0$ at every step, so that $Q \le Q(0) = 1$ and $Q \ge Q(15)$ on the interval. It returns $Q(15) = 1.64670698 \times 10^{-4}$ and $\theta(15) = -0.38956516$ with radii below 2.4×10^{-23}. A run at step 5×10^{-2} and order eight, an independent choice of both parameters, returns wider enclosures that contain these. Since the six bounds are round numbers placed outside the enclosure rather than readings of it, each can be checked against the intervals just given; $Q \le 1$ is the initial value itself, and is attained only at $\tilde{\alpha} = 0$. The programs are in the supplementary code. □

Beyond the horizon the margin decreases without bound while $\theta \to -\varphi(0)$ and $Q \to 0$. Since the solution already keeps $\theta \le -0.30$ and $Q \le 10^{-3}$ from $\tilde{\alpha} = 14.5$ on, the choice of end time is not delicate. Nothing below uses these last two readings.

Lemma A.13 (the remainder on the invariant box and its derivatives)**.** *Write $\mathbb{1}^c$ for the inactive event $\{x \le \theta\}$, $\tau = \max(y, 0)$, and*

$$\Delta_R = \mathbb{E}[\chi y] - \big(\tfrac{1}{2} - R\big), \qquad \Delta_Q = 2\,\mathbb{E}[\chi x] - (R - 2Q), \qquad \Delta_\theta = \big(\varphi(0) + \theta\big) - \mathbb{E}[\chi],$$

so that the right-hand sides of (2)–(4) are the linear-regression flow of the proof below, plus $\eta^2\mathbb{E}[\chi^2]$ in the norm equation, plus $\eta\Delta$. Then for every state with $Q > 0$,

$$\begin{aligned} \Delta_Q &= 2Q\,\Phi(m) - 2\,\mathbb{E}[\tau x \mathbb{1}^c], \\ \Delta_R &= R\,\Phi(m) - \mathbb{E}[\tau y \mathbb{1}^c], \\ \Delta_\theta &= \theta\,\Phi(m) + \sqrt{Q}\,\varphi(m) + \mathbb{E}[\tau \mathbb{1}^c]. \end{aligned}$$

Let

$$A = \{(Q, R, \theta) :\ 0.1\eta \le Q \le 0.002,\ \ 0 \le R \le \eta/\delta,\ \ -0.45 \le \theta \le -0.25\},$$

which is non-empty only for $\eta \le 0.02$, and assume $\eta \le 0.0324\,\delta^2$, so that $\rho = R/\sqrt{Q} \le \sqrt{10\eta}/\delta \le 0.57$ and $s = \sqrt{1 - \rho^2} \ge 0.82$ on A. Then on A each component of Δ is at most $3\,\varphi(m)$ in absolute value, and each first derivative of Δ in Q, R and θ is at most, again in absolute value,

$$g(Q, \theta) = \varphi(m)\,(1 + |m|^3)\,(1 + 1/\sqrt{Q}), \qquad m = \theta/\sqrt{Q}.$$

Proof. For any correlation between x and y, $\mathbb{E}[\tau] = \varphi(0)$, $\mathbb{E}[\tau y] = \frac{1}{2}$ and $\mathbb{E}[\tau x] = R/2$, while the atoms of Lemma A.1 give

$$\begin{aligned} \mathbb{E}[(x - \theta)^+] &= \sqrt{Q}\,\varphi(m) - \theta\,\bar{\Phi}(m), \\ \mathbb{E}[(x - \theta) y \mathbb{1}] &= R\,\bar{\Phi}(m), \qquad \mathbb{E}[(x - \theta) x \mathbb{1}] = Q\,\bar{\Phi}(m). \end{aligned}$$

Substituting $\chi = \tau\mathbb{1} - (x - \theta)^+$ and $\mathbb{E}[\tau\mathbb{1}] = \varphi(0) - \mathbb{E}[\tau\mathbb{1}^c]$ with its two analogues into the three definitions gives the three displayed expressions, for every state with $Q > 0$.

Write $x = \sqrt{Q}\,(\rho u + s z)$ with $u = y$ and z independent standard Gaussians, so that $\mathbb{1}^c = \{\rho u + s z \le m\}$ and, with $c(u) = (m - \rho u)/s$,

$$\mathbb{E}[\tau \mathbb{1}^c] = I_1, \qquad \mathbb{E}[\tau y \mathbb{1}^c] = I_2, \qquad \mathbb{E}[\tau x \mathbb{1}^c] = \sqrt{Q}\,I_3,$$

where I_1, I_2, I_3 depend on (m, ρ) alone:

$$\begin{aligned} I_1 &= \int_0^\infty u\,\varphi(u)\,\Phi(c(u))\,du, \qquad I_2 = \int_0^\infty u^2\,\varphi(u)\,\Phi(c(u))\,du, \\ I_3 &= \int_0^\infty u\,\varphi(u)\,\big[\rho u\,\Phi(c(u)) - s\,\varphi(c(u))\big]\,du. \end{aligned}$$

Since $\rho \geq 0$, $s \leq 1$ and $m < 0$, for every $u \geq 0$

$$c(u) \leq m - \rho u \leq m < 0,$$

so $\Phi(c(u)) \leq \Phi(m)$, $\varphi(c(u)) \leq \varphi(m)$ and $|c(u)| \geq |m|$. With $\int_0^\infty u\,\varphi(u)\,du = \varphi(0)$, $\int_0^\infty u^2\,\varphi(u)\,du = \frac{1}{2}$ and Mills' inequality $\Phi(m) \leq \varphi(m)/|m|$, and using $|m| \geq 5.59$, $Q \leq 0.002$ and $|R| \leq \sqrt{Q}$ on A,

$$I_1 \leq \varphi(0)\,\Phi(m), \qquad I_2 \leq \tfrac{1}{2}\,\Phi(m), \qquad |I_3| \leq \tfrac{\rho}{2}\,\Phi(m) + \varphi(0)\,\varphi(m) \leq 0.5\,\varphi(m),$$

and each component of Δ is at most $3\,\varphi(m)$.

All six partial derivatives have closed forms. That is shorter than estimating them under the integral sign and avoids the factor $1/s^2$ that differentiating c in ρ produces. Write $v = \rho u + sz$, so that (u, v) is standard bivariate normal with correlation ρ and

$$I_1 = \mathbb{E}[u^+ \mathbb{1}\{v \leq m\}], \qquad I_2 = \mathbb{E}[(u^+)^2 \mathbb{1}\{v \leq m\}], \qquad I_3 = \mathbb{E}[u^+ v\,\mathbb{1}\{v \leq m\}].$$

Each is $\int_{-\infty}^{m} \mathbb{E}[\,\cdot \mid v\,]\,\varphi(v)\,dv$, so the derivative in m is the conditional expectation at $v = m$ times $\varphi(m)$. There $u \sim N(\rho m, s^2)$. For the derivative in ρ, the bivariate normal density obeys $\partial_\rho \varphi_2 = \partial_u \partial_v \varphi_2$, and two integrations by parts move both derivatives onto the integrand. With

$$H = \rho m\,\Phi(\rho m/s) + s\,\varphi(\rho m/s), \qquad H_2 = (\rho^2 m^2 + s^2)\,\Phi(\rho m/s) + \rho m\, s\,\varphi(\rho m/s),$$

the two conditional moments $\mathbb{E}[u^+ \mid v = m]$ and $\mathbb{E}[(u^+)^2 \mid v = m]$, this gives

$$\begin{aligned}
&\partial_m I_1 = \varphi(m)\,H, && \partial_\rho I_1 = -\varphi(m)\,\Phi(\rho m/s),\\
&\partial_m I_2 = \varphi(m)\,H_2, && \partial_\rho I_2 = -2\,\varphi(m)\,H,\\
&\partial_m I_3 = m\,\varphi(m)\,H, && \partial_\rho I_3 = \Pr(u > 0,\, v \leq m) - m\,\varphi(m)\,\Phi(\rho m/s).
\end{aligned}$$

On A the mean ρm is at most zero, and H and H_2 increase in it, so $H \leq s\,\varphi(0) \leq 0.4$ and $H_2 \leq s^2/2 \leq \frac{1}{2}$. With Mills' inequality $\Phi(m) \leq \varphi(m)/|m|$ and $|m| \geq 5.59$ the six derivatives are at most $0.4\,\varphi(m)$, $\varphi(m)$, $0.5\,\varphi(m)$, $0.8\,\varphi(m)$, $0.4\,|m|\,\varphi(m)$ and $(|m| + 1/|m|)\,\varphi(m)$ in absolute value, and in particular

$$\begin{aligned}
&|\partial_m I_1| \leq 0.6\,\varphi(m), && |\partial_\rho I_1| \leq (1 + \tfrac{1}{2}|m|)\,\varphi(m),\\
&|\partial_m I_2| \leq 0.8\,\varphi(m), && |\partial_\rho I_2| \leq (1 + \tfrac{1}{2}|m|)\,\varphi(m),\\
&|\partial_m I_3| \leq (\tfrac{1}{2} + \tfrac{1}{2}|m|)\,\varphi(m), && |\partial_\rho I_3| \leq (1 + |m|)\,\varphi(m).
\end{aligned}$$

The parameters reach Δ through m and ρ and through the explicit factors Q, R and $\sqrt{Q}$, with

$$\partial_Q m = -\frac{m}{2Q}, \quad \partial_\theta m = \frac{1}{\sqrt{Q}}, \quad \partial_Q \rho = -\frac{\rho}{2Q}, \quad \partial_R \rho = \frac{1}{\sqrt{Q}},$$

and all others zero. Because every substitution below uses $1/Q = m^2/\theta^2$ and $1/\sqrt{Q} = |m|/|\theta|$, $|\theta|$ is never replaced by a bound at one end of its range and by a bound at the other in the same estimate. Those same two substitutions also give $g \geq \varphi(m)\,m^4/|\theta|$. Collecting terms gives all nine derivatives the same way. For $\partial_Q \Delta_\theta$,

$$\begin{aligned}
|\partial_Q \Delta_\theta| &\leq \frac{|\theta|\,|m|\,\varphi(m)}{2Q} + \frac{(1 + m^2)\,\varphi(m)}{2\sqrt{Q}} + \frac{0.6\,|m|\,\varphi(m)}{2Q} + \frac{0.57\,(1 + \frac{1}{2}|m|)\,\varphi(m)}{2Q}\\
&\leq \varphi(m)\,|m|^3 \Big(\frac{1.02}{|\theta|} + \frac{0.494}{\theta^2}\Big) \leq \varphi(m)\,\frac{m^4}{|\theta|} \cdot \frac{1}{|m|}\Big(1.02 + \frac{0.494}{|\theta|}\Big) \leq 0.54\,g,
\end{aligned}$$

using $|m| \geq 5.59$ and $|\theta| \geq 0.25$ in the last step. The same accounting gives $|\partial_Q \Delta_R| \leq 0.45\,g$ and $|\partial_Q \Delta_Q| \leq 0.13\,g$, and the derivatives in R and in θ carry $1/\sqrt{Q}$ rather than $1/Q$ and are smaller by a factor $|m|$. Every first derivative of Δ is therefore at most g on A. □

Proposition A.14 (small learning rate at fixed drift)**.** *For every $\delta > 0$ there is $\eta_{\text{low}}(\delta) > 0$ such that for $0 < \eta < \eta_{\text{low}}$ the order-parameter equations (2)–(4) from the trained state satisfy $m(\alpha) < m_\star$ for all $\alpha \geq 0$: the unit never enters dormancy, at any horizon. The trajectory also converges to a fixed point that lies within 3×10^{-4} in each coordinate of the reference point $R_\infty = \eta/(2(\delta + \eta))$, $\theta_\infty = -\varphi(0)$ and $Q_\infty = R_\infty/2 + \eta\,\mathbb{E}[\chi^2]/2$ with the second moment taken at that same point, whose squared weight length and overlap are $O(\eta)$ at fixed δ. At that fixed point the margin is below -3: an active unit whose output stays near the mean teacher output and varies little with the input.*

Proof. Write $\epsilon = \eta/\delta$ and $\varphi(0) = 1/\sqrt{2\pi} = 0.399$. The argument has three parts: the trajectory reaches a small box in finite rescaled time, the box is forward invariant for small η, and the flow inside the box is a contraction whose fixed point is close to the reference point of the statement.

The moments are defined where $R^2 \le Q$, a region whose boundary contains the trained state $(1, 1, 0)$. No open neighborhood of it therefore carries the vector field, and a singular-perturbation theorem cannot be applied at the initial time. That boundary is repelling: on $R = \pm\sqrt{Q}$ the two Gaussians are proportional, and $d(Q - R^2)/d\tilde{\alpha} = \eta\,\mathbb{E}[\chi^2] + 2Q/\epsilon > 0$, which at the trained state is $2/\epsilon$. A solution from that state nonetheless exists: extend the field continuously to a neighborhood by clamping ρ to $[-1, 1]$, take a Peano solution of the extension, and note that the computation above makes $Q - R^2$ increase at once. That solution therefore enters the interior immediately and solves the original system for $\tilde{\alpha} > 0$. We bound the overlap directly and carry it as a perturbation of (38), which also makes the threshold on η explicit.

At $R = 0$ the Gaussians x and y are independent, so $\mathbb{E}[\chi y] = \frac{1}{2}\,\bar{\Phi}(m) > 0$ at every state with $Q > 0$. At a first zero of R the derivative would therefore be positive. Since $R(0) = 1$, the overlap stays non-negative. For $R \ge 0$ we have $\mathbb{E}[\chi y] = \mathbb{E}[\tau y \mathbb{1}] - R\,\bar{\Phi}(m) \le \mathbb{E}[\tau y] = \frac{1}{2}$, so $dR/d\tilde{\alpha} \le \frac{1}{2} - R/\epsilon$ and

$$0 \le R(\tilde{\alpha}) \le \frac{\epsilon}{2} + \Big(1 - \frac{\epsilon}{2}\Big) e^{-\tilde{\alpha}/\epsilon} \le \frac{\epsilon}{2} + e^{-\tilde{\alpha}/\epsilon}.$$

Because the ratio $2e^{-15/\epsilon}/\epsilon$ increases in ϵ on $(0, 1/4]$ and equals 7×10^{-26} at $\epsilon = 1/4$, we have $e^{-15/\epsilon} \le 3.5 \times 10^{-26}\,\epsilon$ and $R(15) \le 0.51\,\epsilon$, half of what A allows.

Write (Q_*, θ_*) for the solution of (38) and $\Lambda = \ln\sqrt{Q}$. The overlap and the η^2 term are all that separate the slow pair (Λ, θ) of (37) from the reduced pair. Only $\mathbb{E}[\tau \mathbb{1}]$ and $\mathbb{E}[\tau x \mathbb{1}]$ depend on R, and Price's theorem gives

$$\partial_R \mathbb{E}[\chi] = \frac{\varphi(m)}{\sqrt{Q}}\,\Phi(\rho m/s), \qquad \partial_R \mathbb{E}[\chi x] = \Pr(y > 0,\, x > \theta) + m\,\varphi(m)\,\Phi(\rho m/s),$$

the first in $[0, \varphi(m)/\sqrt{Q}]$ and the second in $[-0.242, 1.242]$, since $|m\,\varphi(m)| \le \varphi(1)$. Both stay bounded as $|\rho| \to 1$, which is what the singular-perturbation route could not deliver at the initial time. The mean value theorem therefore gives, for every $|R| \le \sqrt{Q}$,

$$\big|\mathbb{E}[\chi x](Q, R, \theta) - \mathbb{E}[\chi x](Q, 0, \theta)\big| \le 1.25\,|R|, \qquad \big|\mathbb{E}[\chi](Q, R, \theta) - \mathbb{E}[\chi](Q, 0, \theta)\big| \le \frac{\varphi(m)}{\sqrt{Q}}\,|R|.$$

With $0 \le \mathbb{E}[\chi^2] \le S = \frac{1}{2} + Q + \theta^2 + 2\theta\varphi(0)$ for $R \ge 0$, the slow pair obeys $(\Lambda, \theta)' = f(\Lambda, \theta) + p$ with f the reduced field and

$$|p_\Lambda| \le \frac{1.25\,R}{Q} + \frac{\eta\,S}{2Q}, \qquad |p_\theta| \le \frac{\varphi(m)}{\sqrt{Q}}\,R.$$

For as long as the full solution stays in the tube $\{|\Lambda - \Lambda_*| \le 0.1,\ |\theta - \theta_*| \le 0.02\}$, the segment joining it to the reduced solution stays in the tube's cross-section, and the forced comparison inequality proved in Lemma A.12 gives componentwise

$$\big|(\Lambda, \theta) - (\Lambda_*, \theta_*)\big|(\tilde{\alpha}) \le \int_0^{\tilde{\alpha}} \Pi(\tilde{\alpha}, \sigma)\,|p(\sigma)|\,d\sigma,$$

with $\hat{J}$ now bounding the Jacobian over the cross-section. Substituting the overlap bound and $\int_0^{\tilde{\alpha}} e^{-\sigma/\epsilon}\,d\sigma \le \epsilon$ leaves a bound $\epsilon\,C_a + \eta\,C_b$ on that difference, with the two vectors evaluated over the tube in the same interval arithmetic as Lemma A.12,

$$C_a = (1.30 \times 10^4,\ 153), \qquad C_b = (1.45 \times 10^3,\ 58).$$

Both components fall below the radii of the tube once

$$\eta \le \frac{\delta}{1.3 \times 10^5 + 1.45 \times 10^4\,\delta},$$

the Λ component being the binding one. Since a first time at which the solution left the tube would contradict that strict inequality, it never leaves. Existence of a solution is enough here; uniqueness is not needed, because the estimate holds for every solution.

On the tube the margin is at most 0.39, far below $m_\star$. At $\tilde{\alpha} = 15$ the tube widens the $Q \in [10^{-4}, 10^{-3}]$ and $\theta \in [-0.42, -0.30]$ of Lemma A.12 to $Q \in [8.1 \times 10^{-5}, 1.3 \times 10^{-3}]$ and $\theta \in [-0.44, -0.28]$, and $R(15) \le 0.51\,\epsilon \le \epsilon$ from above. The whole state at $\tilde{\alpha} = 15$ therefore lies in A, because the constraint $Q \ge 0.1\eta$ asks only $\eta \le 8.1 \times 10^{-4}$, which the threshold above already gives.

We show that A is forward invariant for small η. In A the margin, the inactive fraction and the alignment obey

$$\begin{aligned} m &\le m_A = -0.25/\sqrt{0.002} = -5.59, \\ p &= \bar{\Phi}(5.59) < 1.2 \times 10^{-8}, \\ \rho &= R/\sqrt{Q} \le \sqrt{10\eta}/\delta. \end{aligned}$$

Write $\mathbb{1}$ for the activity indicator $\mathbb{1}[x > \theta]$, $\mathbb{1}^c$ for its complement, and $\chi = \tau\,\mathbb{1} - (x-\theta)^+$ with $\tau = \max(y, 0)$. Four identities hold for any correlation between x and y:

$$\mathbb{E}[\tau] = \varphi(0), \qquad \mathbb{E}[\tau y] = \tfrac{1}{2}, \qquad \mathbb{E}[\tau x] = R/2, \qquad \mathbb{E}[(x-\theta)x\mathbb{1}] = Q\,\bar{\Phi}(m).$$

The tail terms obey, by Cauchy–Schwarz with $\mathbb{E}[\tau^2] = \frac{1}{2}$ and $\mathbb{E}[\tau^2 y^2] \le 3$,

$$\begin{aligned} \mathbb{E}[\tau\mathbb{1}^c] &\le \sqrt{p/2} < 10^{-4}, \\ \mathbb{E}[\tau y\mathbb{1}^c] &\le \sqrt{3p} < 2 \times 10^{-4}, \\ \mathbb{E}[x^2\mathbb{1}^c] &= Q\left(\bar{\Phi}(|m|) + |m|\,\varphi(|m|)\right) < 4 \times 10^{-7}\,Q, \\ \mathbb{E}[\tau x\mathbb{1}^c] &\le \sqrt{\tfrac{1}{2}\,\mathbb{E}[x^2\mathbb{1}^c]} < 4.5 \times 10^{-4}\sqrt{Q} \le 2.1 \times 10^{-5}, \end{aligned}$$

and on the active side

$$\mathbb{E}[(x-\theta)^+] = \sqrt{Q}\left(\varphi(m) - m\,\bar{\Phi}(m)\right) \le |\theta| + \sqrt{Q}\,\varphi(m) \le |\theta| + 10^{-8}.$$

On the face $\theta = -0.25$,

$$\begin{aligned} \mathbb{E}[\chi] &\ge \varphi(0) - 10^{-4} - 0.25 > 0.14, \\ \frac{d\theta}{d\alpha} &= -\eta\,\mathbb{E}[\chi] < 0. \end{aligned}$$

On the face $\theta = -0.45$, since $(x-\theta)^+ \ge x - \theta$ and $\mathbb{E}[x] = 0$,

$$\begin{aligned} \mathbb{E}[\chi] &\le \varphi(0) - \mathbb{E}[(x-\theta)^+] \le 0.399 - 0.45 < 0, \\ \frac{d\theta}{d\alpha} &> 0. \end{aligned}$$

The same two inequalities hold at every state with $\theta < -\varphi(0)$, so the half-space $\theta \ge -\varphi(0)$ is forward invariant for every η and δ: below $-\varphi(0)$ the threshold rises. On the face $Q = 0.002$, using $\mathbb{E}[\chi^2] \le 3\,(\mathbb{E}[\tau^2] + Q + \theta^2) \le 2.2$,

$$\begin{aligned} \mathbb{E}[\chi x] &= \mathbb{E}[\tau x\mathbb{1}] - Q\,\bar{\Phi}(m) \\ &\le R/2 + 2.1 \times 10^{-5} - 0.002\,(1-p) \\ &\le \frac{\eta}{2\delta} - 0.00197, \\ \frac{dQ}{d\alpha} &= 2\eta\,\mathbb{E}[\chi x] + \eta^2\,\mathbb{E}[\chi^2] \\ &\le \eta\Big(\frac{\eta}{\delta} - 0.00394\Big) + 2.2\,\eta^2 < 0 \qquad \text{for } \eta < \frac{0.0039}{1/\delta + 2.2}. \end{aligned}$$

On the face $Q = 0.1\eta$ the inactive probability is $p = \bar{\Phi}(0.25/\sqrt{0.1\eta}) \le \exp(-0.31/\eta)$. Since A is non-empty only for $\eta \le 0.02$, and since $\mathbb{E}[(\tau - x + \theta)^4] \le 27\,(\mathbb{E}[\tau^4] + \mathbb{E}[x^4] + \theta^4) \le 42$ there,

$$\begin{aligned} \mathbb{E}[\tau x\mathbb{1}^c] &\le \sqrt{3Q}\,e^{-0.15/\eta} \le 0.01\,\eta, \\ \mathbb{E}[\chi^2] &\ge \mathbb{E}[(\tau - x + \theta)^2] - \mathbb{E}[(\tau - x + \theta)^2\mathbb{1}^c] \ge \mathrm{Var}(\tau) - R - 6.5\,e^{-0.155/\eta} \ge 0.25, \\ \frac{dQ}{d\alpha} &= 2\eta\,\mathbb{E}[\chi x] + \eta^2\,\mathbb{E}[\chi^2] \ge 2\eta\Big(\frac{R}{2} - Q - 0.01\eta\Big) + 0.25\,\eta^2 \ge 0.03\,\eta^2 > 0, \end{aligned}$$

the second line using $\mathrm{Var}(\tau) = \frac{1}{2} - \varphi(0)^2$, $R \le \sqrt{Q} \le 0.045$ and Cauchy–Schwarz on the inactive event. On the face $R = \eta/\delta$, Lemma A.13 gives $\mathbb{E}[\chi y] = \frac{1}{2} - R\,\bar{\Phi}(m) - \mathbb{E}[\tau y \mathbb{1}^c] \le \frac{1}{2}$, so

$$\frac{dR}{d\alpha} = \eta\,\mathbb{E}[\chi y] - \delta R \le \frac{\eta}{2} - \eta = -\frac{\eta}{2} < 0.$$

On the face $R = 0$, $\rho = 0$ and $\mathbb{E}[(x-\theta)y\mathbb{1}] = 0$, so that

$$\mathbb{E}[\chi y] \ge \tfrac{1}{2} - 2 \times 10^{-4} > 0,$$
$$\frac{dR}{d\alpha} > 0.$$

Every face is therefore crossed inward for $\eta < \eta_1(\delta) = 0.0039/(1/\delta + 2.2)$, and the trajectory stays in A for all later times, where $m \le -5.59 < m_\star$.

We show that the flow on A is a contraction. In A write the vector field as $L + \eta\,\Delta(Q, R, \theta)$, where L collects the linear-regression flow and the η^2 term:

$$\frac{dR}{d\alpha} = \eta\,(\tfrac{1}{2} - R) - \delta R + \eta\,\Delta_R,$$
$$\frac{dQ}{d\alpha} = \eta\,(R - 2Q) + \eta^2\,\mathbb{E}[\chi^2] + \eta\,\Delta_Q,$$
$$\frac{d\theta}{d\alpha} = -\eta\,(\varphi(0) + \theta) + \eta\,\Delta_\theta.$$

Lemma A.13 writes Δ out and bounds each of its components by $3\,\varphi(m)$ and each of its first derivatives in Q, R and θ by g, with

$$g(Q,\theta) = \varphi(m)\,(1 + |m|^3)\,(1 + 1/\sqrt{Q}), \qquad m = \theta/\sqrt{Q}.$$

Substituting $1/\sqrt{Q} = |m|/|\theta|$ writes g as $\varphi(m)\,(1+|m|^3)\,(1 + |m|/|\theta|)$, which at fixed $|m|$ decreases in $|\theta|$. Because $|\theta| \ge 0.25$ on A, this gives $g \le k(|m|)$ with $k(a) = \varphi(a)(1+a^3)(1+4a)$. The logarithmic derivative of k is $-a + 3a^2/(1+a^3) + 4/(1+4a)$, which is below $-a + 4/a$ and so negative for $a > 2$. Since $|m| \ge a_0 = 0.25/\sqrt{0.002} = 5.5902$ on A, the supremum sits at a_0, the corner with the largest Q and the smallest $|\theta|$:

$$\sup_A g = g(0.002, -0.25) < 2.7 \times 10^{-4},$$

a bound that does not depend on η. For L the Jacobian is

$$\mathrm{diag}\big(-(\delta+\eta),\, -2\eta,\, -\eta\big)$$

plus the off-diagonal entry $\partial_R(dQ/d\alpha) = \eta$ and the three derivatives of $\eta^2\,\mathbb{E}[\chi^2]$, each of which is at most $2\eta^2$ in absolute value. For those three, start from the exact identity

$$\mathbb{E}[\chi^2] = \mathrm{Var}(\tau) + (\varphi(0) + \theta)^2 + Q - R - F, \qquad F = \mathbb{E}[(\tau - x + \theta)^2 \mathbb{1}^c].$$

The first four terms contribute 1, -1 and $2(\varphi(0)+\theta)$, and the last of these is at most 0.3 in absolute value on A. Of the three parameters only θ moves the inactive event, and the integrand of F does not vanish on its boundary, so

$$\partial_\theta F = 2\,\mathbb{E}[(\tau - x + \theta)\mathbb{1}^c] + \mathbb{E}[\tau^2 \mid x = \theta]\,\frac{\varphi(m)}{\sqrt{Q}}.$$

The derivatives in Q and in R act on the law of x rather than on the event, and in the standardized variable $u = x/\sqrt{Q}$ they carry the chain factors $\partial_Q m = -m/(2Q)$ and $\partial_R \rho = 1/\sqrt{Q}$ of Lemma A.13. The derivative in Q is therefore the largest of the three, by the same factor $|m|/(2\sqrt{Q})$ that separates $1/Q$ from $1/\sqrt{Q}$ in that lemma. Its leading part is $\mathbb{E}[\tau^2 \mid x = \theta]\,\varphi(m)\,|m|/(2Q)$. On A we have $\mathbb{E}[\tau^2 \mid x = \theta] \le \frac{1}{2}$, attained at $\rho = 0$. Because $a^3\varphi(a)$ decreases for $a \ge \sqrt{3}$, we also have $\varphi(m)\,|m|/(2Q) = \varphi(m)\,|m|^3/(2\theta^2) \le 8\,|m|^3\varphi(m) \le 8a_0^3\varphi(a_0) = 9.13 \times 10^{-5}$. That part of the derivative is therefore below 4.6×10^{-5}. Every other part of every one of the three derivatives carries $\varphi(m)/\sqrt{Q} = \varphi(m)|m|/|\theta| \le 4a_0\varphi(a_0) = 1.46 \times 10^{-6}$, by the same argument with $a\,\varphi(a)$ in place of $a^3\varphi(a)$, or one of the tail bounds of Lemma A.13. The largest of those bounds enters

as $2\,\mathbb{E}[\tau \mathbb{1}^c] \le 2\sqrt{p/2} < 2 \times 10^{-4}$. All three derivatives of F are therefore below 10^{-3}, and each derivative of $\mathbb{E}[\chi^2]$ is within that of 1, -1 or $2(\varphi(0) + \theta)$. Hence the bound of 2 used above holds with a factor two to spare. In the maximum norm with equal weights on R and Q the logarithmic norm of the Jacobian of L is at most $-\eta + 6\eta^2$, because the row for Q carries the off-diagonal η and all three derivatives of $\eta^2\mathbb{E}[\chi^2]$. Because every entry of $D\Delta$ is bounded by g, its induced (row-sum) norm is at most $3 \sup_A g < 8.1 \times 10^{-4}$, and the remainder adds at most $8.1 \times 10^{-4}\,\eta$. For $\eta < 0.04$ the flow on A is therefore a contraction with rate at least 0.75η, so it has a unique fixed point in A to which every trajectory in A converges. Write p_* for the state at which L vanishes:

$$R_\infty = \frac{\eta}{2(\delta + \eta)}, \qquad \theta_\infty = -\varphi(0),$$
$$Q_\infty = \frac{R_\infty}{2} + \frac{\eta}{2}\,\mathbb{E}[\chi^2](Q_\infty, R_\infty, \theta_\infty).$$

The last line is an equation in Q_∞ rather than a formula for it. On A we have $0.25 \le \mathbb{E}[\chi^2] \le 2.2$, since the identity above gives $\mathbb{E}[\chi^2] \ge \mathrm{Var}(\tau) - \sqrt{Q} - F \ge 0.29$ and $\mathbb{E}[\chi^2] \le \mathrm{Var}(\tau) + (\varphi(0) + \theta)^2 + Q \le 0.37$ there. That right-hand side therefore takes values in $[0.125\,\eta,\ \eta(1/(4\delta) + 1.1)]$, and over the drift speeds and rates that η_{low} allows the upper end is at most 1.25×10^{-3}. That range sits inside the $[0.1\eta,\ 0.002]$ of A, and the right-hand side has Q-derivative $\eta/2$ times a quantity of modulus at most 2. For $\eta < 1$ it therefore maps the interval into itself and contracts it. The equation therefore has exactly one solution, and p_* lies in A: its threshold is -0.399 and its overlap is between 0 and η/δ. The true fixed point lies within

$$\frac{\|\eta\Delta\|}{0.75\eta} = \frac{\|\Delta\|}{0.75}$$

of it in the maximum norm. Because Lemma A.13 bounds $\|\Delta\|$ on A by $3\varphi(m) \le 2 \times 10^{-7}$, the distance is below 3×10^{-7} in each coordinate, well inside the 3×10^{-4} of the statement. At that point the margin is

$$\frac{\theta}{\sqrt{Q}} \le \frac{-0.399 + 10^{-3}}{\sqrt{\eta\,(1/(4\delta) + 1.1) + 3 \times 10^{-4}}} < -3 \qquad \text{for } \eta < \eta_1,$$

which is the last statement of the proposition; $\eta_{\mathrm{low}}(\delta)$ is the smallest of $\eta_1(\delta)$, the bound 0.04 of the contraction step, the bound $0.0324\,\delta^2$ of Lemma A.13, and the threshold $\delta/(1.3 \times 10^5 + 1.45 \times 10^4\,\delta)$ of the tube step. That threshold is at most 6.9×10^{-5} at every drift speed and so also delivers the $\epsilon \le 1/4$ and the $0.1\eta \le 1.34 \times 10^{-4}$ that the same step uses. □

At finite rates the full order-parameter equations behave as the reduced system does. At $(\delta, \eta) = (0.3, 0.1)$, $(0.3, 0.05)$, $(1.0, 0.1)$, and $(0.1, 0.1)$ the margin peaks at 0.24, 0.27, 0.30, and 0.12. The state settles at θ between -0.30 and -0.40 with an alignment of 0.2 to 0.6 (step 0.5, $300/\eta$ units).

A.7 Threshold rate, minibatches, and weight decay

Lemma A.15 (the threshold rate does not move the fixed points)**.** *If the threshold is stepped at $\kappa\eta/d$ for some $\kappa > 0$ while the weights keep the step η/d, the order-parameter equations become (2), (3), and $d\theta/d\alpha = -\kappa\eta\,\mathbb{E}[\chi]$. Their fixed points, the map Ψ, the fold locus, its anchors, and its end point are the same for every $\kappa > 0$.*

Proof. Per step the threshold increment is $-\kappa(\eta/d)\chi$. Its mean over d steps is $-\kappa\eta\,\mathbb{E}[\chi]$ per unit of α, and its fluctuation has variance $O(d^{-2})$ per step. The threshold therefore remains a deterministic order parameter with only its equation changed. At a fixed point,

$$0 = -\kappa\eta\,\mathbb{E}[\chi] \quad \Longleftrightarrow \quad \mathbb{E}[\chi] = 0 \qquad (\kappa\eta > 0),$$

which is the same threshold balance as at $\kappa = 1$, while the norm and alignment balances do not contain κ. Three balances determine Ψ in Proposition 3, and the fold locus, its anchors, and its end point are functions of Ψ alone. □

Lemma A.16 (minibatches rescale the diagram)**.** *Take b independent inputs per step, the mean gradient over the batch, and one teacher move per step. Then the order-parameter equations change*

in one place: the η^2 term of (3) becomes $\eta^2\mathbb{E}[\chi^2]/b$. Consequently the fixed-point map at batch size b is $\Psi_b = b\,\Psi$. The boundary and every landmark scale as $(\delta, \eta) \mapsto (b\delta, b\eta)$, that is $\eta_c^{(b)}(\delta) = b\,\eta_c(\delta/b)$, with anchors $b\,\eta_0$ and $2b/\pi$.

Proof. Let $(\xi_i, \chi_i)_{i=1}^b$ be the inputs and error signals of one step, independent across i given the state. Weight and threshold increments are

$$\Delta w = \frac{\eta}{bd}\sum_{i=1}^{b}\chi_i\,\xi_i, \qquad \Delta\theta = -\frac{\eta}{bd}\sum_{i=1}^{b}\chi_i.$$

Mean increments of R, θ, and the first-order part of Q are unchanged, because each is linear in the batch average:

$$\mathbb{E}[\Delta R] = \frac{\eta}{bd}\sum_{i=1}^{b}\mathbb{E}[\chi_i\,\xi_i\cdot B] = \frac{\eta}{d}\,\mathbb{E}[\chi y], \qquad \mathbb{E}[\Delta\theta] = -\frac{\eta}{d}\,\mathbb{E}[\chi], \qquad \mathbb{E}[2w\cdot\Delta w] = \frac{2\eta}{d}\,\mathbb{E}[\chi x].$$

Batch size b enters through the squared increment:

$$\begin{aligned}\mathbb{E}\big[\|\Delta w\|^2\big] &= \frac{\eta^2}{b^2d^2}\Big(\sum_{i=1}^{b}\mathbb{E}\big[\chi_i^2\|\xi_i\|^2\big] + \sum_{i\neq j}\mathbb{E}[\chi_i\xi_i]\cdot\mathbb{E}[\chi_j\xi_j]\Big)\\ &= \frac{\eta^2}{b^2d^2}\Big(b\,d\,\mathbb{E}[\chi^2]\,(1+O(d^{-1})) + b(b-1)\,\big\|\mathbb{E}[\chi\,\xi]\big\|^2\Big)\\ &= \frac{\eta^2\,\mathbb{E}[\chi^2]}{b\,d} + O(d^{-2}),\end{aligned}$$

where the second equality uses the decomposition behind (7). The error χ depends on ξ only through its two standard coordinates (u, z) in the student–teacher plane, and the remaining $d-2$ coordinates are independent standard Gaussians. Hence $\mathbb{E}[\chi^2\|\xi\|^2] = (d-2)\,\mathbb{E}[\chi^2] + \mathbb{E}[\chi^2(u^2+z^2)] = d\,\mathbb{E}[\chi^2]\,(1+O(d^{-1}))$. The cross terms use that $\mathbb{E}[\chi\,\xi]$ lies in the span of w and B with coefficients bounded at fixed (Q, R, θ). Over d steps this gives $dQ/d\alpha = 2\eta\,\mathbb{E}[\chi x] + \eta^2\mathbb{E}[\chi^2]/b$, and the teacher move once per step contributes $-\delta R$ as in Lemma A.3. At a fixed point the threshold balance and hence $Q^*(m,\rho)$ are unchanged. The norm balance gives

$$0 = 2\eta\,\mathbb{E}[\chi x] + \frac{\eta^2}{b}\,\mathbb{E}[\chi^2] \quad\Longleftrightarrow\quad \eta_b^* = -\frac{2b\,\mathbb{E}[\chi x]}{\mathbb{E}[\chi^2]} = b\,\eta^*,$$

and the alignment balance gives $\delta_b^* = \eta_b^*\,\mathbb{E}[\chi y]/(\rho\sqrt{Q^*}) = b\,\delta^*$. Hence $\Psi_b = b\,\Psi$ pointwise in (m,ρ). Its fold locus $\det D\Psi_b = b^2\det D\Psi = 0$ is the same set of pairs (m,ρ), and its image, the anchors, and the end point are multiplied by b. □

Lemma A.17 (weight decay inside the order-parameter equations)**.** *Add the decay $w \leftarrow w - (\eta/d)\,\lambda_{\mathrm{wd}}\,w$ to every step and leave the threshold undecayed. This adds $-\gamma_{\mathrm{wd}}R$ to (2) and $-2\gamma_{\mathrm{wd}}Q$ to (3) with $\gamma_{\mathrm{wd}} = \eta\lambda_{\mathrm{wd}}$, and leaves (4) unchanged.*

Proof. Weight decay changes w by $\Delta_{\mathrm{wd}}w = -(\eta\lambda_{\mathrm{wd}}/d)\,w$, so

$$\begin{aligned}\Delta_{\mathrm{wd}}R &= \Delta_{\mathrm{wd}}w\cdot B = -\frac{\eta\lambda_{\mathrm{wd}}}{d}\,R,\\ \Delta_{\mathrm{wd}}Q &= 2w\cdot\Delta_{\mathrm{wd}}w + \|\Delta_{\mathrm{wd}}w\|^2 = -\frac{2\eta\lambda_{\mathrm{wd}}}{d}\,Q + O(d^{-2}),\end{aligned}$$

which over d steps add $-\gamma_{\mathrm{wd}}R$ and $-2\gamma_{\mathrm{wd}}Q$ per unit of α. Cross terms with the gradient step are $O(d^{-2})$ per step and vanish in the limit, and the threshold is unchanged. □

Remark A.18 (what the three lemmas predict). By Lemma A.15 the dormancy boundary is a statement about a threshold that is slow on the time scale of the weights, while stability, basins, and entry times may depend on κ. Appendix D.4 measures that dependence.

Under Lemma A.16, averaging gradient noise moves the phase diagram rather than removing it: dormancy at batch b needs a learning rate b times larger at a drift b times faster, the finite-learning-rate mechanism with noise scale η^2/b. This rescaling includes a time transformation: the vector field

at $(b\delta, b\eta)$ with batch b is b times the field at (δ, η) with batch one. Trajectories coincide with time divided by b, $T_b(b\delta, b\eta) = T_1(\delta, \eta)/b$. Fixed-point landmarks scale as stated, while a first-passage contour at horizon H for batch b corresponds to horizon bH at batch one. Appendix D.4 tests this scaled fold.

The $-2\gamma_{\mathrm{wd}}Q$ term of Lemma A.17 acts directly against the η^2 term. Appendix D.4 integrates the decayed dynamics against the measured fates.

A.8 Two units sharing one output error

Appendix D.3 reports that below the single-unit fold exactly one of two units sharing an output error goes dormant. This subsection extends the same order-parameter treatment to that pair and locates the instability that selects the one.

Two student units w_1, w_2 with thresholds θ_1, θ_2 track two teacher units B_1, B_2, each teacher drifting by its own walk of Section 4.1 at the common speed δ. Write $x_k = w_k \cdot \xi$, $y_m = B_m \cdot \xi$, $\mathbb{1}_k = \mathbb{1}\{x_k > \theta_k\}$, and $m_k = \theta_k/\sqrt{Q_k}$ for the margin of unit k. The student output is $\sigma = \sum_k (x_k - \theta_k)\mathbb{1}_k$, the teacher output is $\tau = \sum_m y_m \mathbb{1}\{y_m > 0\}$, and both units are updated from the one residual:

$$\chi_k = (\tau - \sigma)\,\mathbb{1}_k, \qquad w_k \leftarrow w_k + \frac{\eta}{d}\chi_k\,\xi, \qquad \theta_k \leftarrow \theta_k - \frac{\eta}{d}\chi_k. \tag{39}$$

Nine order parameters close the description: the two norms $Q_k = \|w_k\|^2$, the student overlap $C = w_1 \cdot w_2$, the four alignments $R_{km} = w_k \cdot B_m$, and the two thresholds.

Lemma A.19 (order-parameter equations for the shared pair)**.** *In the limit $d \to \infty$ at fixed $\alpha = \mu/d$, with $B_1 \cdot B_2 = 0$ at $\alpha = 0$, the nine order parameters follow*

$$\frac{dQ_k}{d\alpha} = 2\eta\,\mathbb{E}[\chi_k x_k] + \eta^2\,\mathbb{E}[\chi_k^2], \tag{40}$$

$$\frac{dC}{d\alpha} = \eta\big(\mathbb{E}[\chi_1 x_2] + \mathbb{E}[\chi_2 x_1]\big) + \eta^2\,\mathbb{E}[\chi_1\chi_2], \tag{41}$$

$$\frac{dR_{km}}{d\alpha} = \eta\,\mathbb{E}[\chi_k y_m] - \delta R_{km}, \tag{42}$$

$$\frac{d\theta_k}{d\alpha} = -\eta\,\mathbb{E}[\chi_k], \tag{43}$$

and the teacher pair stays orthogonal, $d(B_1 \cdot B_2)/d\alpha = -2\delta\,(B_1 \cdot B_2)$. The expectations are over (x_1, x_2, y_1, y_2), jointly Gaussian with $\mathrm{Var}\,x_k = Q_k$, $\mathrm{Cov}(x_1, x_2) = C$, $\mathrm{Cov}(x_k, y_m) = R_{km}$, $\mathrm{Var}\,y_m = 1$, and $\mathrm{Cov}(y_1, y_2) = 0$.

Proof. The learning step of (39) gives, per step,

$$\Delta Q_k = 2w_k \cdot \Delta w_k + \|\Delta w_k\|^2 = \frac{2\eta}{d}\chi_k x_k + \frac{\eta^2}{d^2}\chi_k^2\|\xi\|^2,$$

$$\Delta C = w_1 \cdot \Delta w_2 + w_2 \cdot \Delta w_1 + \Delta w_1 \cdot \Delta w_2 = \frac{\eta}{d}\big(\chi_2 x_1 + \chi_1 x_2\big) + \frac{\eta^2}{d^2}\chi_1\chi_2\|\xi\|^2,$$

$$\Delta R_{km} = \Delta w_k \cdot B_m = \frac{\eta}{d}\chi_k y_m, \qquad \Delta\theta_k = -\frac{\eta}{d}\chi_k.$$

Because each χ_k depends on ξ only through the four coordinates (x_1, x_2, y_1, y_2), with the remaining $d - 4$ coordinates independent standard Gaussians, $\mathbb{E}[\chi_k\chi_l\|\xi\|^2] = d\,\mathbb{E}[\chi_k\chi_l]\,(1 + O(d^{-1}))$ by the count of Lemma A.16. Accumulating d steps per unit of α gives (40), (41) and (43), and the first term of (42). Increment fluctuations are $O(d^{-1})$ per unit of α by the same second-moment count. The teacher move contributes $-\delta R_{km}$ to (42) by Lemma A.3 applied to the pair (w_k, B_m), and nothing to Q_k, C or θ_k. For the teacher overlap, one step replaces B_m by $\cos(\omega)B_m + \sin(\omega)z_m$ with $z_m \perp B_m$ isotropic and z_1, z_2 independent, so

$$\mathbb{E}\big[B_1' \cdot B_2' - B_1 \cdot B_2\big] = \big(\cos^2\omega - 1\big)\,B_1 \cdot B_2 = -\frac{2\delta}{d}\,B_1 \cdot B_2 + O(d^{-2}),$$

the cross terms vanishing because each z_m has mean zero given the current state. Over d steps this is $d(B_1 \cdot B_2)/d\alpha = -2\delta\,(B_1 \cdot B_2)$, and an orthogonal pair therefore stays orthogonal. □

The moments. Expanding $\tau - \sigma$ makes every first-order expectation in (40) to (43) a sum of terms $\mathbb{E}[\, z\, \mathbb{1}\{p > a\}\, \mathbb{1}\{q > b\}\,]$ with z of degree at most two in the Gaussian pair (p, q). Conditioning removes any third variable, since $\mathbb{E}[z \mid p, q]$ is linear in (p, q). What remains are the six atoms of Lemma A.1, now with both thresholds general rather than the second held at zero, and the closed forms hold with the orthant probability written as two Owen terms rather than one. Three independent routes agree on them: Owen's T against an adaptive bivariate integrator at 300 random points, largest difference 2×10^{-16}; the six atoms against 30-digit quadrature at eight parameter sets including correlations $1 - 10^{-9}$ and $-(1 - 10^{-8})$, largest difference 10^{-16}; and the assembled first moments against Monte Carlo over 4×10^6 samples at six states, agreeing within the sampling noise of 2×10^{-4}.

Three and four indicator functions enter the second-order expectations $\mathbb{E}[\chi_k^2]$ and $\mathbb{E}[\chi_1 \chi_2]$, whose reduction passes through the trivariate normal distribution function. We evaluate them by two-dimensional Gauss–Legendre quadrature, 60×60 nodes per quadrant. Every evaluation of the right-hand side also recomputes the first moments by that quadrature and compares them with the closed forms. Over all integrations reported here the largest difference is 8×10^{-6}.

Symmetry. The equations are invariant under exchanging the two units together with the two teachers, and separately under exchanging the two teachers alone, since the teachers enter only through $\sum_m y_m \mathbb{1}\{y_m > 0\}$ and drift at one speed. That group is $\mathbb{Z}_2 \times \mathbb{Z}_2$, and the nine coordinates split into a four-dimensional fully symmetric block (Q, C, the mean alignment and θ), a one-dimensional teacher-antisymmetric block, a three-dimensional unit-antisymmetric block ($Q_1 - Q_2$, $\theta_1 - \theta_2$ and the alignment difference), and one doubly antisymmetric coordinate. Started from the trained state, in which each unit matches its own teacher, the symmetric flow converges to a fixed point with $R_{kk} = R_{km}$: the specialized memory is erased, because the teacher-antisymmetric coordinate has no source term and only the drift decay $-\delta$. Finite-dimensional runs do the same, their run-average of $R_{kk} - R_{km}$ falling from 0.07 to within ± 0.01 of zero over the first 500 units of learning time. Over the whole 3000-unit horizon at $\eta = 0.05$, and over the first 600 at $\eta = 0.1$, the symmetric flow matches the run-average of Q, R and θ to within one percent.

The symmetric state is transversely unstable. At $(\delta, \eta) = (0.3, 0.1)$ the symmetric fixed point is $Q = 0.161$, $C = 0.004$, $R_{kk} = R_{km} = 0.099$, $\theta = -0.346$ and $m = -0.86$ for each unit, both of them active. A central-difference Jacobian there, blocked by the symmetry, has vanishing cross blocks to 10^{-12}. Its symmetric block has only negative eigenvalues, and its antisymmetric block has a positive one. Across a grid of five drift speeds by five learning rates the antisymmetric eigenvalue is positive at all 25 cells, from 5×10^{-4} at $\eta = 0.05$ to 2.2×10^{-2} at $(0.15, 0.5)$, while the symmetric block stays stable throughout. The leading antisymmetric mode is mostly $\theta_1 - \theta_2$ with smaller components along $Q_1 - Q_2$ and the alignment difference, which is the picture of one unit raising its threshold while the other lowers it. Dropping the η^2 term leaves the eigenvalue positive but smaller by a factor 3.5 at $(0.3, 0.1)$, against the factor 3.1 by which the batch-mean gradient delays entry in Appendix D.3.

By itself the threshold direction supplies no instability. Differentiating $\mathbb{E}[\chi_1]$ along the antisymmetric perturbation $\theta_{1,2} = \theta \pm a$ splits its rate into two exact pieces:

$$\frac{d}{da}\frac{d\theta_1}{d\alpha} = -\eta\Big[\Pr(x_1 > \theta_1,\, x_2 \le \theta_2) - \frac{\varphi(m_1)}{\sqrt{Q_1}}\,\mathbb{E}[\tau - \sigma \mid x_1 = \theta_1]\Big]. \tag{44}$$

The first piece restores, and it counts only the region where unit 1 is active alone, because where both are active the partner's opposite move cancels the change in the residual. An isolated unit has the whole active set here instead, which is why it is stable in this direction. Feedback comes from the second piece, through samples leaving the active set at a threshold where the residual is positive. Under a batch-mean gradient the two are equal to ten digits at the fixed point, at 0.0351, 0.0616, 0.0826 and 0.0959 for $\eta = 0.05$, 0.1, 0.2 and 0.5, which makes the entry vanish. At states off the fixed point they differ by factors of several. The instability therefore comes through the coupling: a unit with the higher threshold grows in norm, a unit with the larger norm contributes more output and meets a smaller residual, and its threshold rises further. With the η^2 term the entry (44) is slightly negative and the positive eigenvalue is unchanged: the same loop produces the instability in both cases.

What the instability produces. Integrating the full nine equations from the trained state with a small antisymmetric threshold bias $\theta_{1,2} = \pm a_0$, $a_0 = 0.01$, every cell below the single-unit fold

ends with exactly one margin past $m_\star = 2$ and the other unit active and unchanging thereafter, and the two cells above it end with both units past the cutoff. This reproduces the pattern of Table 6 cell by cell. Entry times are 5460, 1773, 745, 350 and 1220 at $\eta = 0.05$, 0.1, 0.2, 0.5 and 0.8 with $\delta = 0.3$, against simulated means of 6927, 2079, 739, 309 and 841. Removing the η^2 term delays entry at $\eta = 0.1$ from 1773 to 4000, against 2079 to 6428 in the batch-mean runs, and removes entry at 0.05 within 12000 units, as those runs also do. Dependence on the bias is the linear one: $a_0 = 0.001$, 0.01 and 0.045 give 2858, 1773 and 1087, and the first gap of 1085 is $\ln(10)/\lambda = 1086$ at the eigenvalue computed above. Entry time is thus the relaxation of the symmetric state plus the time $\ln(1/a_0)/\lambda$ for the antisymmetric mode to reach order one. In a finite-dimensional run the bias comes from fluctuations of order $d^{-1/2}$, which is the size used here.

The growth rate is measurable. Fitting $\ln|\theta_1 - \theta_2|$ over the window where the difference lies between 0.02 and 0.3 in $d = 500$ runs gives 0.00056 ± 0.00009 against the eigenvalue 0.00052 at $\eta = 0.05$ (eight runs), 0.00236 ± 0.00051 against 0.00212 at 0.1 (sixteen runs), and 0.0150 ± 0.0058 against 0.0174 at 0.5 (eight runs), the spreads being standard deviations across runs. Which unit wins is set by the noise: unit 1 in ten of the sixteen runs at $\eta = 0.1$. Aligning the six recorded trajectories of Figure 8 and the integrated one on their own entry times, the norm, alignment and threshold of both the entering and the surviving unit agree within one standard deviation of the runs at all eight sample times from 600 before entry to 1000 after.

The loser reaches the cutoff by the opposite movement of the norm. The isolated unit reaches it with a norm that has grown: integrated from the trained state at $\delta = 0.3$, Q rises from 1 to 1.44 and 1.72 by first passage at $\eta = 0.8$ and 1.2, and the threshold outruns it (Figure 6). The loser reaches the cutoff with a norm that has fallen. At $\delta = 0.3$ its Q is 0.018, 0.038, 0.082 and 0.297 at its own first passage for $\eta = 0.05$, 0.1, 0.2 and 0.5, its alignment with either teacher is 0.01 or below, and its threshold has turned positive, from 0.01 at the start to 0.30, 0.43, 0.66 and 1.35. Its margin crosses the cutoff because the threshold turns positive while the weight length shrinks, and not because the weight length grows and the threshold outruns it.

What separates the two routes is the shared error. Until the symmetry breaks the two units follow the symmetric flow together: at $\eta = 0.1$ both norms fall from $Q = 1$ to $Q = 0.16$ and their alignments stay equal. After it the winner takes the residual, its alignment with both teachers rising to 0.126 and its norm settling at $Q = 0.166$, while the loser is left with an error signal that collapses. Because (40) has no drift term, a norm moves only through the error, and for the loser that motion is downward at every cell integrated. At the moment the loser enters, the winner still holds an alignment of 0.24 to 0.42 with its own teacher, and the pair ends with one unit tracking both teachers and one tracking neither.

Section 5.4 reports that the dormant units of the layer keep chance alignment and shrink their weight vectors, from a median of 0.66 to 0.38. That pairs them with this route rather than with the isolated unit, whose norm grows into the cutoff. A layer of 128 units that trains its output weights is not the pair of units at a fixed output weight that the equations above describe, and the match is one of signature rather than of setting.

B Experimental Setup

Every simulation number in Section 5 and Appendix D rests on the grids, run counts, horizons, initial states, and controls below.

Throughout the paper a criterion is "stated in advance", or "on file", when we entered its quantity and pass threshold in our dated criteria file before its data were collected. Criteria added after the original set are marked "added later"; they were likewise entered before their own data existed, except one that Table 7 marks as run before its reading rule was written.

Each experiment family draws its runs by splitting one base seed, and the seed is written into every output file (main grid 1, dimension scans 61, two units 21, end-point and plateau scans 11, network grid 1, magnitude reruns 17).

The order-parameter predictions of the main grid integrate (2)–(4) by fourth-order Runge–Kutta at step 0.01 from the trained state. Entry is the first crossing of $m_\star$, linearly interpolated within the

step. The scans of Appendix B.7 use step 0.5 unless stated. Appendix D.5 lists the criteria that fail or hold with a qualification. Our full register is in the supplement.

The leading-order scheme integrates the order parameters $X = (Q, R, \theta)$ as $dX = F(X)\, d\alpha + \sqrt{\Sigma(X)/d}\, dW$ with F the right-hand side of the order-parameter equations and Σ the fluctuation covariance of Appendix D.2 (drift channel included). Both terms are evaluated by Gauss–Legendre quadrature with 96 nodes per panel over the active side of the input, with the teacher side split at its kink. The diffusion approximation adds the $O(1/d)$ norm-drift term at every step by 48×48 Gauss–Hermite quadrature in the student–teacher plane. Both schemes are integrated by Euler–Maruyama from the trained state, and each path is frozen at first passage unless stated. Entry is the margin reaching $m_\star$ within the horizon, and the 50% crossing is obtained by linear interpolation.

B.1 Main grid and boundary scan

Main grid: $\delta \in \{0.1, 0.15, 0.2, 0.3, 0.45, 0.65\} \times \eta \in \{0.7, 0.8, 0.9, 1.0, 1.1, 1.2\}$ at $d = 500$ with 100 runs per cell. The horizon of each cell is $\min(3000,\ 4\, T_{\mathrm{pred}})$ units of learning time, where T_{pred} is its predicted entry time. Entry is when the empirical margin $\theta/\|w\|$ first reaches 2. Every entering cell enters in all 100 runs before its horizon, so the horizon censors no entering run (a non-entering run is right-censored at 3000). Non-entering cells have an infinite predicted time and run to 3000. Each run starts from the trained state, $(Q, R, \theta) = (1,\ 1 - 10^{-6},\ 0)$. That alignment offset keeps the initial state off the measure-zero edge $s = 0$. Order within a step: the sample is labeled by the current (pre-move) teacher, the student updates, and the teacher then moves.

Boundary scan: $\delta = 0.3$ fixed and η from 0.50 to 0.68 in steps of 0.02, with 100 runs each at the same horizon. A refinement on file before it ran adds η from 0.665 to 0.690 in steps of 0.005 with 100 runs each. A scan at $d = 1000$ over the same horizon covers η from 0.680 to 0.710 in the same steps (100 runs each, seed 131). Its entry counts are 7, 13, 30, 57, 81, 87 per 100: a monotone rise across the transition.

An either-or criterion on file (called the signature disjunction in Table 7) pairs this boundary-band test with a finite-size decomposition; either alternative is enough. Its premise above the boundary is a mean entry time varying by less than 25% between $d = 60$ and 480. In the dimension scan the mean entry time varies strongly with d. This contradicts that premise: the finite-size alternative fails, and the signature outcome rests on the boundary alternative alone.

B.2 Dimension scans

Dimension scans: the cell $(\delta, \eta) = (0.3, 0.8)$ at $d \in \{30, 60, 120, 240, 480, 960, 1920\}$, and a second cell $(0.15, 0.9)$ at $d \in \{60, \ldots, 960\}$, 100 runs each, horizon 3000. At the second cell, as at the first, the mean entry time rises monotonically toward its predicted limit (524.5 at $d = 60$ toward 1543.7). We fit the amplitude of the $d^{-1/2}$ law on the largest dimensions below 1920, where the local slope approaches $-1/2$. We ran the $d = 1920$ cell only after fixing its predicted band. Over the last measured octave the deficit of the mean entry time from its deterministic limit has local log-log slope -0.545. A free-exponent fit over $d \geq 240$ (run-level bootstrap, 1000 resamples) gives -0.45 $[-0.49, -0.42]$ at $(0.3, 0.8)$ and -0.50 $[-0.56, -0.44]$ at $(0.15, 0.9)$. The first interval excludes $-1/2$ by 0.01 on the shallow side, consistent with the local slope still moving toward $-1/2$ within the fitted range.

Dimension scans at the slowest and fastest grid speeds use the cells $(0.1, 1.2)$ and $(0.65, 0.8)$ at $d \in \{500, 1000, 2000\}$ with 100 runs per cell and horizon 3000 for the new dimensions.

Diffusion across the scan: the leading-order scheme and the diffusion approximation at $(0.3, 0.8)$ for $d \in \{30, 60, 120, 240, 480, 960, 1920\}$, 1000 paths, horizon 3000, step 0.5 (0.25 at $d \leq 60$), seed 3. They are reported as the mean entry time against the measured mean at each dimension and the log-log slope of the deficit from the deterministic limit against d for both. Their outcome is in Appendix D.2.

B.3 Negative controls

Two negative controls accompany every scan of the solvable model and of the network (the batch-32 test carries a drift-free row throughout and the jitter control at its one entering cell). Drift off: the

teacher is held fixed. Jitter: at each step the teacher is reset to $B = \cos(\omega)B_0 + \sin(\omega)z$ with a fresh unit vector $z \perp B_0$ and $\cos(\omega) = 1 - \delta/d$. The displacement of each teacher from the fixed reference direction B_0 then equals the per-step displacement of the drifting condition at the same δ. The squared distance between two consecutive jitter teachers is about $2\omega^2$, twice that between consecutive drifting teachers. The control therefore has at least the per-step noise of drift while nothing accumulates. In the solvable model the jitter control is run at both $\delta = 0.3$ and the strongest drift, $\delta = 0.65$, with six learning rates per control and 100 runs per cell. A third control, which deletes the η^2 term of the norm equation, is defined with its outcome in Appendix D.3. Matched-residual control: the teacher is held fixed and Gaussian noise of standard deviation $\sigma_{\rm n}$ is added to its output. Per cell, $\sigma_{\rm n}$ is set from $\mathbb{E}[\chi^2]$ of the drifting condition in the order-parameter equations (at its tracking state where one exists, otherwise the time average along the deterministic trajectory from the trained state to first passage). The matched-residual control runs at the twelve cells of the controls of the solvable model, with $d = 500$, 100 runs, horizon 3000, and one seed set. Two noise levels are reported in Appendix D.3: $\sigma_{\rm n}\sqrt{2}$ matching $\mathbb{E}[\chi^2]$ at the activity 0.5 of the trained state, and $2\sigma_{\rm n}$ at the six $\delta = 0.3$ cells. Independent targets: the target is $\max(0, z)$ with z a standard normal drawn afresh at every step independently of the input, the teacher unused. Protocol: $d = 500$, 100 runs, horizon 3000, $\eta \in \{0.5, 0.6, 0.7, 0.8, 1.2\}$, reported as the entry fraction, the mean entry time, and the 50% crossing by interpolation against the $\rho = 0$ envelope $2/\pi$. Noisy fixed teacher at the independent residual: the teacher held fixed with Gaussian label noise of standard deviation 0.584 ($\sigma_{\rm n}^2 = 1/2 - 1/(2\pi)$, the variance of $\max(0, z)$), same cells, runs, and horizon as the independent targets.

B.4 Basin sweep and random initializations

Basin sweep, on file before it ran: every cell of the main grid and the boundary scan (46 cells) is integrated in the order-parameter equations from a 108-point initial-condition lattice with horizon 1.2×10^4. Its lattice is $\rho \in \{0.1, \ldots, 0.9\} \times m \in \{0, 0.5, 1, 1.5\} \times Q \in \{0.25, 1, 4\}$. A near-trained subset, also on file, is $\rho \geq 0.7$ and $m \leq 0.5$.

Random initializations: $w \sim \mathcal{N}(0, \mathbf{I}/d)$ and $\theta = 0$ at $d = 500$ with 100 runs per cell at three cells chosen in advance, $(0.3, 0.5)$, $(0.3, 0.8)$, and $(0.3, 1.1)$. These runs are compared with the order-parameter equations started from the matched state $(Q, R, \theta) = (1, 1/\sqrt{d}, 0)$. A random isotropic initialization has a signed overlap centred at zero with typical magnitude $d^{-1/2}$. A matched state takes that magnitude with positive sign; a later check averages the order-parameter prediction over the signed initial overlap. Outcomes, including this check, are in Appendix D.1.

B.5 The network experiment

Student: one hidden layer of 128 ReLU units with biases. A unit computes $g(w \cdot x - \vartheta)$. Its bias ϑ corresponds to the threshold θ with the same sign ($\vartheta = -b$ for a bias b added to the pre-activation). Input weights, biases, and the output layer are all trained at raw learning rate $\eta_{\rm raw}$ by plain online stochastic gradient descent without momentum, with batch size one on the squared loss. The learning rate of the model is $\eta = \eta_{\rm raw} d$ (Appendix D.4). Initialization: student weights $\mathcal{N}(0, 1/d)$, biases zero, and output weights $\mathcal{N}(0, 1/128)$. Teacher: a random network with 32 hidden units at input dimension 32, whose input-weight rows drift and whose other parameters are frozen. The teacher has rows $\mathcal{N}(0, 1/d)$, biases $\mathcal{N}(0, 0.1^2)$, and output weights $\mathcal{N}(0, 1/32)$. Inputs are standard Gaussian in $\mathbb{R}^d$. The loss is $\frac{1}{2}(\text{output} - \text{target})^2$ (its batch mean at batch 32), and neither network has an output bias.

Drift moves each teacher input-weight row on its sphere with per-step overlap $1 - \delta/d$, the convention of Section 4.1. Under it the overlap of a teacher with its earlier self decays as $e^{-\delta\alpha}$ in learning time. Grid: $\eta_{\rm raw} \in \{0.00375, 0.0075, 0.015, 0.03\} \times \delta \in \{0.075, 0.15, 0.3, 0.6\}$ plus a drift-free row and a jitter row with no accumulation at each learning rate, with 10 runs per cell and 3×10^5 steps. Every 10^4 steps, per-unit activity is measured on 4096 fresh inputs. A unit is dormant when its activity is at most 0.023, matching Definition 1.

A run whose final training loss exceeds ten times its initial loss or is not finite is flagged divergent and excluded from dormancy statistics. In the grid of Figure 3b divergence occurs only above the largest learning rate retained, so that grid contains no divergent runs. But the per-unit diagnostic and the magnitude rerun at $\eta_{\rm raw} = 0.03$ below each drop one or two runs per cell under this rule.

Every dropped run has non-finite weights from its first log point, so its zero activity is numerical divergence rather than dormancy. Appendix D.3 gives the kept counts. The learning-rate scan below reports its own divergence ceiling.

Unlike the fixed unit output of the solvable model, the output layer of the student is trained. Drift can therefore be partly absorbed by output weights, and hidden units compete for it.

Per-unit diagnostic: at $(\delta, \eta_{\text{raw}}) = (0.6, 0.03)$ over ten runs, the weight norm, bias, activity, and best cosine to a teacher row of every hidden unit are logged every 10^4 steps. Input dimension 512: the same network with 512 inputs (teacher and student) at the same model learning rates $\eta = \eta_{\text{raw}} d \in \{0.12, 0.24, 0.48, 0.96\}$ and drift speeds as the $d = 32$ grid, with ten runs per cell and 3×10^5 steps. A rerun at the learning time of the $d = 32$ grid (9,375 units, 4.8×10^6 steps) covers six cells: $(\delta, \eta) \in \{0.3, 0.6\} \times \{0.48, 0.96\}$ and the two drift-off controls at $\eta = 0.48$ and 0.96. A width-matched rerun (added later) repeats these six cells with 2048 student units in place of 128, with the same seeds, steps, and learning rates. The student then keeps four units per input, as at $d = 32$. Batch 32: the same network with the batch-mean gradient and one teacher move per step, on the grid $\delta \in \{0.15, 0.6, 2.4\}$ plus a drift-free row by $\eta_{\text{raw}} \in \{0.03, 0.06, 0.12, 0.25, 0.5\}$, with ten runs per cell and 3×10^5 steps. A jitter control runs at $(\delta, \eta_{\text{raw}}) = (2.4, 0.12)$ (ten runs, seed 233). It runs twice independently on different hardware (GPU and CPU): one copy completes all twenty cells and the other the fourteen drift cells other than $(2.4, 0.5)$. Four times the budget: the two smallest raw rates $\{0.00375, 0.0075\}$ at $\delta \in \{0.15, 0.6\}$ rerun for 1.2×10^6 steps with the same seed and ten runs. Intermediate models: the two largest raw rates of the grid $\{0.015, 0.03\}$ at $\delta \in \{0.15, 0.6\}$ with ten runs per cell and 3×10^5 steps. One change is made at a time: a trained start (32 student units initialized at the weights, biases, and output weights of the teacher), output weights frozen at their initial values, or independent units, meaning 32 student units each trained by the squared loss on the ReLU output of its own teacher unit with no shared readout.

Per-unit entry is the first log point (every 10^4 steps) at which the activity of a unit is at most 0.023. The norm ratio of a unit is its norm at the log point before entry over its initial norm. At each cell the single-unit prediction integrates the order-parameter equations at $\eta = \eta_{\text{raw}} d$ from the trained state and from $(Q, R, \theta) = (1, 1/\sqrt{32}, 0)$ over 9375 units. Magnitude-based score: the largest-rate row ($\eta_{\text{raw}} = 0.03$ at the four drift speeds, the drift-free and the jitter control), ten runs, seed 17, with the mean post-activation magnitude of each unit over the 4096 evaluation inputs logged at every log point. The score of Sokar et al. (2023) divides this magnitude by the layer mean at the final log point. Magnitude score across the grid: every cell of the $d = 32$ grid and the drift-off and jitter controls at each rate rerun with per-unit magnitudes logged, ten runs, seed 17, 3×10^5 steps, scored as above at the final log point. Recovery on a fresh teacher: the same cells and the drift-free network at both rates run the 3×10^5 drift steps and then 10^5 further steps on a fresh fixed random teacher at the same raw rate, with loss and dormant fraction logged every 500 steps in the second phase.

B.6 Network learning-rate scan, two units, and a trained output weight

Learning-rate scan across drift speeds: the drifting-teacher network experiment over $\delta \in \{0.075, 0.15, 0.3, 0.6, 1.2, 2.4\}$ and eight raw learning rates η_{raw} from 0.005 to 0.085 (logarithmic grid, factor ≈ 1.5 between neighbors) with eight runs per cell. A criterion stated in advance defines the network onset $\eta_c^{\text{net}}(\delta)$ as the smallest raw learning rate at which the seed-mean final dormant fraction reaches 5%, linearly interpolated within each drift column.

The three largest learning rates diverge in every cell (144 of 384 runs), which caps the usable grid at $\eta_{\text{raw}} \leq 0.025$. The measured onset declines monotonically and saturates: $\eta_c^{\text{net}} = 0.0188$, 0.0159, 0.0132, 0.0128, 0.0124, 0.0122 across the six drift speeds. Converted with $\eta = \eta_{\text{raw}} d$ at $d = 32$ these are 0.60 to 0.39. Two units: a two-unit student $\sigma = \sum_{k=1,2} g(w_k \cdot \xi - \theta_k)$ tracks a two-unit teacher $\tau = \sum_{m=1,2} g(B_m \cdot \xi)$ whose orthonormal vectors drift independently at the common speed δ, each by the random walk of Section 4.1 with its own displacement. All student parameters share the rate η/d, and the runs start from the specialized trained state: unit k at overlap $1 - 10^{-6}$ with B_k, unit norm, threshold zero. Per-unit entry uses the margin criterion of Definition 1 ($d = 500$, 50 runs per cell, seed 21, horizon 3000). Cells are $\delta = 0.3$ with $\eta \in \{0.5, 0.8, 1.0, 1.2\}$, the cell $(\delta, \eta) = (0.15, 0.7)$, and a drift-free control at $\eta = 1.2$. Three further cells chosen in advance, $\eta \in \{0.05, 0.1, 0.2\}$ at $\delta = 0.3$, probe the onset of the competition channel. A repeat on file before it ran scales student and teacher outputs by $1/K$, with $K = 2$ the number of units (same protocol,

50 runs per cell). A four-fold horizon: the cells $(0.3, 0.05)$ and $(0.3, 0.1)$ rerun over 12000 units with the same seed and 50 runs; the isolated unit is run at $(0.3, 0.05)$ and $(0.3, 0.2)$ over 12000 units ($d = 500$, 100 runs, seed 11). Batch 100 and 500: the same two cells with the batch-mean gradient over 100 inputs per step (both rates) and 500 ($\eta = 0.1$), one teacher move per step, 32 runs, horizon 12000.

Matched bridge: two units as described above at $\delta = 0.3$ and $\eta \in \{0.1, 0.5, 0.8\}$ ($d = 500$, 32 runs, seed 41, horizon 12000) in three variants. Variants: independent residuals, with each unit matching its own teacher unit; the shared residual; and the shared residual with the two output weights trained from 1 at the same raw step. Reported: units and runs entered, the entry time of the first entering unit, and the final mean thresholds and output weights. Trainable output weight, single unit: the model of Section 4.1 with output $v\,\sigma$ and error $e = \tau - v\,\sigma$, updates $w \leftarrow w + (\eta/d)\,v\,e\,\mathbb{1}\,\xi$, $\theta \leftarrow \theta - (\eta/d)\,v\,e\,\mathbb{1}$, and $v \leftarrow v + (\eta/d)\,e\,\sigma$ with v initialized at 1 (so $\chi = v\,e\,\mathbb{1}$ replaces the error signal of (1)). Runs: $(\delta, \eta) \in \{(0.3, 0.8), (0.3, 1.2), (0.3, 0.5), (0.1, 1.0), (0.65, 0.8)\}$, $d = 500$, 100 runs, horizon 3000, seed 11, reported as entries and the final mean output weight and threshold. Four-variable closure: with $e = \tau - v\,\sigma$ the averaged dynamics are $\dot{R} = \eta v\,\mathbb{E}[e\,y\,\mathbb{1}] - \delta R$, $\dot{Q} = 2\eta v\,\mathbb{E}[e\,x\,\mathbb{1}] + \eta^2 v^2\,\mathbb{E}[e^2\mathbb{1}]$, $\dot{\theta} = -\eta v\,\mathbb{E}[e\,\mathbb{1}]$, and $\dot{v} = \eta\,\mathbb{E}[e\,\sigma]$, the expectations by the quadrature of Appendix C.3. They are integrated by fourth-order Runge–Kutta at step 0.5 over 3000 units from $(1, 1, 0, 1)$ at the four entering cells, reported as the final state and the largest margin along the way, against the final means of the trained-weight runs. Output weight held fixed: the same cells with $v = 0.15$ held constant (no update of v), 100 runs, horizon 3000, seed 11, reported as entries and the final mean threshold. Reduction check: the original unit at rate $v^2\eta$ from the initial state $(Q, R, \theta) = (v^2, v, 0)$ and from the function-matched trained state $(1, 1, 0)$, at the same cells, 100 runs, horizon 3000, seed 11.

B.7 End-point scans, horizon scans, and the action minimization

Interpolation sensitivity: at every comparison of Table 2 the 50% crossing by linear interpolation of the per-rate entry fractions is compared with the centre of a probit fit (maximum likelihood over the scanned rates, centre and width on a grid of 10^{-4}). A locally refined microscopic scan at $\delta = 1.2$, horizon 12000, rates 0.562, 0.564, 0.566, 0.568 (100 runs each, same protocol) resolves the sharpest transition.

Post-end-point onset in the order-parameter equations: for $\delta > \delta_{\mathrm{cp}}$ the onset is computed by bisection (tolerance 10^{-3}) on first passage of the margin to $m_\star$ within $\alpha \le 1.2 \times 10^4$, integrating from the trained state at step 0.5 up to $\delta = 1.2$. At $\delta \in \{2, 3, 5\}$, and for the conventions $m_\star \in \{1.5, 2.5\}$ at $\delta \in \{0.7, 1, 1.5, 2, 3\}$, the bisection is run at steps 0.5 and 0.25 with three rounds of sixteen rates and resolution 4×10^{-5}. Longer horizons (4.8×10^4, 1.92×10^5, 7.68×10^5) use a learning-rate grid step of 0.003 or bisection resolution 2.4×10^{-4} with integration step 1.0. At $\delta = 3.0$ the integration step is 0.5, because step 1.0 leaves the stability region of the integrator there ($\delta h > 2.78$).

Direct test of the non-monotonicity of the onset at finite d, on file before it ran: $\delta \in \{0.45, 0.59, 1.2\}$ with eleven learning rates each in steps of 0.01, at $d = 500$ with 100 runs per cell and the 3000-unit horizon. A follow-up likewise on file uses a four-fold horizon ($\alpha \le 12000$) and 100 runs per cell with lower learning-rate windows: 0.555 to 0.635 at $\delta = 0.45$ and 0.535 to 0.615 at 0.59 in steps of 0.01, and 0.585 to 0.605 in steps of 0.005 at $d = 1000$. Further scans use $d = 2000$ at the same horizon, and $d \in \{1000, 4000\}$ for the dimension extrapolation added later.

Action minimization: minimizing the path action under the fluctuation covariance of Appendix D.2 fails its check, which repeats it at doubled path resolution and from a perturbed initialization. Discretization, initialization, and optimizer settings are with the code.

Direct scaling tests of the fluctuation-driven escape, on file before they ran: $(\delta, \eta) = (0.3, 0.60)$ with 500 runs at each of $d \in \{120, 240, 480\}$. A nearer cell $(0.3, 0.62)$ is run at $d \in \{60, 90, 105, 120, 135, 150, 240, 480\}$ with 500 runs at $d \in \{60, 120, 240, 480\}$ and 1000 at $\{90, 105, 135, 150\}$. All are within the 3000-unit horizon. The diffusion approximation is run at the same cell and dimensions (1000 paths, 4000 at $d \ge 150$, step 0.5, the norm-drift term included). It is reported as entry fractions against the measured ones and a weighted least-squares slope of the log fraction over the nonzero dimensions (weights $np/(1-p)$) with a binomial-likelihood slope over all dimensions.

Crossing scans under the leading-order scheme: step 0.5 with 1000 paths per learning rate. The crossing-scan cells are $\delta = 0.3$ at $d \in \{500, 1000\}$ (horizon 3000, η from 0.60 to 0.72), $\delta = 0.45$ at $d = 500$, and $\delta = 0.59$ at $d \in \{500, 1000\}$ (horizon 1.2×10^4).

The leading-order scheme over the main grid: the scheme at every cell ($d = 500$, horizon 3000, 1000 paths per cell, seed 3), reported as the entry fraction and the mean first-passage time of the entered paths. Two additions: the step-0.25 refinement at $(0.1, 1.0)$, $(0.3, 0.8)$, and $(0.65, 1.2)$; and the persistent-entry run, integrated without freezing after first passage and measured with the 50-unit persistence window of Appendix B.8 at all thirty-six cells. The diffusion approximation over the main grid: the same runs at the thirty-six cells and at the five boundary-crossing cells of the leading-order scheme. Convergence of the diffusion approximation: at $(0.1, 1.0)$ ($d = 500$, 1000 paths, horizon 3000, seed 3) the scheme is rerun at step 0.25, with 96×96 quadrature, and with the $O(1/d)$ ordering term $-(\delta\eta/d)\,\mathbb{E}[\chi y]$ of the overlap drift added. The crossing scan at $\delta = 0.3$, $d = 500$ is rerun with that term. Both are reported as entry fraction, restricted mean, persistent-entry fraction, and 50% crossing against the 48×48, step-0.5 runs. Beyond the end point: crossing scans under the diffusion approximation at $d = 500$, 1000 paths, seed 3, within horizon 3000 at $\delta \in \{0.45, 0.59, 1.2\}$ and within 12000 at $\delta = 1.2$ (rates 0.50 to 0.58); the same scheme at $\delta = 0.59$ and $d \in \{2000, 4000\}$ within 12000 over the rate grid of the $d = 1000$ scan. They are compared with the measured short-horizon crossings and with a microscopic scan at $\delta = 1.2$, horizon 12000, $d = 500$, 100 runs per rate at $\eta \in \{0.50, 0.52, 0.53, 0.54, 0.55, 0.56, 0.562, 0.564, 0.566, 0.568\}$ (seed 11) joined to the 0.57 and 0.58 cells of the earlier window scan. Intervals come from binomial resampling of the per-rate fractions of the diffusion (2000 resamples, monotone interpolation) and from the posterior of the measured fractions under the same interpolation. Distribution-level comparison: the same thirty-six cells rerun with the first-passage time and persistent-entry start of every path kept, against the per-run first-passage times of the main grid and the per-run persistent-entry starts of the persistence rerun. Per cell the comparison reports the 25, 50, and 75% quantiles of both (paths and runs that do not enter placed at the horizon) and the Kolmogorov distance between the two first-passage distributions. It also reports the entry and persistent fractions against the measured 95% Wilson interval. Our integrator also counts the steps at which Q is floored at 10^{-6} and at which R^2 exceeds Q.

Slow-drift cells across dimensions: the leading-order scheme at $\delta = 0.1$, $\eta \in \{1.0, 1.2\}$ for $d \in \{250, 1000, 2000\}$ (500 from the main-grid run), 1000 paths, horizon 3000, step 0.5, seed 3. Comparison: microscopic runs at $(0.1, 1.0)$ for $d \in \{250, 1000\}$ with 100 runs and seed 201, and the dimension-scan runs at $(0.1, 1.2)$, reported as entry fractions and median first passage on both sides. Slow-drift plateau: microscopic runs at $\delta \in \{0.003, 0.01\}$ and $\eta \in \{1.8, 1.9, 2.0, 2.1, 2.2\}$ ($d = 500$, 50 runs, horizon 12000, seed 11) with two stationary controls at the same five rates with the same runs, horizon, and seed: a fixed teacher with label noise of standard deviation 0.2 from the trained state, and a noiseless fixed teacher from a random initialization ($w \sim \mathcal{N}(0, \mathbf{I}/d)$, $\theta = 0$). Reported: entry fractions, mean entry times, and the 50% crossing in η against the fold by continuation. Onset neighbourhood: microscopic runs at $\delta = 0.1$, $\eta \in \{0.925, 0.95, 0.975, 1.025, 1.05\}$ ($d = 500$, 100 runs, horizon 3000, seed 11; 1.0 from the main grid) and the leading-order scheme at the same six rates (1000 paths), reported as entry fractions, median first passage, and the 50% crossing in η on both sides.

Moment check: at the states of the deterministic $(0.1, 1.0)$ trajectory with margins 0.5, 1, 1.5, and 2, a weight vector and threshold realizing (Q, R, θ) at $d = 500$ and five replicates of 2×10^5 fresh inputs give the mean one-step increment of (Q, R, θ) under the gradient step, teacher fixed for the step. It is scaled to learning time with its standard error, against the closed-form gradient-channel drift and the quadrature the diffusion integrates. Overlap-drift perturbation: the leading-order scheme at $(0.1, 1.0)$ with a constant 0.003 or 0.006 added to the overlap drift, 1000 paths, horizon 3000, step 0.5.

Small rates over a long horizon: the order-parameter equations from the trained state at $\eta \in \{0.1, 0.2, 0.3, 0.4\}$ and $\delta \in \{0.1, 0.3, 0.65\}$ integrated to 10^5 units, reported as the final margin, the late slope of the margin, and the Jacobian eigenvalues at the final state. Same-horizon contour: the deterministic onset of the order-parameter equations from the trained state within 3000 units at seventeen speeds from 0.05 to 3.0, and within 1.2×10^4 at eight speeds from 0.45 to 1.2, by bisection to 3×10^{-5} with step 0.5.

Fold versus dormancy boundary: the cutoff curve $\Psi(m_\star, \rho)$ for ρ from 0.001 to 0.999 against the fold interpolated at the same δ; fixed points, stability, and margins on η from 0.005 below to 0.06 above the fold in steps of 0.0025 at $\delta \in \{0.1, 0.3, 0.45, 0.55, 0.59\}$; trajectories from the trained state at 0.001, 0.003, 0.01, and 0.03 above the fold at the same speeds, and at $(0.59, 0.564)$, over 4×10^5 units at step 0.5. Refinement where the boundary leaves the fold: at $\delta \in \{0.548, 0.55, 0.553\}$ the fold, the maximum of η^* on the δ level set of Ψ along the tracking branch, is located from the closed forms with the cutoff curve $\Psi(m_\star, \cdot)$. The fixed points of the order-parameter equations are sought on a learning-rate grid of step 10^{-4} from 0.0002 below to 0.0012 above the fold. Trajectories from the trained state are integrated (fourth-order Runge–Kutta, step 0.5) over 4×10^5 units at 0.0001, 0.0002, 0.0003, 0.0005, 0.0008, and 0.0012 above the fold. End-point fixed points: the fixed points of the order-parameter equations on $\eta \in [0.550, 0.600]$ in steps of 0.0025 at $\delta \in \{0.59, 0.5939, 0.595, 0.60, 0.62\}$ with their Jacobian eigenvalues, and trajectories from the trained state at $\delta \in \{0.5939, 0.595\}$ and $\eta \in \{0.5623, 0.5635, 0.5645, 0.5656, 0.568, 0.575\}$ over 2×10^5 units at step 0.5.

B.8 Conventions, robustness variants, and the threshold rate

Persistence-based entry: the main grid rerun under the original protocol and seeds with two additional quantities recorded per run. The two quantities are the start of the first interval of at least 50 units of learning time spent continuously at or above $m_\star$ (persistent entry) and the minimum margin after first passage. The timing comparison also uses windows of 25 and 100 units beside the 50-unit one.

Label noise: Gaussian noise of standard deviation $\sigma_y \in \{0.1, 0.3\}$ added to the teacher output at the cells $(0.3, 0.8)$ and $(0.3, 0.5)$ with $d = 500$, 100 runs, and horizon 3000.

Weight decay: the decay $w \leftarrow w - (\eta/d)\,\lambda_{\mathrm{wd}}\,w$ at every step with $\lambda_{\mathrm{wd}} \in \{0.01, 0.1\}$ at the four cells $(0.3, \eta)$, $\eta \in \{0.5, 0.6, 0.8, 1.2\}$ ($d = 500$, 100 runs, horizon 3000). We integrate the decayed order-parameter equations of Lemma A.17 from the trained state over the same horizon at the same eight cells (four learning rates times two decay strengths). The diffusion approximation under weight decay: the same scheme with $-\gamma_{\mathrm{wd}}R$ and $-2\gamma_{\mathrm{wd}}Q$ added to the drift (Lemma A.17), $\gamma_{\mathrm{wd}} = \eta\lambda_{\mathrm{wd}}$, at $\delta = 0.3$, $\eta \in \{0.5, 0.8, 1.2\}$, $\lambda_{\mathrm{wd}} \in \{0.01, 0.1\}$, $d = 500$, 1000 paths, horizon 3000, seed 3, reported as entry and persistent-entry fractions against the microscopic runs of Appendix D.4.

Minibatch: $b \in \{2, 4\}$ samples per step with the batch-mean gradient and one teacher move per step, at $\delta = 0.3$ ($d = 500$, 100 runs per cell, 3000-unit horizon). Test cells against the scaled fold of Lemma A.16 are $\eta \in \{0.8, 2.0\}$ for $b = 2$ and $\eta \in \{3.2, 5.6\}$ for $b = 4$. Its criterion stated in advance is at most 2 entries in the tracking cells and at least 98 in the entry cells.

Population-gradient reference: $b \in \{32, 500\}$ samples per step with the batch-mean gradient and one teacher move per optimizer step, at $\delta \in \{0.3, 0.65\}$ and $\eta \in \{0.7, 1.2, 2.0\}$ ($d = 500$, 32 runs per cell, horizon 1.2×10^4 units, the (Q, R, θ) means logged at 60 points). Learning time counts optimizer steps divided by d, so a batch-b run sees b times more inputs per unit of learning time at the same drift per step. The runs are compared against the order-parameter equations without the η^2 term over the same horizon and against the scaled fold of Lemma A.16. That fold lies at about 60 for $b = 32$ and 10^3 for $b = 500$.

Threshold rate: the threshold stepped at $\kappa\eta/d$ with $\kappa \in \{0.5, 2\}$. In the order-parameter equations the horizon-limited onset is found by bisection over η within $\alpha \le 1.2 \times 10^4$ at $\delta \in \{0.3, 0.45, 1.2\}$, with a four-fold horizon rerun at $\delta = 0.3$. In the microscopic runs at $d = 500$ (100 runs per cell, 3000-unit horizon) the cells are $(0.3, 0.8)$ and $(0.3, 0.5)$, chosen in advance. Two further conventions, a threshold stepped at $\eta/\sqrt{d}$ and at raw η, are run at the same two cells.

Recovery after the drift stops: at $(\delta, \eta) \in \{(0.3, 0.8), (0.3, 1.2), (0.65, 0.8)\}$ with $d = 500$ and 100 runs, the teacher drifts for 3000 units and is then frozen for 3000 units at the same learning rate, with the margin logged every 50 units in both phases. A run recovers when its margin falls below $m_\star$ at any second-phase log point after an entry.

C Evaluation Metrics and Tests of the Derivations

Every measured quantity and statistical procedure the paper reports is defined below, from the dormancy criteria of Appendix C.1 to the tests of the derivations of Appendix A.

C.1 Dormancy criteria and entry conventions

In the solvable model a run is dormant when its empirical margin $\theta/\|w\|$ reaches $m_\star = 2$, the population criterion of Definition 1 evaluated on the finite-d state. Entry is the first learning time at which the margin reaches $m_\star$ (first passage). A persistence alternative of Appendix D.2 instead takes the start of the first interval of ℓ units of learning time spent continuously at or above $m_\star$.

In the network experiments the activity of a unit is its activation probability: the fraction of fresh inputs drawn at each log point on which its pre-activation is positive. A unit is dormant when this activity is at most 0.023, the absolute criterion. In the layer-normalized score of Sokar et al. (2023) the mean post-activation magnitude of a unit is divided by the layer mean of that quantity, with cutoffs $\tau_{\mathrm{S}} \in \{0.025, 0.1\}$. Appendix D.3 reports an activity-based version next to that score and to the absolute criterion; it divides activation probabilities instead. The entry time of a unit in the network is the first log point with activity at most 0.023. The dormant fraction of a run is the fraction of hidden units dormant at the end of training, and the dormant fraction of a cell is the seed mean over its non-divergent runs.

C.2 Entry-time statistics and tests

The mean entry time of a cell is the mean over its entering runs with its standard error. Its entry fraction is the number of entering runs over the number of runs.

We measure agreement between predicted and measured entry times by the Spearman rank correlation across cells and by the median relative error $|T_{\mathrm{pred}} - T_{\mathrm{meas}}|/T_{\mathrm{meas}}$, before and after one global calibration factor c. The paper uses two timing statistics and names each where it appears: the conditional mean $\mathbb{E}[T \mid T \le H]$ over entering runs and the restricted mean $\mathbb{E}[\min(T, H)]$ over all runs. The calibration factor c is the in-sample median of the ratio of measured to predicted times, so leave-one-speed-out and split-half versions test whether it carries over to held-out cells.

A 50% crossing of a learning-rate scan comes from interpolating the entry fraction linearly between adjacent scan points. Its binomial posterior 95% interval comes from sampling the Jeffreys posterior of each per-rate entry fraction of the refined scan under a monotonicity constraint. We judge each outcome against the band stated in advance, not against the interpolated crossing.

We test exponential suppression in d by a weighted linear fit of the log entry fraction against d over the nonzero cells, with zero-count cells entering as 95% upper bounds only, and by a binomial-likelihood fit over all cells (added later).

In the network experiments all paired comparisons are two-sided Wilcoxon signed-rank tests over the ten seeds of a cell pair. With four drift speeds tested, a Bonferroni factor of four leaves every reported p significant at 0.05. Bootstrap 95% intervals are reported for the largest-rate drift-minus-control differences.

C.3 Tests of the derivations

Every closed form and each of the structures listed below is covered by a numerical test with an explicit pass threshold. For the derivation-level tests the threshold is a numerical-precision floor. For the predictive tests the threshold was on file before the data existed. Tests are listed in dependency order.

Consistency test of the closed forms: all atoms and moments of Appendix A.1 are compared against adaptive quadrature and against Monte Carlo with 4×10^6 samples, across a grid of (m, ρ, Q) including the $s \to 0$ edge. The criteria are relative error $\le 10^{-6}$ against quadrature (observed $\le 3 \times 10^{-9}$, the resolution of the reference quadrature itself) and three standard errors against Monte Carlo.

Predictive test of the order-parameter equations: predicted trajectories of (Q, R, θ) are compared with microscopic runs of the update rule (1) at $d = 4000$ with 24 runs per configuration, across drift speeds and learning rates spanning the phase diagram (the configurations are listed in the test script of the supplement). Its criterion is that the prediction lies within four combined standard errors at $\geq 95\%$ of checkpoints in every configuration.

Consistency test of the parameterization: Proposition 3 is evaluated on fixed points found independently by continuation (root-finding on the right-hand side of the order-parameter equations) across the scanned (δ, η) range. Here the criterion on file is relative error 10^{-8} (observed 10^{-14}). The fold locus $\det D\Psi = 0$ reproduces the continuation boundary $\eta_c(\delta)$ to 10^{-4} at six drift speeds.

Consistency test of the asymptotics, covering the four limits of Appendix A.5: $\eta^* \to 2$ on C_0 ($\leq 10^{-8}$ at all checked margins), $\rho_0 \to 1/\sqrt{2}$ with its rate, the $2/\pi$ approach, and the location and value of η_0 (10^{-6}). The cancellations in the $2/\pi$ limit are checked symbolically and numerically.

Predictive tests of the rate law and the scaling: the entry-time integration of Section 4.5 is compared with the simulation grid, and the $d^{-1/2}$ fit makes an out-of-sample prediction whose band was fixed before the run. Both are reported in Section 5.

Predictive test of the basin structure: the reading of the fold as a basin boundary is compared with the initial-condition sweep and the random-initialization simulations (Appendix D.1).

One script runs the three consistency tests and the order-parameter test end to end and exits nonzero on any failure. It is part of the supplementary code together with all simulation and analysis programs.

D Additional Experiments

Experiments that Section 5 cites without showing are gathered here by the question they serve. Protocols are in Appendix B, metrics in Appendix C, and the criteria stated in advance in Appendix D.5 (Table 7).

D.1 Question 1: where the boundary lies

Boundary crossing. Section 5.2 states that the measured crossing lies within 0.003 of the diffusion prediction; the protocol is the boundary scan of Appendix B.1, and Table 2 holds the comparison. On the refined scan the 50% crossing falls between $\eta = 0.675$ $(30/100)$ and 0.680 $(57/100)$, and the count reaches $87/100$ at $\eta = 0.69$. These fractions give $\eta = 0.679$ with posterior 95% interval $[0.677, 0.681]$ (Appendix C.2). This crossing agrees with the coarse-scan estimate 0.678. The boundary band ± 0.06 was stated in advance as a tolerance on the fold: a half-width of three grid steps. At $d = 500$ and horizon 3000 the observed shift of the crossing above the fold $(+0.049)$ lies within this band.

Each measured crossing lies between the horizon-free dormancy boundary and the deterministic onset of the order-parameter equations from the trained state within the horizon of the scan itself (Table 2; the boundary is the fold at $\delta = 0.3$ and 0.45, at most 0.0033 above it at 0.59, and 0.5597 at 1.2). At every scanned speed the finite-d crossing lies below the onset because fluctuations advance entry and above the boundary because the horizon censors slow entry near it. The $+0.049$ shift at $\delta = 0.3$ relative to the fold is a -0.052 shift relative to the same-protocol deterministic onset. At the main-grid horizon 3000 the deterministic onset runs from 1.388 at $\delta = 0.05$ through 0.731 at 0.3 to 0.712 at the end point and 0.702 at $\delta = 3$; it is monotone and above the dormancy boundary throughout, although the fold itself ends at the end point (Figure 1, thin dashed). Along the computed boundary the ratio δ/η varies by more than three orders of magnitude, so the boundary is no level set of it.

Diffusion approximation: the order parameters $X = (Q, R, \theta)$ follow $dX = F(X)\, d\alpha + \sqrt{\Sigma(X)/d}\, dW$ with F the right-hand side of the order-parameter equations, W a three-dimensional Wiener process and Σ the fluctuation covariance of Appendix D.2. The equations are integrated by the Euler–Maruyama scheme at step 0.5 with 1000 paths per learning rate (protocol in Appendix B.7). Entry is the margin reaching $m_\star$ within the horizon of the scan, and the 50% crossing is obtained by linear interpolation of the entry fraction. In full: $F(X)$ is the right-hand side of

(2)–(4) and $\Sigma(X)$ the covariance of d times the increment of X per step under the same Gaussian closure, so that Σ/d is the covariance per unit of learning time. Both are evaluated at the current state at every step, F from the closed forms and Σ by Gauss–Legendre quadrature over the active side of the input. The noise of the scheme is $\Sigma^{1/2}z$ with the symmetric square root (eigenvalues below zero set to zero) and z standard normal. Each step is 0.5 units of learning time (halving it moves the readings by at most 0.6%, Appendix D.2). Q is floored at 10^{-6} and no projection onto $Q \geq R^2$ is applied. Neither the floor nor $R^2 > Q$ occurs at any step of the $36{,}000$ paths of the main grid.

This closure assumes increments of order $1/d$ whose accumulated fluctuations are Gaussian to leading order (the microscopic increments themselves need not be). It is therefore meant for large d and for states away from $Q = 0$.

The norm increment contains one more term of order $1/d$ that the closure omits: $(\eta/d)^2\chi^2\|\xi\|^2$ has mean $\eta^2\mathbb{E}[\chi^2]$ plus $(\eta^2/d)\,\mathbb{E}[\chi^2(y^2 + z^2 - 2)]$, with (y, z) the standard coordinates of the input in the student–teacher plane. The extra term arises because $\|\xi\|^2$ fluctuates with χ through those two coordinates. The diffusion approximation adds this term to the norm drift at every step by two-dimensional Gauss–Hermite quadrature. Every diffusion reading below includes this term unless it is labelled leading order; the dimension scan of Appendix D.2 reports both. The overlap increment contains an $O(1/d)$ term of the same order from the update ordering, $-(\delta\eta/d)\,\mathbb{E}[\chi y]$, because the teacher moves after the student updates. This ordering term is not part of the diffusion approximation, and adding it moves the readings below by less than their own resolution (the convergence check at the end of this part).

At $(0.3, 500)$ the diffusion approximation predicts 0.679 against the measured 0.679 $[0.677, 0.681]$, inside the interval. At $(0.45, 500)$, $(0.59, 500)$, $(0.59, 1000)$, $(0.59, 2000)$, and $(0.59, 4000)$ it predicts 0.591, 0.580, 0.588, 0.594, and 0.598 against 0.590, 0.580, 0.588, 0.592, and 0.598 (Table 2). The leading-order scheme gives 0.684 (outside the interval) and 0.597, 0.586, 0.591. Its offsets of 0.003 to 0.007 are due to the omitted norm-drift term. Corrected differences are 0.000, 0.001, 0.000, and 0.000, whereas leading-order differences are 0.0055, 0.0070, 0.0055, and 0.0029. The corrected shift from the same-horizon deterministic onset matches the measured shift (-0.052 against -0.052 at $\delta = 0.3$). The leading-order scheme recovers 60 to 90% of it and lies above the measured crossing at every cell. At $(0.3, 1000)$ the diffusion approximation predicts 0.694 against a measured 0.693 $[0.691, 0.695]$, above the $d = 500$ value, as the measured crossings at $\delta = 0.59$ also rise with d.

Every difference is below 0.003, on both sides of the end point (Table 2). Beyond the end point at $\delta = 1.2$ (protocol in Appendix B.7) the diffusion approximation predicts 0.637 $[0.636, 0.638]$ within horizon 3000 against the measured 0.640 $[0.636, 0.642]$. Within 12000 it predicts 0.565 $[0.565, 0.566]$ against the measured 0.566 $[0.565, 0.568]$. The latter crossing is resolved by a scan at rates 0.50 to 0.568 joined to the earlier 0.57 and 0.58 cells (91 and 100 of 100). It gives 0 of 100 up to 0.55; 4, 11, 33, 49, and 64 of 100 at 0.56, 0.562, 0.564, 0.566, and 0.568. At $\delta = 0.45$ and 0.59 within 3000 it predicts 0.656 and 0.648 against the measured 0.656 $[0.653, 0.658]$ and 0.648 $[0.645, 0.650]$. At four digits the corrected crossing lies inside the measured interval at nine of the eleven comparisons. At $(0.59, 2000, 12000)$ it lies 0.002 above the crossing of a scan with grid step 0.01. At $(0.45, 500, 12000)$ the prediction 0.5914 lies 0.001 above the linearly interpolated crossing and 0.0015 below the probit centre 0.5929 of the same scan.

Diffusion intervals are 95% intervals from binomial resampling of the path fractions; the measured ones are posterior intervals under monotonicity. A probit fit of the entry fraction against the rate gives a crossing within 0.0003 of the linearly interpolated one for every diffusion scan and within 0.0025 for the measured scans (protocol in Appendix B.7). The largest is at $(0.45, 500, 12000)$, grid step 0.01. The locally refined scan at $(1.2, 500, 12000)$ (11, 33, 49, and 64 of 100 at 0.562 to 0.568) resolves the sharpest transition directly. It widens that measured interval to 0.002 and leaves the corrected crossing inside it.

Unconditional readouts over the main grid place paths that do not enter within the horizon at the horizon. The restricted mean $\mathbb{E}[\min(T, H)]$ of the diffusion approximation then has median relative error 1.3% over the thirty entering cells (largest 4.8% at $(0.65, 0.9)$) and Spearman 0.999 against the measured restricted means. The leading-order scheme gives 3.7%, with its largest error 22% at $(0.1, 1.0)$. There it enters in 0.78 of the paths against 100 of 100 runs, while the diffusion approxi-

mation enters in 0.995. By contrast the diffusion entry fraction is within 0.1 of the measured one at every cell.

For the leading-order scheme, halving the integration step to 0.25 at three cells moves the entry fraction by at most 0.021 and the mean first passage by at most 0.6% ($(0.1, 1.0)$: 0.781 to 0.802 and 2511 to 2521 units; $(0.3, 0.8)$: 1122 to 1115; $(0.65, 1.2)$: 264 to 265), within 0.03 and 3%. At $(0.1, 1.0)$ the diffusion approximation is converged: every shift lies within 0.02, 3%, and 0.002 (protocol in Appendix B.7). Halving the step moves its entry fraction from 0.995 to 0.993 and its restricted mean by 0.06%, and its persistent-entry fraction (Appendix D.2) from 0.965 to 0.982. 96×96 quadrature in place of 48×48 moves them by less than 0.001 and 0.01%. Adding the ordering term of the overlap drift moves them by less than 0.001 and 0.2%, and moves the crossing at $(0.3, 500)$ by 2×10^{-5}.

Table 2: Boundary crossings against the fold and the deterministic onset within the horizon at every speed, dimension, and horizon compared. Brackets are 95% intervals (diffusion: binomial resampling of the path fractions; measured: posterior under monotonicity). Beyond the end point there is no fold.

δ	d	horizon	fold	deterministic onset	diffusion approximation	measured crossing
0.3	500	3000	0.629	0.731	0.6789 [0.6783, 0.6795]	0.6788 [0.6775, 0.6806]
0.3	1000	3000	0.629	0.731	0.6939 [0.6933, 0.6944]	0.6930 [0.6917, 0.6948]
0.45	500	3000	0.581	0.717	0.6558 [0.6548, 0.6565]	0.6563 [0.6539, 0.6580]
0.45	500	12000	0.581	0.614	0.5914 [0.5912, 0.5916]	0.5904 [0.5900, 0.5908]
0.59	500	3000	0.563	0.712	0.6476 [0.6470, 0.6482]	0.6479 [0.6450, 0.6499]
0.59	500	12000	0.563	0.607	0.5797 [0.5793, 0.5800]	0.5798 [0.5793, 0.5803]
0.59	1000	12000	0.563	0.607	0.5879 [0.5877, 0.5881]	0.5879 [0.5874, 0.5885]
0.59	2000	12000	0.563	0.607	0.5935 [0.5933, 0.5937]	0.5915 [0.5908, 0.5923]
0.59	4000	12000	0.563	0.607	0.5977 [0.5976, 0.5979]	0.5976 [0.5974, 0.5979]
1.2	500	3000	none	0.705	0.6372 [0.6365, 0.6379]	0.6396 [0.6365, 0.6420]
1.2	500	12000	none	0.597	0.5654 [0.5652, 0.5656]	0.5661 [0.5652, 0.5671]

The slow-drift plateau at finite dimension. Section 4.4 anchors the boundary at η_0 as $\delta \to 0^+$; here the drift speed moves to $\delta \in \{0.003, 0.01\}$ with 50 runs at a 12000-unit horizon, and two stationary controls at the same rates are the comparison. The finite-dimension crossings lie within 0.1 of the fold. Microscopic runs near the plateau use $d = 500$, 50 runs, horizon 12000 (protocol in Appendix B.7). For $\delta = 0.003$ they enter in 0 of 50 at $\eta \le 1.9$ and 50 of 50 at $\eta \ge 2.0$, mean entry 3849, 1865, 1199 at 2.0, 2.1, 2.2; for $\delta = 0.01$ in 0 of 50 at 1.8 and 50 of 50 from 1.9 (2435 to 636). The 50% crossings are therefore 1.95 and 1.85 against 1.968 and 1.87 of the fold, within 0.1.

The stationary controls at the same five rates, runs, horizon, and seed identify the plateau. A fixed teacher under label noise 0.2 enters in 0 of 50 at $\eta \le 1.9$ and 50 of 50 from 2.0 (mean entry 5217, 2282, 1383), a 50% crossing at 1.95. A noiseless fixed teacher from a random initialization converges onto the teacher at $\eta \le 1.9$, neither converges nor enters at 2.0 (norm 15 and margin 1.6 at the horizon), and enters in 50 of 50 from 2.1 (3220, 1721). Its crossing is at 2.05. The part of the plateau specific to drift is the difference: at $\delta = 0.003$ the onset in rate coincides with the noisy stationary onset at the grid resolution of 0.1 while entry is faster (mean 3849 against 5217 at 2.0), and at $\delta = 0.01$ it lies 0.1 below. Both lie 0.1 to 0.2 below the noiseless stationary onset, the direction Proposition A.10 gives for the asymptotic stationary branch. The slow-drift anchor $\eta_0 = 2.022$ is the largest rate at which the stationary branch C_0 of Proposition A.10 exists (the trained state itself is a fixed point at every rate).

The onset across the end point and its horizon dependence. Section 4.4 states that the onset decreases monotonically through the end point at one horizon and that the horizon-free boundary is continuous; the scans move past δ_{cp} and vary the horizon, and Figures 4 and 5 hold the readings. At any common horizon the deterministic onset is monotone across the end point. An apparent minimum arises only from mixing the horizon-free fold value with finite-horizon crossovers. That horizon-free boundary itself is continuous (below). Post-end-point onset: for $\delta > \delta_{\mathrm{cp}}$ the onset plotted in Figure 1 is computed by bisection on first passage of the margin to $m_\star$ within $\alpha \le 1.2 \times 10^4$ (protocol in Appendix B.7). At $m_\star = 2$ it runs from 0.606 at $\delta = 0.6$ through 0.594 at 2 and 0.592 at 3 to 0.591 at 5 (the speeds 2, 3 and 5 integrated at steps 0.5 and 0.25, which agree to four digits).

At the end point the deterministic onset depends on the integration horizon. On the crossover side it moves down as the horizon grows, toward the cutoff-locus value $\Psi(m_\star, \cdot) = 0.5656$ (the learning rate at which the fixed-point margin equals the cutoff) rather than to $\eta_{\mathrm{cp}} = 0.5623$. At $\delta = 0.595$ (just past δ_{cp}) the onset is 0.607 within $\alpha \le 1.2 \times 10^4$, 0.577 within 4.8×10^4, and 0.568 within 1.92×10^5, a gap to η_{cp} of 0.044, 0.015, and 0.006. The learning-rate grid step is 0.003; the first-passage time at $\eta = 0.568$ is 1.29×10^5. That limit lies above η_{cp} by 0.0033 because the fixed point at the end point has margin $1.35 < m_\star$. A scan of the fixed points and their stability (protocol in Appendix B.7) confirms this cause: at every speed from 0.59 to 0.62 a stable fixed point exists at every η from 0.550 to 0.5975. Its margin rises steeply from 0.93 at $\eta = 0.560$ to 1.59 at 0.5625 and 1.95 at 0.565, and crosses $m_\star$ between 0.565 and 0.5675. Trajectories from the trained state at $\delta \in \{0.5939, 0.595\}$ and $\eta \in \{0.5623, 0.5635, 0.5645, 0.5656\}$ converge within 2×10^5 units to margins 1.26 to 1.93 without crossing the cutoff. Trajectories at $\eta = 0.568$ and 0.575 first pass at 1.3×10^5 and 4.75×10^4 units and settle at margins 2.07 and 2.32. The gap between η_{cp} and 0.5656 is therefore set by the margin of a stable state and not by the horizon.

Where the cutoff curve leaves the fold (protocol in Appendix B.7): $\Psi(m_\star, \cdot)$ lies below the fold for $\delta < 0.546$, by 0.053 at $\delta = 0.3$ and 0.011 at 0.46. It crosses the fold at $(\delta, \eta) = (0.5456, 0.5666)$ and lies above it from there to the end point, by 0.0033 at δ_{cp}. Trajectories from the trained state started 0.001 above the fold cross the cutoff at $\delta \in \{0.1, 0.3, 0.45, 0.55\}$ after 7000, 20000, 65000, and 347000 units. At $\delta = 0.59$ the same start converges to a stable state with margin 1.81 (late slope 1.4×10^{-7} per unit), and the cutoff is crossed only from 0.003 above the fold, at 3.8×10^5 units. No fixed point with $\rho \in (0, 1)$ exists just above the fold at $\delta \le 0.3$. At 0.45 and 0.55 the stable state that appears above it has margin 3.1 and 2.2, above the cutoff. At 0.55 the cutoff curve lies 0.0004 above the fold, inside the first step of the 0.0025 scan.

Just past the crossing point $\delta = 0.5456$ a band of learning rates above the fold ends in an active state below the cutoff. A finer scan of the crossing point (protocol in Appendix B.7): the fold and the cutoff curve are 0.56637 and 0.56657 at $\delta = 0.548$, 0.56615 and 0.56653 at 0.55, and 0.56584 and 0.56646 at 0.553. Trajectories from the trained state started 0.0001 to 0.0005 above the fold at 0.548, and 0.0001 to 0.0008 above it at 0.55 and 0.553, converge within 4×10^5 units to margins between 1.79 and 2.00 without crossing the cutoff. Those started 0.0008 and 0.0012 above cross after 3.7×10^5 and 2.8×10^5 units at 0.548. The ones started 0.0012 above cross after 3.0×10^5 and 3.4×10^5 units at 0.55 and 0.553. The width of the band grows with δ as the two curves separate. At 0.59 it has margin 1.66 at the fold and 1.98 at 0.0025 above it.

The horizon-free dormancy boundary is therefore one continuous curve: the fold up to $\delta = 0.546$ and the cutoff curve beyond. The end point of the fold is not a point of it. A trajectory above the fold converges below the cutoff for δ from 0.546 to δ_{cp}, and not at slower drift.

At $\delta \in \{0.45, 0.55\}$ every stable fixed point above the fold on the 0.0025 grid has margin above the cutoff and every trajectory from the trained state crosses it (the grid skips the band of width 0.0004 just above the fold at 0.55, resolved by the fine scan above). At $\delta = 0.59$ the stable fixed points within 0.0025 of the fold have margin below 2 and the trajectories started there converge to them without crossing (Figure 4).

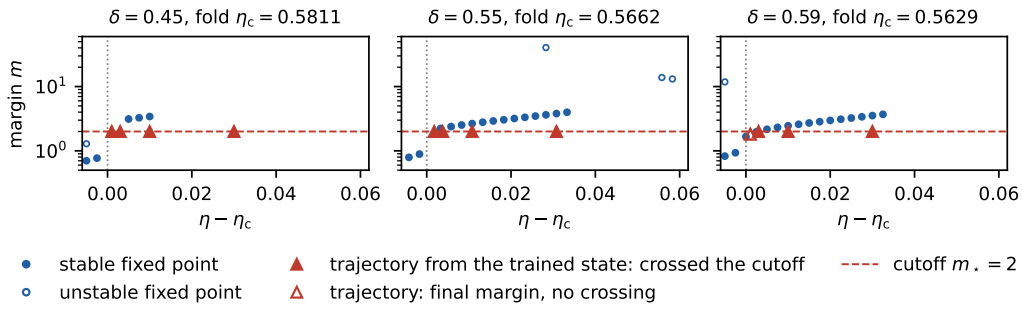


Figure 4: Fixed points near the fold at three drift speeds: margin against the learning rate above the fold; filled: stable; open: unstable; triangles: the trajectories from the trained state (filled: crossed the cutoff; open: final margin without crossing).

The fold branch does not depend on the cutoff convention, while the post-end-point onset does (Figure 5). Convention sensitivity (steps 0.5 and 0.25, agreeing to four digits): at $m_\star = 1.5$ the onset declines from 0.565 at $\delta = 0.7$ through 0.556, 0.549 and 0.545 at 1, 1.5 and 2 to 0.542 at 3, below the end-point value. At $m_\star = 2.5$ no first passage occurs below $\eta = 0.70$, the top of the scan, at any of the five probed drift speeds.

Measured onset across the end point: the scan at $\delta \in \{0.45, 0.59, 1.2\}$ ($d = 500$, the 3000-unit horizon) measures 50% crossings 0.656 $[0.653, 0.658]$, 0.648 $[0.645, 0.650]$, and 0.640 $[0.636, 0.642]$: monotone decreasing, with no interior minimum at this dimension and horizon. On the fold side the crossings drop to 0.590 $[0.590, 0.591]$ at $\delta = 0.45$ and 0.580 $[0.579, 0.581]$ at 0.59, offsets $+0.009$ and $+0.017$ from theory. Finite-window censoring therefore accounts for nearly all of the short-horizon shift. Past the end point the crossing is 0.566 $[0.565, 0.568]$, below the deterministic onset at 0.597. The lowest cell of the window scan (0.57) enters in 91 of 100, so the crossing lies below the window, and the finer scan of the boundary protocol above resolves it. With no fold barrier, fluctuation-activated entry undercuts the deterministic onset. The measured onset at $d = 500$ therefore remains monotone decreasing across the end point at both horizons, and the interior-minimum criterion fails on both.

No minimum appears at any tested common horizon. At the common horizon 1.2×10^4 the onset is 0.607 on the fold side ($\delta = 0.59$) and 0.597 past the end point ($\delta = 1.2$). Under longer common horizons the deterministic onsets on both sides fall together and remain monotone: 0.577 against 0.567 within 4.8×10^4, and 0.568 against 0.561 within 1.92×10^5. The horizon-free boundary (the fold up to $\delta = 0.546$ and $\Psi(m_\star, \cdot)$ beyond) also decreases monotonically through the end point.

Horizon asymptotics: the long-horizon onset is a δ-dependent curve that coincides with the cutoff curve $\Psi(m_\star, \cdot)$ rather than tending to η_{cp}. Lengthening the horizon four-fold from 1.92×10^5 to 7.68×10^5 moves the deterministic onset by at most 0.0007 at $\delta \in \{0.8, 1.2, 2.0\}$ ($0.5632 \to 0.5625$, $0.5600 \to 0.5598$, $0.5576 \to 0.5573$; bisection resolution 2.4×10^{-4}, step 1.0). That is below 0.002 at each speed. That curve decreases slowly with δ, within 0.0002 of $\eta_{\mathrm{cp}} = 0.5623$ at $\delta = 0.8$ and below it from $\delta = 1.2$ onward. It does not approach η_{cp} as a common limit: the three limits spread by 0.0051. This curve is a closed-form first-passage threshold rather than a bifurcation. Within the bisection resolution it coincides with the image curve $\Psi(m_\star, \rho)$ of Proposition 3 at $m = m_\star$, the locus where the margin of the fixed point equals the cutoff. The curve gives 0.5626, 0.5597, 0.5574, 0.5562 at $\delta = 0.8$, 1.2, 2.0, 3.0 against the measured 0.5625, 0.5598, 0.5573, 0.5564. Along that curve $\eta^* \to \eta^*(m_\star, 0) = 0.554$ as $\delta \to \infty$, and $\to 2/\pi$ as $m_\star \to \infty$. That limit is the sense in which $2/\pi$ is the envelope. At $\delta = 3.0$ the step-0.5 run gives an onset of 0.5564 at both horizons (shift $< 2.4 \times 10^{-4}$, the bisection resolution), below the $\delta = 2.0$ limit and consistent with the slow decrease in δ. At $\delta = 2.0$ the step-0.5 rerun reproduces the step-1.0 values to all quoted digits, so the step-1.0 results above are integrator-converged. Figure 5 collects the onset-versus-horizon data.

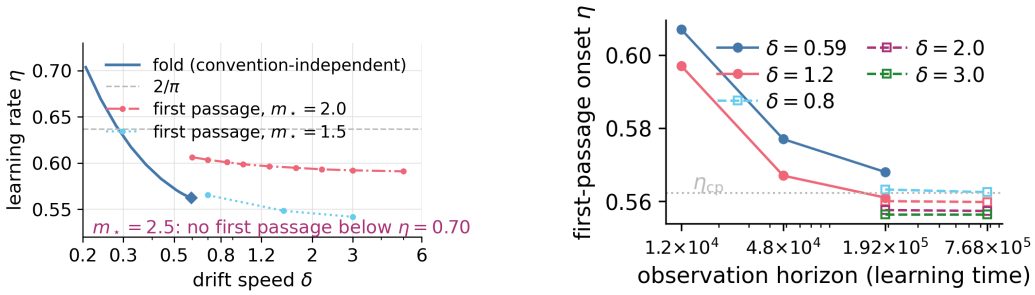


Figure 5: Left: the post-end-point deterministic onset (horizon 1.2×10^4) for $m_\star = 2$ and 1.5 against the convention-independent fold (solid, diamond at its end point) and the $2/\pi$ line (dashed); at $m_\star = 2.5$ no first passage occurs below $\eta = 0.70$. Right: deterministic onset versus observation horizon; circles: $\delta = 0.59$ and 1.2; open squares: the four speeds of the horizon-asymptotics scan; dotted: η_{cp}.

Basin sweep and random initializations. Section 4.4 states that the fold bounds tracking from the trained state and nearby starts; only the initial state changes (protocol in Appendix B.4), and the entry counts by lattice cell are the reading. Within the scanned box the fold is the boundary of tracking

for starts with $Q \leq 1$. Above the fold (33 cells) every lattice trajectory enters, and no bounded non-entering trajectory appears anywhere in the sweep, with no sign of competing attractors or cycles. Below the fold the near-trained subset ($\rho \geq 0.7$, $m \leq 0.5$) produces zero entries in ten of thirteen cells. In the three cells nearest the boundary (within 0.013 of η_c) exactly its inflated-norm corner $Q = 4$ enters, 3 of 18 initial conditions each. Every $Q \leq 1$ member of the subset tracks in every below-fold cell. The tracking basin contracts in the norm direction as the fold approaches, as it must since the tracking state annihilates at the fold.

From random finite-d initializations the fate matches the order-parameter equations started from the matched state $(Q, R, \theta) = (1, 1/\sqrt{d}, 0)$ at all three cells chosen in advance ($w \sim \mathcal{N}(0, \mathbf{I}/d)$, $\theta = 0$; $d = 500$, 100 runs per cell). The two above-fold cells enter in 100 of 100 runs each, with mean entry times 1067 ± 155 and 350 ± 70 against 1827 and 541 from the order-parameter equations. The below-fold cell produces zero entries while the order-parameter equations track. Averaging the prediction over 40 signed overlaps from $\mathcal{N}(0, 1/d)$ in place of the typical overlap $1/\sqrt{d}$ leaves both times at 1827 and 541 to the nearest unit. Entry classifications agree. The measured times lie 42% and 35% below the order-parameter equations, a finite-size acceleration of the same order as on the main grid before calibration. This comparison tests the order-parameter equations outside the trained-state regime, because a random start has near-zero alignment.

Anatomy of the near-boundary entry time. The entry-time law of Section 4.5, its divergence at the fold, and its fitted exponent are examined here with the order-parameter equations alone; the window times and exponents below are the reading. Write $\epsilon = \eta - \eta_c(\delta)$ for the distance above the boundary, which at these speeds is the fold. Below the boundary the leading eigenvalue obeys $\lambda_{\max}^2 \propto (\eta_c(\delta) - \eta)$ with fitted exponent 1.05 ± 0.01 close to the fold. Farther from the fold the exponent increases. The normal-form coefficients are nonzero at every scanned drift speed.

At a representative drift speed in the folded regime ($\delta = 0.3$) the passage from the trained state to dormancy splits into three consecutive segments, each timed separately: an approach to the vicinity of the annihilated fixed point, the passage through the narrow window around it, and a creep from that window to the cutoff. The approach segment is short and nearly independent of ϵ. As $\epsilon \to 0$ the creep segment saturates above 10^4 time units, far longer than the approach but bounded. The fold-window passage instead diverges. Its local exponent (of the window time against ϵ, fitted over one decade around each point) moves from about -0.7 at $\epsilon = 10^{-3}$ to -1 at $\epsilon = 10^{-1}$. Both are more negative than the $\epsilon^{-1/2}$ ghost scaling of a generic saddle-node bottleneck over the tested finite-distance range.

Integrating the one-dimensional normal form of the fold between the measured entry and exit states of the window reproduces every measured window time (within 1.2% at the small-ϵ end, 18% at $\epsilon = 10^{-1}$, median 2.2%). The second eigenvalue of the fold Jacobian is larger by many orders of magnitude, so slow-mode mixing is excluded. The naive ghost formula $\pi/\sqrt{ab\epsilon}$ from the normal-form coefficients alone, without edge matching, overestimates the same window times by 23% to 470% across the grid, so an edge correction of this size masks the asymptotic exponent over this range.

D.2 Question 2: how long entry takes

Calibration checks and residuals. Section 5.3 states that the entry-time law ranks the entering cells; the protocol is the main grid of Appendix B.1 with the dimension scans of Appendix B.2, and Figure 2b and Figure 3a hold the readings. The 95% interval of the Spearman correlation over the thirty entering cells is $[0.93, 0.99]$. After the single fitted factor $c = 0.642$ over 30 cells the median error is 7.9% (95% interval $[3.9, 11.0]\%$ from a cell bootstrap with the factor refitted in each resample) and the largest residual 25%. The mean entry time at $(0.3, 0.8)$ rises from 249 at $d = 30$ to 1452 at 1920 against the predicted limit 1830 and the band 1460 ± 40 predicted before the run. The entry-fraction log-slope in d below the boundary at $(0.3, 0.62)$ is -0.055 ± 0.002. Horizons: in every entry cell all 100 runs enter well within their horizon, and the per-cell horizon rule ($\min(3000, 4T_{\text{pred}})$) plays no censoring role in any reported number.

The finite-d deficit of the deterministic law is the fluctuation advance of first passage. The Gaussian closure reproduces it to a few percent. Diffusion approximation of the entry times, leading-order scheme (protocol in Appendix B.7): the stochastic order-parameter equations of Appendix D.1 are

integrated at $d = 500$ with 1000 paths per cell over the main-grid horizon 3000. At every one of the thirty entering cells they give a mean first-passage time with no fitted parameter. Against the measured 100-run means over entering runs Spearman is 0.999, the median error 3.5% (largest 17.3% at $(0.1, 1.0)$, where the leading-order entry fraction is 0.78), and the measured-over-scheme ratio 0.85 to 1.02 with median 0.97. The six tracking cells give leading-order entry fraction 0 against measured 0 at five of them and 2 of 100 at $(0.15, 0.8)$. The fitted factor $c = 0.642$ calibrates the deterministic law and is not needed once the fluctuations are integrated.

The diffusion approximation reproduces the shape of the entry-time distribution and not only its mean. Distribution-level comparison (protocol in Appendix B.7) keeps the first-passage time of every path. The Kolmogorov distance between the first-passage distribution of the diffusion approximation and the measured one is at most 0.2 at all thirty entering cells. Its restricted mean has median relative error 1.3% and its conditional mean over entering paths 1.27%. The leading-order scheme has quantile errors of 3.4% in the median and Kolmogorov distance at most 0.2 at 24 of the 30. Six cells of the leading-order scheme lie above 0.2: $(0.1, 1.0)$ at 0.65, $(0.1, 1.1)$ at 0.34, $(0.65, 0.7)$ at 0.27, and $(0.2, 1.0)$, $(0.2, 0.8)$, and $(0.1, 1.2)$ at 0.20 to 0.22. There it runs later than the measured one by 5 to 25% of the median while the distributions are narrow. The leading-order entry fraction lies outside the measured 95% Wilson interval at three cells: $(0.1, 1.0)$, 0.78 against 100 of 100; $(0.3, 0.7)$, 0.945 against 100 of 100; and $(0.15, 0.8)$, 0 against 2 of 100. Its persistent-entry fraction lies outside only at $(0.1, 1.0)$. The leading-order scheme therefore reproduces the shape of the entry-time distribution as well as its mean over four fifths of the grid, and the lag it retains is mostly at the slowest drift speeds.

The lag of the leading-order scheme vanishes with d, but its under-entry at $(0.1, 1.0)$ does not. Across dimensions at those cells (protocol in Appendix B.7; these dimension runs use the leading-order scheme): at $(0.1, 1.2)$ the lag of the leading-order scheme in the median first passage falls from 5.8% at $d = 250$ to 0.7% at 2000. At $(0.1, 1.0)$, where the deterministic first passage equals the horizon, it falls from 36% at $d = 250$ to 8% at 1000 while the leading-order entry fraction stays at 0.78 to 0.82 against a measured 0.97 to 1.0.

The leading-order under-entry (0.78 against 1.0) is confined to the one rate at which the deterministic first passage equals the horizon and the late tail of the scheme falls beyond it. In a neighbourhood of that cell at $\eta \in \{0.925, 0.95, 0.975, 1.0, 1.025, 1.05\}$ and $\delta = 0.1$ (leading-order runs; protocol in Appendix B.7), the diffusion and the microscopic runs agree on no entry at the three lowest rates and full entry at the two highest. The median first passage is 1667 against 1531 at 1.025 and 1290 against 1192 at 1.05. Their 50% crossings in η are 0.991 and 0.9875, so the onset is placed correctly.

The leading-order moments of the closure are not the cause of the under-entry (protocol in Appendix B.7). Conditional on the state, the mean increments of the overlap and the threshold equal the closed moments at every d up to the $O(1/d)$ ordering term of Appendix D.1. The norm increment differs from its closed moment by the $O(1/d)$ term of that appendix. At states of the trajectory at $(0.1, 1.0)$ with margins 0.5, 1, 1.5, and 2 the Monte Carlo means of the exact one-step increments at $d = 500$ (five replicates of 2×10^5 inputs) lie within two standard errors of the closed forms in every component. The quadrature the diffusion integrates reproduces those closed forms to five digits.

At the slowest and fastest grid speeds the ratio of measured to predicted time rises monotonically toward 1 with d. Dimension scans at the slowest and fastest grid speeds: at $(\delta, \eta) = (0.1, 1.2)$ the ratio of measured to predicted time rises from 0.852 at $d = 500$ to 0.921 at 1000 and 0.955 at 2000. At $(0.65, 0.8)$ it rises from 0.584 to 0.692 and 0.782. Each cell has 100 runs, all entering; the $d = 500$ values are the main-grid cells under the per-cell horizon rule. The gap to 1 falls to a third between $d = 500$ and 2000 at the slow speed, $0.148 \to 0.045$. At the fast one it falls just short of halving, $0.416 \to 0.218$, a 48% reduction.

The leading-order scheme reproduces the dimension dependence of the deficit, whereas the diffusion approximation over-corrects below $d \approx 500$. Diffusion approximation across the dimension scan (protocol in Appendix B.2): the leading-order scheme gives the open circles of Figure 3a and the diffusion approximation the filled ones. At $(0.3, 0.8)$ the leading-order mean entry time lies within 6.4% of the measured mean at $d = 30$ and within 4.5% at every larger dimension (1.0% at 1920). The log-log slope of the deficit $1830 - \bar{T}$ against d is -0.354 for the leading-order scheme against the measured -0.342 (bootstrap interval -0.355 to -0.329). The deficit exponent is therefore shal-

lower than $-1/2$ over this range. The diffusion approximation, run at the same seven dimensions, undershoots the measured mean at small dimension by 21.1%, 14.5%, and 5.4% at $d = 30$, 60, and 120. It lies within 2.1% at 240 and 480 and within 0.2% at 960 and 1920 (deficit slope -0.351). Its $O(1/d)$ norm-drift term over-corrects where $1/d$ is no longer small, and the two schemes merge from $d \approx 500$.

Fluctuation covariance, action minimization, and the activated channel. Section 4.5 states that below the boundary entry at finite d is a fluctuation-driven escape; here the cell moves below the fold and the dimension is varied. The escape interpretation predicts entry rates that fall with dimension at fixed environment.

The stochastic-gradient channel contributes the covariance of the per-example increments $(\eta\chi y,\ 2\eta\chi x+\eta^2\chi^2,\ -\eta\chi)$, and the teacher drift adds $2\delta(Q-R^2)$ to the variance of R (Lemma A.3). These components are listed in the (R, Q, θ) order of the order-parameter equations; the integrator stores the state as (Q, R, θ) and permutes between the two orders.

A path-action minimization below the boundary failed its robustness checks (Table 7); no barrier value is reported.

The scan at $(\delta, \eta) = (0.3, 0.60)$ gives zero entries in 500 runs at each of $d \in \{120, 240, 480\}$ within the 3000-unit horizon, so it yields no exponent and no rate against which a barrier estimate could be checked (Table 7).

Below the boundary the entry fraction falls exponentially with d. The follow-up at the nearer cell $(\delta, \eta) = (0.3, 0.62)$, chosen for intermediate entry probability, gives entry fractions 0.820, 0.189, 0.054, 0.012, and 0.001 at $d \in \{60, 90, 105, 120, 135\}$ (500, 1000, 1000, 500, and 1000 runs), and zero entries in 1000, 500, and 500 runs at $d = 150$, 240, 480. Zero-count cells enter as 95% upper bounds only and are excluded from the fit. Across the five nonzero dimensions a weighted linear fit of the log entry fraction against d (weights $np/(1-p)$) gives slope -0.055 ± 0.002 per unit of dimension, an exponential-in-d suppression. A binomial-likelihood fit over all eight dimensions including the three zero-count cells gives slope -0.059 (95% profile interval $[-0.062, -0.056]$), slightly steeper, as expected once censored cells are included. Beyond the criterion, the local slope steepens from -0.05 to -0.17 per unit across the window, and the decay is exponential or faster there.

The diffusion approximation reproduces this suppression (protocol in Appendix B.7). At the same cell and dimensions it gives entry fractions 0.860, 0.239, 0.081, 0.023, 0.005, 0.0015, 0, and 0 (1000 paths, 4000 from $d = 150$), within 0.05 of the measured ones everywhere. Its weighted slope of the log fraction is -0.051 ± 0.001 per unit of dimension, inside the interval $[-0.062, -0.050]$ stated in advance. The binomial-likelihood slopes are -0.055 $[-0.057, -0.053]$ against the measured -0.059 $[-0.062, -0.056]$: the scheme decays slightly more slowly than the unit itself.

Below-boundary contrast: a cell 0.03 below the fold produces no entries in 1500 runs across three dimensions, while a cell just below the boundary produces 2 entries in 100 runs at $d = 500$. The deep cell shows suppression beyond observability and the near cell a small hazard, in the direction of the activated channel of Section 4.5.

Recrossings and the persistence of first passage. For the first-passage convention of Definition 1, the margin is read after first passage at two entering cells. First passage is persistent in the order-parameter equations, and the recrossings seen at finite d are fluctuation effects. At finite d the margin fluctuates: at $d = 500$ every entered run at $(\delta, \eta) = (0.3, 0.8)$ and $(0.3, 1.2)$ recrosses the cutoff at least once after first passage (median post-entry minimum margin 1.64 and 1.69; no run returns below 1.42). Finite-d entry times are therefore first-passage times of a fluctuating margin, while dormancy deepens on average as Proposition A.4 describes. A deterministic recrossing test integrates the order-parameter equations from the trained state at the thirty entering cells at step 0.5 and horizon 3000. The margin reaches $m_\star$ within the horizon in twenty-seven of them. Since the deterministic onset lies above the horizon at the three cells with $\eta = 0.7$, their measured entries are fluctuation-advanced. In none of those twenty-seven does the margin return below $m_\star$ after first passage.

Recovery after the drift stops. Section 3.2 records that a unit with zero activation cannot recover by gradient descent alone; here the teacher is frozen after 3000 units of drift instead of moving throughout, and the margins after the freeze are the reading. The persistence above holds only

under continued drift: once the teacher is frozen every entered unit returns below the cutoff. A recovery run drifts for 3000 units at $(\delta, \eta) \in \{(0.3, 0.8), (0.3, 1.2), (0.65, 0.8)\}$ and then freezes the teacher for 3000 units at the same learning rate. It uses $d = 500$ and 100 runs per cell (protocol in Appendix B.8). All 100 entered units at each cell return below the cutoff, and all end below it. The mean margin of the entered runs is 2.21, 2.64, and 2.20 when the drift stops, falls to 0.13, 1.01, and 0.09 after 200 units, and ends at 10^{-7} or less. The mean final state is $(Q, R, \theta) = (1.00, 1.00, 0.00)$, the teacher itself. The order-parameter equations do the same: from the trained state the margin at the end of the drift phase is 2.17, 2.60, and 2.18. After the drift is switched off it falls below $m_\star$ after 27, 84, and 28 units, and the state ends at $(0.993, 0.997, 0.001)$. The dormant state is therefore not a resting state of the model once the teacher stops moving: the exponentially small gradients of Proposition A.4 suffice for recovery once the drift no longer opposes them.

Persistence-based entry. Definition 1 reads entry as a first passage; here entry is read instead as the start of the first 50-unit interval above the cutoff, and Table 3 holds the timing comparison. Neither criterion of the rerun (protocol in Appendix B.8) was met. The entered-or-not classification of the 36 cells agrees between first passage and persistent entry in 33 cells but not in the three near-boundary cells at $\eta = 0.7$. There every run crosses $m_\star$ within the horizon, but only 10, 45, and 57 of 100 (at $\delta = 0.3$, 0.45, 0.65) hold it continuously for 50 units before the horizon ends. In the entering cells the median persistent-entry time exceeds the median first-passage time by a factor of 1.54 (per-cell range 1.09 to 1.86). Once across, the margin fluctuates around $m_\star$ before settling above it, and the factor of 1.54 measures that interval. This settling interval is what the finite-d recrossings recorded above register. The minimum margin after first passage is 1.35 over all 3000 entered runs. A hysteresis definition of entry with any release level at or below 1.35 classifies every cell exactly as first passage does.

Under the persistence definition the uncalibrated law is already close: Table 3 compares the persistent-entry time with the uncalibrated order-parameter entry time T over the 27 entering cells whose deterministic trajectory crosses the cutoff within the horizon. The uncalibrated median error is 5.1% at the 50-unit window, 12.7% at 25 and 14.5% at 100, against 56.9% for first passage over the same cells. The uncalibrated law is therefore already close under the persistent definition. First passage, the definition stated in advance, remains the primary comparison. The fitted factor depends on the entry definition (0.637 over these 27 cells for first passage, 0.987 for the 50-unit window).

The diffusion approximation also predicts the persistent-entry times (protocol in Appendix B.7). It is integrated without freezing after first passage and read with the same 50-unit window. It gives persistent-entry means with median relative error 2.0% (largest 5.7%) and Spearman 0.993 against the measured persistent entry over the 28 cells where at least half the runs enter persistently. No persistence fraction differs from the measured one by more than 0.1. At leading order the figures are 1.7% and 0.995, and its own persistent-over-first-passage factor is 1.536 against the measured 1.523 over the same 27 cells. At leading order one fraction gap exceeds 0.1, at $(0.1, 1.0)$.

Table 3: Persistent-entry time (first interval of ℓ units continuously at or above $m_\star$) against the uncalibrated order-parameter entry time over the 27 comparable cells, first passage for reference; error, median of $|T - T_{\text{meas}}|/T_{\text{meas}}$. Below the rule: the diffusion approximation against the same measurements with no fitted factor; means conditional on entry.

entry definition	Spearman	uncalibrated median error
first passage	0.964	56.9%
persistent, $\ell = 25$	0.991	12.7%
persistent, $\ell = 50$	0.991	5.1%
persistent, $\ell = 100$	0.987	14.5%
first passage, diffusion approximation (30 cells)	0.999	1.3%
persistent, $\ell = 50$, diffusion approximation (28 cells)	0.993	2.0%

D.3 Question 3: the learning rate and the network

Trainable output weight. Sections 4.3 and 5.4 state that a trained output weight removes dormancy; here the output weight is trained or held small instead of fixed at one, and the entry counts and final weights are the reading. The model fixes the output weight of the unit at one (Section 4.1); here that weight is trained instead (protocol in Appendix B.6: output $v\,\sigma$, v started at 1 and updated at the

same raw step, $d = 500$, 100 runs, horizon 3000). No run enters at $(0.3, 0.8)$, $(0.3, 1.2)$, $(0.1, 1.0)$, or $(0.65, 0.8)$, where the fixed-weight unit enters in 100 of 100. None enters at $(0.3, 0.5)$ either. The output weight falls to 0.14 to 0.16 and the threshold to between -2.5 and -2.8 at every cell, the same resolution as in the two-unit pair with trained output weights below. The output scale absorbs the mismatch and the unit stays active. The reading stated for this control (fewer than a tenth entering at every entering cell) applies: the isolated-unit mechanism is not observed without the fixed output scale at any tested cell.

Holding the output weight fixed at $v = 0.15$ instead of training it (protocol in Appendix B.6) gives the same outcome at the same cells: no entry in 100 runs with the threshold again between -2.6 and -2.7. The protection therefore lies in the output scale at which the weight settles; the adaptation of the weight itself is not needed. The reduction is exact for $v > 0$: with $\tilde{w} = vw$ and $\tilde{\theta} = v\theta$ the output is $g(\tilde{w} \cdot \xi - \tilde{\theta})$ and the updates are those of the original unit at rate $v^2\eta$. The fixed-v unit is therefore the original unit at rates 0.018 to 0.027 started from the scaled state $(Q, R, \theta) = (v^2, v, 0)$. Run that way at the same cells with 100 runs, the original unit enters in 0 of 100 at all four cells. Its final state equals the fixed-v final state scaled by v to four decimals with the same seeds. This is the small-rate finding of Section 4.5 from a shrunken aligned start. For a trained weight the transformed variables acquire the terms $w\,\Delta v$ and $\theta\,\Delta v$, which remain in the averaged dynamics.

The protection does not depend on the shrunken start. The unit is also started from the state that represents the teacher exactly, which under the reduction is the trained state $(1, 1, 0)$ at rate $v^2\eta$ (protocol in Appendix B.6). It again enters in 0 of 100 at all four cells and settles at the same fixed point.

The averaged dynamics with the output weight as a fourth order parameter reproduce the resolution (protocol in Appendix B.6; Table 7). From $(Q, R, \theta, v) = (1, 1, 0, 1)$ the margin never exceeds 0.07 at the four cells, and the final norm, overlap and output weight lie within 0.09 of the run means (v settles at 0.17 to 0.23 against 0.14 to 0.16). The threshold settles 0.3 to 0.9 less negative than in the runs, a discrepancy of the closure.

The unit stays active (activity 0.95 to 1.00) at alignment 0.13 to 0.41, of little more use than a constant. At the mean final states of the runs the closed forms of Lemma 2 give $\epsilon_g = 0.149$ at $(0.1, 1.0)$ and 0.163 to 0.169 at the $\delta = 0.3$ and 0.65 cells for the trained weight. With the weight held at 0.15 they give 0.150 and 0.162 to 0.169, against 0.170 for the constant predictor.

Small rates over a long horizon. Section 5.4 states that the order-parameter equations settle into an active state at small rates at every scanned speed; the horizon extends to 10^5 units at rates $\eta \leq 0.4$, and the final margins and Jacobian eigenvalues are the reading. At $\eta \in \{0.1, 0.2, 0.3, 0.4\}$ and $\delta \in \{0.1, 0.3, 0.65\}$ the order-parameter equations from the trained state converge within 10^5 units to a fixed point whose Jacobian eigenvalues are all negative and whose margin is negative, -0.07 to -1.63 (protocol in Appendix B.7). The late slope of the margin is 0 to the precision of the integrator; the margin along the way never exceeds 0.29. At finite d, entry at these rates is a fluctuation escape (Section 4.5).

Trajectories of the order-parameter equations. The last link of the mechanism in Section 4.3 says the threshold rises faster than the weight vector; the order-parameter equations are integrated at one drift speed for one tracking and two entering cells, and Figure 6 holds the trajectories. In all three cells the alignment falls first and the threshold rises. Figure 6: Q, $\rho = R/\sqrt{Q}$, θ, and $m = \theta/\sqrt{Q}$ along the order-parameter equations from the trained state at $\delta = 0.3$; tracking cell $\eta = 0.6$, entering cells 0.8 and 1.2. At 0.6 the norm settles below 1 and the margin at 0.43. At 0.8 and 1.2 the norm inflates to 1.44 and 1.72 by first passage and the threshold outruns it; the margin crosses 2 at 1831 and 416 units. The trajectories show the last three links of the mechanism of Section 4.3 in temporal order.

Norm-channel control. For the second link of the mechanism in Section 4.3 (the inflation of the weight vector), the squared increment is removed from the norm at every step, and the entry counts and terminal norms are the reading. After every weight update $w \leftarrow w + (\eta/d)\chi\xi$ the weight vector is rescaled so that $\|w\|^2$ equals $Q_{\text{old}} + 2(\eta/d)\chi\, x$, the value it would take without the squared increment $(\eta/d)^2\chi^2\|\xi\|^2$. The threshold update and the teacher move are unchanged. The rescaling multiplies the whole updated vector by $c = 1 - \eta^2\chi^2/(2Qd) + o(d^{-1})$, deleting the squared increment from the norm and shrinking the overlap $R = w \cdot B$ by the same factor. In the limit the intervention therefore obeys (3) without its $\eta^2\mathbb{E}[\chi^2]$ term and (4) unchanged. Its overlap equation

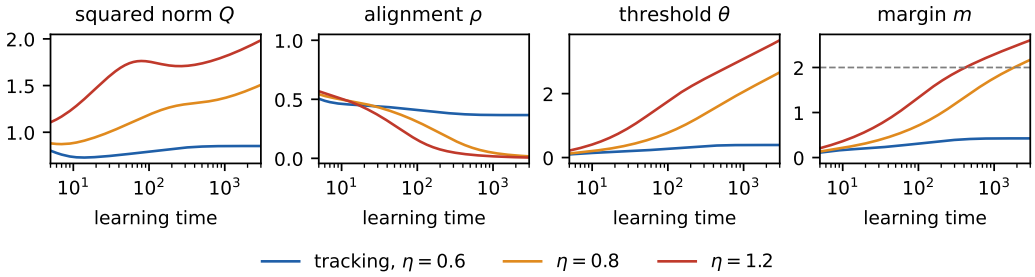


Figure 6: Order-parameter equations from the trained state at $\delta = 0.3$: norm, alignment, threshold and margin against learning time (log scale) at $\eta = 0.6$ (tracking), 0.8 and 1.2; dashed, the cutoff $m_\star = 2$.

is (2) with the extra term $-\eta^2 \mathbb{E}[\chi^2]\, R/(2Q)$, a state-dependent weight decay at rate $\eta^2 \mathbb{E}[\chi^2]/(2Q)$ that remains as $d \to \infty$.

No run enters at cells $(0.3, 0.8)$ and $(0.3, 1.2)$ ($d = 500$, 100 runs each, 3000-unit horizon; criterion at most 2 entries per cell). The order-parameter equations with the η^2 term deleted track at both cells. At $\delta = 0.3$ they have no onset up to $\eta = 2.2$ (the top of the scan) within a 1.2×10^4 horizon. Their terminal state there is a low-norm, negative-threshold unit ($Q \approx 0.38$, $\theta \approx -0.24$ at $\eta = 0.8$): the gracefully degraded state of Section 4.5.

Checkpoint means over the 100 runs fail their pointwise tolerance of 0.02 against (3) minus its η^2 term and (2) unchanged (deviations up to 0.27), as the extra overlap term requires. Against the corrected limiting system the same means agree within 0.007 at $(0.3, 0.8)$ and 0.011 at $(0.3, 1.2)$. The terminal norms are 0.25 and 0.22 against the simulated 0.24 and 0.21. The deviations are due to the missing overlap term and not to a finite-d effect, and the corrected system tracks at both cells with a negative margin throughout. The per-step rescaling is therefore a quantitative realization of its own limiting equations, which differ from the η^2-deleted equations by that overlap term.

A matched residual without drift produces no dormancy: this control (protocol in Appendix B.3) holds the teacher fixed and adds Gaussian label noise of standard deviation σ_{n} chosen so that $\mathbb{E}[\chi^2]$ matches the value of the drifting cell at the trained state. It runs at the twelve cells $\delta \in \{0.3, 0.65\} \times \eta \in \{0.7, \ldots, 1.2\}$ of the existing controls ($d = 500$, 100 runs, horizon 3000). The matched values $\mathbb{E}[\chi^2]$ are taken from the order-parameter equations (at the tracking state where one exists, otherwise the time average along the deterministic trajectory to first passage) and range from 0.032 to 0.055. The batches use $\sigma_{\mathrm{n}}\sqrt{2}$ (matched at the trained state, where the activity factor is 0.5) at all twelve cells and $2\sigma_{\mathrm{n}}$ at the six $\delta = 0.3$ cells. No run enters under the matched noise or under twice the matched noise (0 of 1200 and 0 of 600), while the drifting cells enter in 100 of 100 each. The mean margin under matched noise stays between 0.07 and 0.27, and under twice the matched noise between 0.21 and 0.49, far below $m_\star = 2$.

An input-independent target enters near the $2/\pi$ envelope, so dormancy does not require cumulative drift. Independent stationary targets (protocol in Appendix B.3): the target is a ReLU of a standard normal drawn independently of the input, so the residual cannot be learned. The unit enters in 0 of 100 runs at $\eta = 0.5$, 3 of 100 at 0.6, and 100 of 100 at 0.7, 0.8, and 1.2 (mean entry 1727, 951, and 289), a 50% crossing at 0.648 against the fully misaligned envelope $2/\pi = 0.637$ of Proposition A.8. The entered states have $R \approx 0$ and margins 2.0 to 2.6. That crossing lies at a rate where the drifting cells themselves enter dormancy. It arises when the target offers no input-dependent signal to align with (an unrelated target) or when that signal keeps moving away under drift.

Under additive noise the unit follows a stationary signal and does not go dormant even at the residual of the independent target. A fixed teacher with label noise $\sigma_{\mathrm{n}} = 0.584$ (the residual variance 0.341 of the independent target; protocol in Appendix B.3) gives 0 of 100 entries at every η from 0.5 to 1.2. The norm Q inflates from 1.5 to 3.9 and the threshold rises from 0.25 to 1.44. The overlap nevertheless grows with the norm (alignment 0.87 to 0.95) and the margin stays between 0.20 and 0.73. Residual magnitude alone therefore does not produce dormancy at these rates (up to 1.2; the plateau controls of Appendix D.1 enter from $\eta = 2.0$). With a learnable teacher the norm and the

overlap grow together and the margin stays low; with an unrelated target the overlap stays near zero and the threshold outruns the norm.

Task value of the degraded state. For the definition of tracking in Section 4.2, the closed-form generalization error at the tracking states is the reading, against the constant predictor. The degraded unit outperforms the constant predictor only at slow drift. Two reference errors bracket the degraded state. A unit with zero output has generalization error $\mathbb{E}[\tau^2]/2 = 1/4$ exactly on the unit-tracking task. This error is the reference for a dormant unit, which is still active on at most 2.3% of inputs. The best constant predictor outputs $\mathbb{E}[\tau] = 1/\sqrt{2\pi}$, and its error is $\frac{1}{2}\,\mathrm{Var}(\tau) = 1/4 - 1/(4\pi) \approx 0.170$ (derived here: the teacher is a ReLU of a standard Gaussian variable, with $\mathbb{E}[\tau] = 1/\sqrt{2\pi}$ and $\mathbb{E}[\tau^2] = 1/2$). The closed form of Lemma 2 at the degraded tracking state below the boundary gives $\epsilon_g = 0.152$ at $(\delta, \eta) = (0.3, 0.3)$, 0.177 at $(0.65, 0.3)$, 0.193 at $(1.5, 0.3)$, 0.200 at $(3.0, 0.3)$, and 0.2492 at $(1.5, 0.55)$ near the continuous crossover. Every value lies below the dormant error $1/4$. Only the slow-drift value 0.152 lies below the 0.170 of the constant predictor, so at the three faster drift speeds and near the crossover the degraded unit tracks worse than a constant output.

The network experiment. For the network statements of Section 5.4, the model changes from the isolated unit to a 128-unit network trained at batch one (protocol in Appendix B.5), and Figure 3b and Table 4 hold the readings. At fixed learning rate the dormant fraction grows monotonically with drift speed across this grid (0.34 to 0.55 at the largest rate).

Final tracking loss across the sixteen drift cells is set by the drift speed, not by the dormant fraction: the cell-mean loss correlates with δ at Spearman 0.97 but with dormancy only at 0.22 (permutation $p = 0.40$; seed means over the kept runs). Within each drift condition the loss is flat to within 10% while dormancy moves from zero to over half.

The difference between the largest and smallest learning rates exceeds the run-to-run variation at every drift speed. Ten independent runs per cell across the grid give standard deviations of the dormant fraction of at most 0.093 (Figure 3b, with the cell values in the supplement). The largest and smallest learning rates differ at paired $p = 0.002$ at each drift speed. The bootstrap 95% intervals for the top-rate drift-minus-control differences are $[0.29, 0.39]$, $[0.43, 0.54]$, $[0.49, 0.58]$, and $[0.50, 0.59]$ across the four drift speeds.

Along an equal-ratio diagonal (four grid cells with δ/η fixed while η spans a factor of eight) the measured dormancy is 0, 0, 0.027, 0.545. Dormancy therefore differs among cells with the same δ/η, though it is not strictly increasing along the diagonal (two cells are zero).

Four times the budget (protocol in Appendix B.5): over 1.2×10^6 steps the dormant fraction stays 0 in every run at $\eta_{\mathrm{raw}} = 0.00375$ at both drift speeds. At 0.0075 it reaches 0.025 at $\delta = 0.15$ (runs 0 to 0.125) and 0.098 at 0.6, with runs 0 to 0.375. It is still 0 at half the budget and rises in the last quarter. Since 9.8% exceeds 5%, the reading stated in advance makes the finite-learning-rate signature a budget statement: the two smallest rates are zero within the budget of the grid and the second smallest enters late.

The dormant units of the network do not follow the single-unit mechanism. Per-unit mechanism diagnostic: of 708 dormant units, 23 have their alignment fall to half its first-checkpoint value before the bias margin $\vartheta/\|w\|$ reaches 1, and 3 have a norm at entry at least 1.5 times its first-checkpoint value. The criterion stated in advance asked for at least 70% of dormant units to show both. The dormant units instead start with no teacher alignment, median cosine 0.35 and unchanged. Their biases rise in all 708, median 0.49 to 0.99; their norms shrink in 96%, median 0.66 to 0.38. The alignment statistic is the best cosine to any of the 32 teacher rows. Its chance level for a random direction in $d = 32$ is 0.36 (median 0.35, 5–95% range 0.24–0.50), so 0.35 is chance. Hence the single-unit erosion step has no aligned starting point to erode. Biases of surviving units rise similarly (0.44 to 0.97) while their norms grow (0.65 to 0.75). The drift-free control produces no dormant units. Figure 7 shows the per-unit trajectories behind these numbers at $(\delta, \eta_{\mathrm{raw}}) = (0.6, 0.03)$, fifteen sampled units per group with group means.

The dormant units of the network enter 4.5 times later than the single-unit law predicts. Median entry for the 708 dormant units is 3750 units of learning time (quartiles 2500 and 5312), against the single-unit prediction $T(0.6, 0.96) = 842$ from (6). A log point is 312.5 units at $d = 32$. The ratio of 4.5 is outside a factor of two, so the entry-time law does not transfer to this network either. No

unit is censored at the first log point. At input dimension 32 the single unit itself enters well before its $d \to \infty$ time (Appendix D.2), and the finite-d correction widens this gap rather than closing it.

Table 4: Dormant fraction at the largest raw step $\eta_{\text{raw}} = 0.03$ under the absolute criterion and the cited magnitude-based score, seed means over the kept runs (run sets in the text), with the layer-mean activity. Under the magnitude score at $\tau_{\text{S}} = 0.025$ all other rates are 0 to 0.007 and the controls 0 at every rate.

score	$\delta = 0.075$	0.15	0.3	0.6
absolute (activity ≤ 0.023)	0.355	0.484	0.510	0.553
layer-mean activity	0.107	0.078	0.061	0.050
magnitude-normalized, $\tau = 0.025$ (later rerun)	0.166	0.240	0.240	0.219
magnitude-normalized, $\tau = 0.1$ (later rerun)	0.368	0.451	0.458	0.411

The absolute and activity rows of Table 4 come from the per-unit diagnostic runs (8, 8, 9, and 10 of ten kept at the four drift speeds) and its magnitude rows from the magnitude rerun (10, 9, 9, 9 kept). Every dropped run diverged numerically, with non-finite weights from its first log point.

Under the cited score the finite-learning-rate signature (a rise of at least 5 points at every speed with controls at or below 5%) is met at all four speeds. However, the activity surrogate meets it at two. Magnitude-based score (protocol in Appendix B.5): a rerun of the largest-rate row with per-unit mean post-activation magnitudes logged gives dormant fractions 0.166, 0.240, 0.240, 0.219 at the four drift speeds and 0 in both controls under the cited score $s_i = \mathbb{E}|h_i| / \operatorname{mean}_j \mathbb{E}|h_j|$ at $\tau_{\text{S}} = 0.025$. At $\tau_{\text{S}} = 0.1$ the fractions are 0.37 to 0.46 and the jitter control (Appendix B.3) 0.023. The activity-based version on the same runs gives 0.020 to 0.057; it is computed from the per-unit logs of those runs, which ship with the supplement. Across the whole grid (protocol in Appendix B.5) the cited score at $\tau_{\text{S}} = 0.025$ is 0 at the two smallest rates at every drift speed, 0.001 to 0.007 at $\eta_{\text{raw}} = 0.015$, and 0.166, 0.240, 0.240, 0.219 at 0.03, with the same seeds as the magnitude rerun of the largest-rate row above. The drift-off and jitter controls are at 0 at every rate. The two scores differ by 15 to 20 points. The paper therefore reports the cited score and keeps the activity surrogate as a second reading. The normalized count falls at $\delta = 0.6$ while the absolute count rises because the normalized threshold is 0.025 of a layer-mean magnitude that falls from 0.053 to 0.024 across the four speeds. A fixed magnitude threshold gives 0.26, 0.38, 0.46, 0.46, monotone like the absolute row.

Input dimension 512: the same network with 512 inputs at the same model learning rates shows no dormant unit at any cell or control within 3×10^5 steps. At $d = 512$ this step budget is 586 time units. At the learning time of the $d = 32$ grid (9,375 units) the six rerun cells again show no dormant unit. Layer-mean activity is 0.35–0.50 against 0.05–0.14 at the same $d = 32$ cells. A width-matched rerun with 2048 hidden units (four per input as at $d = 32$) at the same six cells also shows no dormant unit: lowest unit activity 0.12, layer means 0.20–0.48. The absence is therefore not due to a shortage of units per input. Neither rerun shows the signature (it requires 5% at $\eta = 0.96$). The entry-time and mechanism comparisons have no dormant units to score.

A probe at $\delta = 0.6$ raises the raw steps to those of the $d = 32$ grid, $\eta_{\text{raw}} \in \{0.00375, 0.0075, 0.015\}$ ($\eta = 1.92$, 3.84, 7.68). It gives dormant fractions 0, 1.7%, and 16.5% (four of ten runs diverging at the last). The $d = 32$ network enters at the same raw-step scale (0–7% seed means at 0.015, 34–55% at 0.03, Figure 3b). The network onset tracks η_{raw} at the three raw steps probed, so the $\kappa = 1$ conversion of Appendix D.4 does not place the network on the single-unit diagram.

Only removing the shared readout nearly restores the predicted fate. Intermediate models between the unit and the network are a trained start, with the 32 student units initialized at the teacher; frozen output weights; and independent units, each trained on the ReLU output of its own teacher unit with no shared readout. They run at raw steps $\{0.015, 0.03\}$ and $\delta \in \{0.15, 0.6\}$ (protocol in Appendix B.5). The independent units restore that fate at three of the four cells (every unit dormant at $\eta_{\text{raw}} = 0.03$; none at $\delta = 0.15$ and 11% at 0.6 at the small rate, where the model predicts none). No change brings the entry times within a factor 2 of the single-unit prediction with a rising norm at the logging resolution (312 units). The per-cell dormant fractions, entry times and norm ratios of the three models are in the supplement.

The dormant units do not wake on a new fixed target. Recovery on a new fixed teacher (protocol in Appendix B.5): after the 3×10^5 drift steps the teacher is replaced by a fresh fixed random

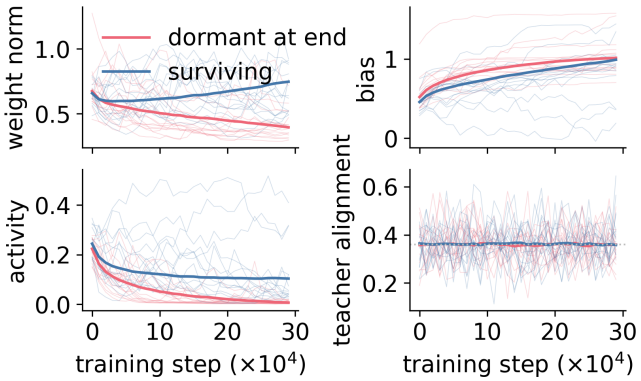


Figure 7: Per-unit network trajectories at $(\delta, \eta_{\text{raw}}) = (0.6, 0.03)$ (ten runs; 708 dormant of 1280 units): weight norm, bias, activity, teacher alignment (dotted: chance level 0.36 of the best-of-32 cosine); thick lines: group means; thin lines: fifteen sampled units per group; groups by final dormancy.

teacher and training continues for 10^5 steps at the same raw step. The recovery run of the unit in Appendix D.2 instead freezes the current teacher. The cells are $\eta_{\text{raw}} \in \{0.015, 0.03\}$ and $\delta \in \{0.15, 0.6\}$, with the drift-free network at each rate. The intended comparison was the steps to halve the post-switch loss between networks matched to within 25% in pre-switch loss. It cannot be applied: the drift-free networks end the first phase at a loss of 0.001 to 0.002 against 0.11 to 0.16 for the drifted ones, and the post-switch loss halves within the first 500 steps in 55 of 58 runs in every condition. At $\eta_{\text{raw}} = 0.03$ the drifted networks keep 42 to 43% of their units dormant through the whole second phase (seed medians, nine or ten runs). They reach a tenth of the switch loss after 3000 and 5000 steps against 1500 for the drift-free network, reach a loss of 0.02 after 8500 and 11500 steps against 4750, and end at 0.0043 and 0.0059 against 0.0032. At $\eta_{\text{raw}} = 0.015$, with at most 6% of units dormant, the drifted networks end at the same loss as the drift-free one. The network adapts through its remaining units.

Two competing units. For the pair statements of Sections 1 and 5.4, two units share one output error instead of one unit alone (protocol in Appendix B.6), and Tables 5 and 6 hold the readings. Below the single-unit fold exactly one of the two units enters in every run, which Appendix A.8 derives from the order-parameter equations of the pair. For reference, the single-unit boundary at $\delta = 0.3$ is $\eta_c = 0.629$, and at $(0.15, 0.7)$ the single-unit model produces 0 entries in 100 runs. The monotonicity criterion passed: the unit-level entry fraction at $\delta = 0.3$ is $1/2$, $1/2$, 1, 1 across the four learning rates, and the drift-free control produces no entries in 100 unit-runs. One clause of the signature criterion asked for at most 5% unit-level entry at the below-boundary cell $(0.3, 0.5)$. The measured value is 50%; the criterion failed. In every drift cell every run has at least one unit enter, and below the single-unit fold it is always exactly one. The losing unit enters quickly (mean 204 at $(0.15, 0.7)$ and 309 at $(0.3, 0.5)$) while its sibling never does; at $(0.15, 0.7)$ an isolated unit never enters. Just above the boundary at $\eta = 0.8$ one unit per run enters (mean 841) while the survivor outlasts the 3000-unit horizon. An isolated unit at that cell enters by 1126 on average. At the two largest learning rates both units enter in every run, each later than the isolated unit at the same cell (first and second unit means 1173 and 1485 versus 477 at $\eta = 1.0$; 495 and 618 versus 271 at 1.2).

At small rates entry through the competition channel is delayed rather than absent. The three further cells probe the learning-rate dependence of the channel: within the 3000-unit horizon no unit enters in 50 runs at $\eta = 0.05$, exactly one unit enters in 48 of 50 at 0.1, and in all 50 at 0.2. The isolated unit at the same drift speed enters in 0 of 100 runs over 12000 units at $\eta = 0.05$ and at 0.2, with final mean margins -1.8 and -0.7, and in 0 of 64 independent units at 0.1 (Table 5). A four-fold horizon (protocol in Appendix B.6) reverses the first reading: within 12000 units one unit enters in 49 of 50 runs at $\eta = 0.05$ and in 50 of 50 at 0.1, mean first entry 6927 and 2079. The mean first entry grows roughly as $\eta^{-1.7}$ between the two rates.

At $\eta = 0.1$ the channel survives a hundredfold attenuation of the term on which the isolated-unit mechanism rests. The batch-mean gradient over 100 inputs per step suppresses the η^2 term a hundredfold (protocol in Appendix B.6). Under it one unit still enters in 32 of 32 runs at $\eta = 0.1$ (mean first entry 6428 against 2079 at batch one, and 6423 at batch 500), while none enters in 32

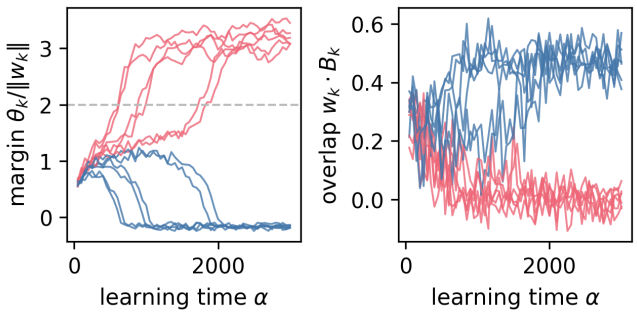


Figure 8: Two units at $(\delta, \eta) = (0.3, 0.8)$, $d = 500$, six runs: margin (left) and own-teacher overlap $w_k \cdot B_k$ (right) per unit; red: the entering unit; blue: its surviving sibling; dashed: $m_\star = 2$.

runs at 0.05 within 12000 units. The term speeds the channel up threefold. Attenuating it further to batch 500 leaves the entry time unchanged (6423 against 6428). The finite-learning-rate finding of Section 5.4 is thus confined to the isolated unit.

The learning rate of the network does not exclude the competition channel. At the $v = 1$ conversion the network rates ($\eta = 0.12$ to 0.96; Appendix D.4) lie above the entry rates of the channel. At the $v^2\eta$ conversion they lie below $\eta = 0.05$, where the channel takes 6927 ± 1742 units.

The surviving unit enters later than an isolated one at the same cell, or not at all within the horizon, so the single-unit dormancy boundary is relevant to the survivor rather than the loser. At $(0.3, 0.8)$ the margin of the entering unit rises through $m_\star = 2$ while its own-teacher alignment collapses, and the alignment of the surviving unit rises as its margin settles slightly negative (six paired trajectories, Figure 8; entry statistics, Table 6).

A matched bridge from independent to competing units isolates the ingredient (protocol in Appendix B.6). Everything else is as above: $d = 500$, orthonormal teacher pair, each unit started at its own teacher, output scale 1, horizon 12000, 32 runs. Two units that each match their own teacher unit reproduce the isolated-unit fates: both enter in 32 of 32 runs at $\eta = 0.8$ (unit entry mean 1093 ± 142 against 1126 for the isolated unit) and none of 64 units enters at 0.5 or 0.1. The shared residual opens the channel: exactly one unit per run enters at 0.8, 0.5, and 0.1 (first entry 810 ± 349, 297 ± 48, and 1965 ± 588). A trainable output weight closes it again. With the two output weights trained from 1 at the same raw step, no unit enters at any of the three rates within the horizon. The output weights fall to 0.13 to 0.19 and both thresholds to between -2.1 and -2.9. In every run at 0.1 and 0.8 both output weights end between 0.03 and 0.23 and both margins below -1.5. The pair therefore resolves the mismatch through its readout and both units stay active (Table 5). The layer studied above has trainable output weights, consistent with its dormant units arising by another mechanism.

Table 5: The matched bridge at $\delta = 0.3$ ($d = 500$, horizon 12000, 32 runs). Columns: units and runs entered (runs with at least one entering unit / with both), the entry time of the first entering unit as mean $\pm$ s.d., and the final mean thresholds $\theta_{1,2}$ and output weights $v_{1,2}$ of the two units; "trained v" trains the output weights.

variant	η	units	runs	first entry	θ_1, θ_2	v_1, v_2
independent	0.1	0/64	0/32	none	$-0.37, -0.37$	1, 1
independent	0.5	0/64	0/32	none	0.07, 0.07	1, 1
independent	0.8	64/64	32 / 32	1093 ± 142 (per unit)	3.46, 3.46	1, 1
shared	0.1	32/64	32 / 0	1965 ± 588	$-0.02, -0.19$	1, 1
shared	0.5	32/64	32 / 0	297 ± 48	0.86, 0.40	1, 1
shared	0.8	32/64	32 / 0	810 ± 349	2.00, 1.29	1, 1
shared, trained v	0.1	0/64	0/32	none	$-2.13, -2.13$	0.18, 0.19
shared, trained v	0.5	0/64	0/32	none	$-2.82, -2.91$	0.14, 0.14
shared, trained v	0.8	0/64	0/32	none	$-2.62, -2.89$	0.13, 0.15

The outputs above are unnormalized sums. Scaling student and teacher outputs by $1/K$ multiplies the squared loss and its gradient by $1/K^2$, so it divides the effective learning rate by $K^2 = 4$. Repeating the two-unit runs with outputs scaled by $1/K$ is consistent with this identity.

Table 6: Two-unit entry times by cell, mean $\pm$ s.d. over entering runs ($d = 500$, 50 runs per cell); "none": no entry within the 3000-unit horizon; the last rows give the four-fold horizon and the batch runs of the two small-rate cells. Last column: the single-unit mean at the same cell; sources for the three small-rate cells are in the text.

(δ, η)	first unit	second unit	isolated unit
$(0.3, 0.05)$	none	none	none
$(0.3, 0.1)$	1932 ± 442	none	none
$(0.3, 0.2)$	739 ± 164	none	none
$(0.15, 0.7)$	204 ± 43	none	none
$(0.3, 0.5)$	309 ± 74	none	none
$(0.3, 0.8)$	841 ± 405	none	1126
$(0.3, 1.0)$	1173 ± 183	1485 ± 188	477
$(0.3, 1.2)$	495 ± 86	618 ± 80	271
$(0, 1.2)$	none	none	none
$(0.3, 0.05)$, horizon 12000	6927 ± 1742 $(49/50)$	none	none
$(0.3, 0.1)$, horizon 12000	2079 ± 491	none	none
$(0.3, 0.05)$, batch 100, 12000	none (32 runs)	none	none
$(0.3, 0.1)$, batch 100, 12000	6428 ± 1335	none	none
$(0.3, 0.1)$, batch 500, 12000	6423 ± 1473	none	none

D.4 Protocol audits

The learning rate of the threshold. For the threshold-rate statement of Section 4.1, the threshold is stepped at κ times the weight rate. We integrate the order-parameter equations numerically from the trained state and compare the onset within $\alpha \leq 1.2 \times 10^4$ at $\kappa = 1$ against $\kappa \in \{0.5, 2\}$. At $\delta = 0.3$ the onset is 0.668, 0.638, and 0.630 for $\kappa = 0.5$, 1, 2 against the fold at 0.6298; at $\delta = 0.45$ it is 0.650, 0.614, 0.595 against 0.581; on the crossover side at $\delta = 1.2$ it is 0.634, 0.597, 0.578. The differences from $\kappa = 1$ exceed 0.005 at every speed. The shift is a horizon effect on the entry time rather than a moved fold. Near the fold the passage time grows as the rate of the threshold shrinks, and a finite horizon censors it. This censoring is why even $\kappa = 1$ lies above the fold by 0.008 and 0.033 at the two speeds. Repeating the $\kappa = 0.5$ bisection at $\delta = 0.3$ with a four-fold horizon ($\alpha \leq 4.8 \times 10^4$) moves its onset from 0.668 to 0.6298, onto the κ-independent fold. Direct simulation at $d = 500$ with $\kappa \in \{0.5, 2\}$ gives the same fate as at $\kappa = 1$ at both cells chosen in advance: 100 of 100 runs enter at $(0.3, 0.8)$ and 0 of 100 at $(0.3, 0.5)$ for both values. The mean entry times are 2060 ± 248 at $\kappa = 0.5$ and 567 ± 108 at $\kappa = 2$ against 1126 at $\kappa = 1$, and the order-parameter entry times 3603 and 943 against 1830.

We run two further conventions at the same two cells. A threshold stepped at $\eta/\sqrt{d}$ has per-step fluctuations of variance $O(1/d)$ that accumulate to $O(1)$ over one unit of learning time, so it is no longer a deterministic order parameter. At $(0.3, 0.8)$ every run enters, at mean time 56 ± 14, and at $(0.3, 0.5)$ none does. A threshold stepped at raw η puts every run into dormancy at both cells within 3 to 5 time units, a threshold runaway driven by the step noise itself rather than by drift.

Under the learning-time and initialization scalings used here, a layer with input dimension d trained at raw step η_{raw} on weights and bias alike is the $\kappa = 1$ case with $\eta = \eta_{\text{raw}} d$. The network experiment of Section 5.4 ($d = 32$; raw steps 0.00375 to 0.03) converts to η from 0.12 to 0.96 at $v = 1$. At the network output-weight scale ($|v| \approx 0.1$) the effective rates are 10^{-3} to 10^{-2}, below the region where the unit enters. The diagram therefore predicts no entry at any grid cell. Minibatching enters only through the η^2 term, as the minibatch protocol below sets out. The phase diagram of this paper is therefore a statement about a threshold that is slow relative to the weights.

Label noise, weight decay, and minibatch scaling. Section 5.4 states that the fates of the unit hold under label noise and minibatch rescaling and are abolished by weight decay; label noise, a decay term, or a batch size is added to the main-grid protocol at two to four cells, and the entry counts are the reading. Label noise leaves the fates unchanged, weight decay suppresses entry, and minibatches rescale the diagram as Lemma A.16 predicts; the exception is the network at batch 32, which enters below the predicted threshold. Label-noise sensitivity: adding Gaussian noise of standard deviation $\sigma_y \in \{0.1, 0.3\}$ to the teacher output leaves the fates at both test cells unchanged (100 of 100 entering at $(0.3, 0.8)$ and 0 of 100 at $(0.3, 0.5)$). Entry accelerates with noise (mean times 1026 and 709 against 1126 clean), consistent with noise enlarging $\mathbb{E}[\chi^2]$ and hence the η^2 term. At these amplitudes the fold does not cross $\eta = 0.5$.

Weight-decay sensitivity: the decay suppresses entry far more strongly than the modest fold shift stated in advance. The test covers four learning rates $\eta \in \{0.5, 0.6, 0.8, 1.2\}$ at $\delta = 0.3$ and two coefficients, eight measured fractions in all. At $\lambda_{\mathrm{wd}} = 0.01$ the cell $(0.3, 0.8)$ enters in 0 of 100 runs against 100 of 100 without decay; $(0.3, 1.2)$ enters in 34 of 100 with mean entry time 1635. At $\lambda_{\mathrm{wd}} = 0.1$ all four rates enter in 0 of 100. The two rates below the fold, $\eta \in \{0.5, 0.6\}$, stay at 0 of 100 with and without decay.

The decayed order-parameter equations integrated from the trained state over the same horizon predict no entry at any of the eight cells. Decay inside the order-parameter equations (Lemma A.17): seven of the eight measured fractions are 0 of 100, and the one at $(0.3, 1.2)$ with $\lambda_{\mathrm{wd}} = 0.01$ is 34 of 100. The decayed order-parameter equations place the decayed onset at $\delta = 0.3$ beyond $\eta = 1.6$ for both coefficients against 0.731 without decay, and make $(0.3, 0.8)$ non-entering already at $\lambda_{\mathrm{wd}} = 6.6 \times 10^{-4}$.

The diffusion approximation with the decay terms of Lemma A.17 (protocol in Appendix B.6) gives entry fraction 0 at the five cells where none of 100 runs enters. At $(0.3, 1.2)$ with $\lambda_{\mathrm{wd}} = 0.01$ it gives 0.51 against the measured 0.34, outside the 0.15 stated for this check. Its persistent-entry fraction there is 0.04, so under decay the entries it predicts are transient recrossings.

The mechanism is that the $-2\gamma_{\mathrm{wd}} Q$ term of Lemma A.17, with $\gamma_{\mathrm{wd}} = \eta\lambda_{\mathrm{wd}}$, cancels the η^2 norm increment. After that the norm never inflates and the threshold never grows without bound. The 34 of 100 entries at $(0.3, 1.2)$, whose deterministic margin ends at 1.38, are consistent in kind with the finite-d fluctuation channel of Section 4.5.

Minibatch scaling: by Lemma A.16 the boundary and every landmark scale as $(\delta, \eta) \mapsto (b\delta, b\eta)$, so $\eta_{\mathrm{c}}^{(2)}(0.3) \approx 1.60$ and $\eta_{\mathrm{c}}^{(4)}(0.3) \approx 4.6$. The simulation agrees in all four cells: at $b = 2$, 0 of 100 runs enter at $\eta = 0.8$ and 100 of 100 at $\eta = 2.0$; at $b = 4$, 0 of 100 at $\eta = 3.2$ and 100 of 100 at $\eta = 5.6$.

Combined with the layer conversion, the rescaling predicts no drift-driven dormancy for the network at $b = 32$. With the layer conversion $\eta = \eta_{\mathrm{raw}} d$ of the threshold-rate audit above, a layer of input width d trained with batch b would under the isolated-unit model enter dormancy only if η_{raw} exceeds about $0.55\, b/d$. The lower envelope of the deterministic onset over all drift speeds is its fast-drift limit 0.554 (Section 4.4). It lies just below the minimum of the fold $\eta_{\mathrm{cp}} = 0.5623$ and is not the fold itself. For the network of Section 5.4 ($d = 32$) at $b = 32$ this is $\eta_{\mathrm{raw}} \geq 0.55$, some twenty times the divergence ceiling of its batch-one scan (Appendix B.6).

Population-gradient reference (protocol in Appendix B.8): no run enters at any of the twelve cells with the batch-mean gradient over $b \in \{32, 500\}$ inputs per step under the same drift, at $\delta \in \{0.3, 0.65\}$ and $\eta \in \{0.7, 1.2, 2.0\}$ with 32 runs per cell over 1.2×10^4 units. At every cell the threshold ends below zero. The final margins are -0.43, -0.28, -0.18 at $\delta = 0.3$ and -0.73, -0.51, -0.35 at 0.65 for $b = 500$; these lie within 0.002 of the order-parameter equations without the η^2 term at the same time, and within 0.03 of them for $b = 32$. The final norms are $Q = 0.20$ to 0.62, below the trained 1, with overlaps $R = 0.29$ to 0.69. So the averaged rule has an active low-norm tracking state as its limit. The scaled fold of Lemma A.16 (near 60 for $b = 32$ and 10^3 for $b = 500$) lies far above the scanned rates, so the lemma predicts no entry there, as observed.

The direct test at $b = 32$ fails the prediction at the fastest drift. Its criteria were no cell with $\eta_{\mathrm{raw}} \leq 0.5$ above 5% dormancy and a drift-free control below 5% at every rate (protocol in Appendix B.5). At $\delta = 2.4$ and $\eta_{\mathrm{raw}} = 0.12$ the dormant fraction is 15–16% in two independent copies (14.9% and 16.3%, up to 55% in single runs). The drift-free control at the same cell is at 0, and the jitter control at 0.08% (one unit in one of ten runs). Every other non-diverging drift cell stays at or below 0.7%, and the controls stay at 0. Both $\eta_{\mathrm{raw}} \in \{0.25, 0.5\}$ diverge in all runs with or without drift, so the batch-32 stability ceiling lies between 0.12 and 0.25 and the predicted threshold 0.55 is out of reach. The two copies agree on all fourteen cells both completed. The network thus shows drift-driven dormancy at batch 32 well below the single-unit placement threshold: the placement, like the mechanism, does not transfer from the unit to the network.

D.5 Criteria stated in advance: failures and qualified outcomes

Table 7 lists each criterion whose outcome was a failure or a qualified reading, and where its full result is reported. Of the 93 criteria, 46 were met, 26 not met, and 21 had a qualified outcome (as

read, mixed, reported, or not assessable). The full register, met criteria included, ships with the supplementary code as `criteria_register.md`.

Table 7: Criteria whose outcome was a failure or a qualified reading, with a status and where the full result is reported. Status: not met; as read (one of two pre-stated readings applied); reported (no pass or fail was set); mixed; not assessable. $^+$ marks criteria added after the original set, all but one entered before their data were collected.

criterion	outcome	status	reported in
Solvable model: timing, boundary, basin, channels			
signature disjunction, finite-size decomposition alternative	not met in the form stated	not met	App. B.1
escape-scaling slope at $(0.3, 0.60)$	out of range	not assessable	App. D.2
basin sweep: near-trained zero entry below the fold$^+$	not met in $3/13$ cells	not met	App. D.1
matched stationary controls at the five plateau rates: their 50% crossings reported next to the drift crossings, no pass or fail$^+$	reported (1.95 noisy, 2.05 noiseless, against 1.95 and 1.85)	reported	App. D.1
onset scan across the end point: interior minimum of the 50%-crossing learning rate in δ^+	not met (monotone)	not met	App. D.1
the same scan at a four-fold horizon$^+$	not met (monotone)	not met	App. D.1
the end-point pair at $d = 2000^+$	not met (monotone)	not met	App. D.1
the end-point pair at $d = 1000$ and 4000, dimension extrapolation$^+$	not met (monotone; deterministic onset under the protocol also monotone)	not met	App. D.1
action robustness at doubled path resolution $(5\%)^+$	not met	not met	App. D.2
four-variable closure with a trained output weight against the trained-weight runs (run before a reading rule was written)$^+$	reported (no entry; threshold 0.3 to 0.9 less negative)	reported	App. D.3
Solvable model, controls, conventions, horizon, calibration and persistence			
norm-channel control: modified trajectories within 0.02 of the modified order-parameter equations$^+$	not met (0.27)	not met	App. D.3
label noise $\sigma_y \in \{0.1, 0.3\}$: fates unchanged at two cells$^+$	as read $(4/4)$	as read	App. D.4
weight decay $\lambda_{\mathrm{wd}} \in \{0.01, 0.1\}$: modest upward shift of the fold at four cells$^+$	suppression stronger than expected ($0/100$ at $(0.3, 0.8)$ for 0.01)	not met	App. D.4
calibration factor transfers across drift speeds: leave-one-speed-out median error $\leq 15\%$ at every speed$^+$	not met (26.6% and 16.8% at $\delta = 0.1, 0.15$; $\leq 10.2\%$ from 0.2 up)	not met	supplement
persistence: cell classification identical under 50-unit continuous dormancy$^+$	not met ($33/36$; hysteresis definition $36/36$)	not met	App. D.2
persistence: median persistent-entry over first-passage time $\leq 1.10^+$	not met (1.54)	not met	App. D.2
persistent-entry timing: uncalibrated error $\leq 15\%$ and Spearman ≥ 0.95 at windows 25 and 100^+	met under the stated denominator (12.7%, 14.5%; Spearman 0.991, 0.987), whereas the first reading, with the prediction as denominator, gave 16.4% at 100	mixed	App. D.2
boundary crossings between fold and same-protocol deterministic onset at three speeds$^+$	as read $(3/3)$	as read	App. D.1

continues on the next page

Table 7, continued.

criterion	outcome	status	reported in
threshold rate κ: horizon-limited onset within 0.005 of $\kappa = 1^+$	not met in the form stated (horizon effect)	not met	App. D.4
Beyond the solvable model			
network equal-ratio diagonal	not met (floor at zero)	not met	App. D.3
two-unit signature ($\leq 5\%$ below boundary)	not met (50%)	not met	App. D.3
two units at $\eta \in \{0.05, 0.1\}$ over a four-fold horizon: entries at 0.05 change the reading to delayed entry$^+$	entries appeared (49/50 within 12000 units)	as read	App. D.3
two units at batch 100 and $\eta \in \{0.05, 0.1\}$: no entry at either rate extends the finite-rate finding to the channel; entry in half the runs at either rate confines it to the isolated unit$^+$	32/32 at 0.1, 0/32 at 0.05: confined	as read	App. D.3
two units at batch 500 and $\eta = 0.1$: entry in half the runs keeps the statement that the channel does not need the η^2 term$^+$	kept (32/32, mean 6423)	as read	App. D.3
network learning-rate scan: interior minimum of $\eta_c^{\mathrm{net}}(\delta)^+$	not seen in range	not met	App. B.6
network per-unit mechanism sequence in $\geq 70\%$ of dormant units$^+$	not met (23/708, 3/708)	not met	App. D.3
network per-unit entry time within $[0.5, 2]\times$ the single-unit prediction$^+$	not met (ratio 4.5)	not met	App. D.3
network layer-normalized dormant fraction ($\tau = 0.025$) within 10 points of the absolute one$^+$	not met (6.6% vs 55.3%)	not met	App. D.3
network normalized-score signature across the grid ($\tau = 0.025$): rise ≥ 5 points at every speed, controls $\leq 5\%^+$	not met at two of four speeds (3.1, 4.1, 6.2, 6.6%; controls 0) under the activity surrogate; met at all four under the cited magnitude score (16.6, 24.0, 24.0, 21.9%; controls 0)	mixed	App. D.3
network at batch 32: no drift-driven dormancy ($\leq 5\%$) at any non-diverging rate$^+$	not met (15–16% at $\delta = 2.4$, $\eta_{\mathrm{raw}} = 0.12$)	not met	App. D.4
network at four times the budget: dormant fraction $\leq 1\%$ at the two smallest rates keeps the signature; any cell above 5% makes it a budget statement$^+$	0 at 0.00375; 2.5% and 9.8% at 0.0075	as read	App. D.3
network with 512 inputs at the same model rates: signature (zero at the two smallest rates, $\geq 5\%$ at the largest; controls $\leq 1\%$) within 586 time units$^+$	not met (no dormant unit anywhere)	not met	App. D.3
network with 512 inputs, matched learning time 9,375: $\geq 5\%$ at $\eta = 0.96$, $\leq 1\%$ at 0.48 and in controls$^+$	not met (no dormant unit anywhere)	not met	App. D.3
network with 512 inputs, entry-time ratio in $[0.5, 2]$ and mechanism sequence in $\geq 70\%$ at $(0.6, 0.96)^+$	not assessable (no dormant unit)	not assessable	App. D.3
network with 512 inputs and 2048 units (four per input, as at $d = 32$), matched learning time: $\geq 5\%$ at $\eta = 0.96$, $\leq 1\%$ at 0.48 and in controls$^+$	not met (no dormant unit anywhere)	not met	App. D.3
network with 512 inputs and 2048 units, entry-time ratio in $[0.5, 2]$ and mechanism sequence in $\geq 70\%$ at $(0.6, 0.96)^+$	not assessable (no dormant unit)	not assessable	App. D.3
Solvable model and network, later additions			

continues on the next page

Table 7, continued.

criterion	outcome	status	reported in
independent stationary targets: entry in half the runs at a rate where the drifting cells enter withdraws the cumulative-drift wording; crossing within 0.05 of $2/\pi^{+}$	entered ($100/100$ from $\eta = 0.7$; crossing 0.648)	as read	App. D.3
noisy fixed teacher at the independent target's residual: entry $\leq 2\%$ at every rate means input–target dependence matters beyond residual magnitude^{+}	$0/500$	as read	App. D.3
single-unit recovery after drift stops: $< 10\%$ of entered units return (persistent) or $> 50\%$ return (dormant while the drift lasts)$^{+}$	second reading ($300/300$ return)	as read	App. D.2
network recovery: halving time of the post-switch loss longer for dormant-heavy networks at matched pre-switch loss^{+}	not assessable (pre-switch losses unmatched; halving within the first 500 steps in $55/58$ runs)	not assessable	App. D.3
intermediate network models: one change brings entry within a factor 2 of the single-unit prediction with a rising norm^{+}	not met (independent units nearly restore the fate, entry $\geq 2.7\times$ faster; the other two move away)	not met	App. D.3
slow-drift cells across dimensions: at $(0.1, 1.0)$ diffusion entry fraction ≥ 0.95 and median lag at most half its $d = 500$ value by $d = 1000^{+}$	lag met (8% against 20%); fraction not met (0.80, leading-order scheme)	mixed	App. D.2
overlap-drift perturbation of 0.003 and 0.006 in the diffusion at $(0.1, 1.0)^{+}$	entry fraction fell to 0.003 and 0 (the cell is marginal)	as read	supplement
$O(1/d)$ norm-drift term: below 10^{-3} at the four states and entry fraction at $(0.1, 1.0)$ within 0.05 of 0.78 means negligible; a move above 0.05 means it matters^{+}	matters (2.1 to 9.4×10^{-4}; $0.78 \to 0.995$ with the state-dependent term, 0.996 with a constant shift)	as read	supplement
output weight held at $v = 0.15$: no entry means the protection is the output scale, entry means the adaptation itself^{+}	no entry (0 of 100 at the four entering cells)	as read	App. D.3
diffusion approximation under weight decay: entry fraction within 0.15 of the measured one at the entering cell^{+}	not met (0.51 against 0.34; the five non-entering cells at 0)	not met	App. D.4
magnitude-based normalized score within 5 points of the activity surrogate at every speed^{+}	not met (15 to 20 points apart; the cited score is reported)	not met	App. D.3