

A Supplementary Material

All code for running simulations, numerically solving the ODEs and generating phase plots is provided in the attached ‘RL-Perceptron.rar’ zip file. All code for running bossfight and pong is provided in the attached ‘procgen_fork.rar’ zip file/

B Relation to vanilla policy gradient

The RL perceptron with its update rule eq. (1) can be grounded in *policy gradient methods* [Sutton et al., 2000] where, at every timestep t , the agent occupies some state s_t in the environment, and receives an observation \mathbf{x}_t conditioned on s_t . An action y_t is then taken by sampling from the policy $\pi(y_t | \mathbf{x}_t)$, and the agent receives a reward accordingly. Policy gradient methods aim to optimise parameterised policies with respect to the total expected reward J . The gradient step for the REINFORCE policy gradient method is given in eq. (9). Our sequential decision-making task can be reformulated in the language of policy gradient methods: at each timestep t , the state s_t of the environment can be one of two states s_+, s_- and $\mathbf{x}_t \sim P(\cdot | s_t)$ is a high dimensional sample representative of the underlying state, with $P(\cdot | s_{\pm}) = \mathcal{N}_{\pm}(\cdot | \mathbf{w}^*)$. Where $\mathcal{N}_{+}(\cdot | \mathbf{w}^*)$ is the $N(\mathbf{0}, \mathbb{I}_D)$ distribution, but with zero-probability mass everywhere except in the half-space whose normal is parallel to \mathbf{w}^* , and $\mathcal{N}_{-}(\cdot | \mathbf{w}^*)$ is correspondingly non-zero in the half-space with a normal that is antiparallel to \mathbf{w}^* — ($N(\mathbf{0}, \mathbb{I}_D)$ has been partitioned in two). The next state s_{t+1} is sampled with probability $P(s_{t+1} | s_t) \equiv P(s_{t+1}) = 1/2$ independently from the decision made by the student at previous steps. At the end of an episode, after all decisions have been made, we update the agent as in eq. (1). Within this framework we can consider both rewards and penalties, i.e. at the end of an episode we may consider a reward (penalty) of size η_1 (η_2) depending on the fulfilment (unfulfilment) of Φ . The introduction of states in this case does not affect the dynamics of the system, but we use them as an exemplary case for more complex setups we plan to instantiate with more states and where actions conditionally affect state transitions without required a partitioning of the Gaussian distribution. Formally, the mapping to the RL setting can be stated by introducing the states and a probabilistic policy $\pi_{\mathbf{w}}(y | \mathbf{x}) = 1/(1 + \exp\{-y\mathbf{w}^T \mathbf{x}\}/\sqrt{D})$. The REINFORCE policy gradient update in this case is

$$\nabla_w J = \left\langle \sum_{t=0}^{T-1} \nabla_{\mathbf{w}} \log \pi_{\mathbf{w}}(y_t | \mathbf{x}_t) \left(\sum_{t'=t+1}^T r_{t'} \right) \right\rangle \approx \left\langle \sum_{t=0}^{T-1} y_t \mathbf{x}_t [\eta_1 \mathbb{I}(\Phi) - \eta_2 (1 - \mathbb{I}(\Phi))] \right\rangle \quad (9)$$

$$\longrightarrow \Delta \mathbf{w} \propto \eta_1 \left\langle \sum_{t=0}^{T-1} y_t \mathbf{x}_t \mathbb{I}(\Phi) \right\rangle - \eta_2 \left\langle \sum_{t=0}^{T-1} y_t \mathbf{x}_t (1 - \mathbb{I}(\Phi)) \right\rangle \quad (10)$$

The approximation in eq. (9) holds in the early phases of learning—when $\mathbf{w}^T \mathbf{x}/\sqrt{D}$ is small

$$\nabla_w \log \pi(y|x) = \nabla_w \log \frac{1}{1 + e^{-y\mathbf{w} \cdot \mathbf{x}}} \quad (11)$$

$$\approx -\nabla_w \log e^{-y\mathbf{w} \cdot \mathbf{x}} = y\mathbf{x} \quad (12)$$

-and gives us the possibility to understand the most complex part of the problem when the student is still learning the rule. In this way, the update in eq. (1) is analogous to the REINFORCE policy gradient in the same way that the perceptron update is analogous to SGD on the squared loss for a perceptron in binary classification.

C Derivations

Thermodynamic Limit: In going from the stochastic evolution of the state vector \mathbf{w} to the deterministic dynamics of the order parameters, we must take the thermodynamic limit. For the ODE involving R we must take the inner product of eq. (1) with \mathbf{w}^* .

$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} + \frac{\eta_1}{\sqrt{D}} \left(\frac{1}{T} \sum_{t=1}^T y_t \mathbf{x}_t \mathbb{I}(\Phi) \right)^{\mu} - \frac{\eta_2}{\sqrt{D}} \left(\frac{1}{T} \sum_{t=1}^T y_t \mathbf{x}_t (1 - \mathbb{I}(\Phi)) \right)^{\mu} \quad (13)$$

$$DR^{\mu+1} = DR^{\mu} + \frac{\eta_1}{\sqrt{D}} \left(\frac{1}{T} \sum_{t=1}^T y_t \mathbf{w}^{*\top} \mathbf{x}_t \mathbb{I}(\Phi) \right)^{\mu} - \frac{\eta_2}{\sqrt{D}} \left(\frac{1}{T} \sum_{t=1}^T y_t \mathbf{w}^{*\top} \mathbf{x}_t (1 - \mathbb{I}(\Phi)) \right)^{\mu} \quad (14)$$

557 we subtract DR^{μ} and sum over l episodes, the LHS is a telescopic sum, and 14 becomes

$$\begin{aligned} \frac{D(R^{\mu+l} - R^{\mu})}{l} &= \frac{\eta_1}{\sqrt{D}} \frac{1}{l} \sum_{i=0}^{l-1} \left(\frac{1}{T} \sum_{t=1}^T y_t \mathbf{w}^{*\top} \mathbf{x}_t \mathbb{I}(\Phi) \right)^{\mu+i} \\ &\quad - \frac{\eta_2}{\sqrt{D}} \frac{1}{l} \sum_{i=0}^{l-1} \left(\frac{1}{T} \sum_{t=1}^T y_t \mathbf{w}^{*\top} \mathbf{x}_t (1 - \mathbb{I}(\Phi)) \right)^{\mu+i} \end{aligned} \quad (15)$$

$$\frac{dR}{d\alpha} = \frac{\eta_1 + \eta_2}{\sqrt{D}} \left\langle \frac{1}{T} \sum_{t=1}^T y_t \mathbf{w}^{*\top} \mathbf{x}_t \mathbb{I}(\Phi) \right\rangle - \frac{\eta_2}{\sqrt{D}} \left\langle \frac{1}{T} \sum_{t=1}^T y_t \mathbf{w}^{*\top} \mathbf{x}_t \right\rangle \quad (16)$$

558 We go from eq. 15 to eq. 16 by taking the limit $D \rightarrow \infty$, $l \rightarrow \infty$ and $l/D = d\alpha \rightarrow 0$. The
 559 RHS of eq. 16 is a sum of a large number of random variables, and by the central limit theorem
 560 is self-averaging in the thermodynamic limit (under the assumption of weak correlations between
 561 episodes), consequently the LHS is self-averaging. A similar procedure can be followed for order
 562 parameter Q , but we instead take the square of eq. ?? and go to the limit described, obtaining:

$$\begin{aligned} \frac{dQ}{d\alpha} &= \frac{2(\eta_1 + \eta_2)}{T\sqrt{D}} \left\langle \sum_{t=1}^T y_t \mathbf{w}^{\top} \mathbf{x}_t \mathbb{I}(\Phi) \right\rangle - \frac{2\eta_2}{T\sqrt{D}} \left\langle \sum_{t=1}^T y_t \mathbf{w}^{\top} \mathbf{x}_t \right\rangle \\ &\quad + \frac{\eta_1^2 - \eta_2^2}{T^2 D} \left\langle \sum_{t,t'=1}^T y_t y_{t'} \mathbf{x}_t^{\top} \mathbf{x}_{t'} \mathbb{I}(\Phi) \right\rangle + \frac{\eta_2^2}{T^2 D} \left\langle \sum_{t,t'=1}^T y_t y_{t'} \mathbf{x}_t^{\top} \mathbf{x}_{t'} \right\rangle \end{aligned} \quad (17)$$

563 recalling the auxiliary variables (known in the literature as the aligning fields)

$$\nu = \frac{\mathbf{w}^{*\top} \mathbf{x}}{\sqrt{D}} \quad \text{and} \quad \lambda = \frac{\mathbf{w}^{\top} \mathbf{x}}{\sqrt{D}}, \quad (18)$$

564 which are sums of N independent terms and by the central limit theorem they obey a Gaussian
 565 distribution, we note

$$\langle \nu \rangle = \langle \lambda \rangle = 0 \quad (19)$$

$$\langle \nu^2 \rangle = D, \quad \langle \lambda^2 \rangle = DQ \quad (20)$$

$$\langle \nu \lambda \rangle = \mathbf{w}^{*\top} \mathbf{w} = DR \quad (21)$$

566 Substituting 18 into 16 and 17 we can rewrite as

$$\frac{dR}{d\alpha} = \frac{\eta_1 + \eta_2}{T} \left\langle \sum_{t=1}^T \nu_t \text{sgn}(\lambda_t) \mathbb{I}(\Phi) \right\rangle - \eta_2 \langle \nu \text{sgn}(\lambda) \rangle \quad (22)$$

$$\begin{aligned} \frac{dQ}{d\alpha} &= \frac{2(\eta_1 + \eta_2)}{T} \left\langle \sum_{t=1}^T \lambda_t \text{sgn}(\lambda_t) \mathbb{I}(\Phi) \right\rangle - 2\eta_2 \langle \lambda \text{sgn}(\lambda) \rangle \\ &+ \frac{\eta_1^2 - \eta_2^2}{T^2 D} \left\langle \sum_{t,t'=1}^T y_t y_{t'} \mathbf{x}_t^\top \mathbf{x}_{t'} \mathbb{I}(\Phi) \right\rangle + \frac{\eta_2^2}{T^2 D} \left\langle \sum_{t,t'=1}^T y_t y_{t'} \mathbf{x}_t^\top \mathbf{x}_{t'} \right\rangle. \end{aligned} \quad (23)$$

567 **Computing Averages:** It remains to compute the expectations in eqs. 22 and 23. All expectations
 568 can be expressed in terms of the constituent expectations given below, which are trivially computed
 569 by considering the Gaussianity of ν and λ and \mathbf{x} :

$$\langle \nu \text{sgn}(\lambda) \rangle = \sqrt{\frac{2}{\pi}} \frac{R}{\sqrt{Q}}, \quad \langle \lambda \text{sgn}(\lambda) \rangle = \sqrt{\frac{2Q}{\pi}}, \quad \langle \nu \text{sgn}(\nu) \rangle = \sqrt{\frac{2}{\pi}}, \quad \langle \lambda \text{sgn}(\nu) \rangle = \sqrt{\frac{2}{\pi}} R \quad (24)$$

$$\frac{1}{D} \left\langle \sum_{t,t'=1}^T y_t y_{t'} \mathbf{x}_t^\top \mathbf{x}_{t'} \right\rangle = \frac{1}{D} \left\langle \left(\sum_{t=1}^T \mathbf{x}_t^\top \mathbf{x}_t + 2 \sum_{t=2}^T \sum_{t'=1}^{t-1} y_t y_{t'} \mathbf{x}_t^\top \mathbf{x}_{t'} \right) \right\rangle \quad (25)$$

$$= T + \mathcal{O}(1/D) \quad (26)$$

570 The terms involving Φ will in general consist of expectations containing step functions $\theta(x)$ (1 for
 571 $x > 0$, 0 otherwise), specifically $\theta(\nu\lambda)$ (1 if student decision agrees with teacher, 0 otherwise) and
 572 $\theta(-\nu\lambda)$ (1 if student decision disagrees with teacher, 0 otherwise). When we encounter these terms,
 573 they can be greatly simplified by considering the following equivalences:

$$\text{sgn}(\lambda)\theta(\nu\lambda) = \frac{1}{2}(\text{sgn}(\lambda) + \text{sgn}(\nu)) \quad \text{and} \quad \text{sgn}(\lambda)\theta(-\nu\lambda) = \frac{1}{2}(\text{sgn}(\lambda) - \text{sgn}(\nu)) \quad (27)$$

574 We show as an example the case where Φ is the condition to get all decisions correct in an episode,
 575 $\mathbb{I}(\Phi) = \prod_{t=1}^T \theta(\nu_t \lambda_t)$, where $\theta(x)$ is the step function (1 for $x > 0$, 0 otherwise). The first term in
 576 Eq. 22 can be addressed:

$$\left\langle \frac{1}{T} \sum_{t=1}^T \nu_t \text{sgn}(\lambda_t) \mathbb{I}(\Phi) \right\rangle \rightarrow \left\langle \frac{1}{T} \sum_{t=1}^T \nu_t \text{sgn}(\lambda_t) \prod_{s=1}^T \theta(\nu_s \lambda_s) \right\rangle \quad (28)$$

$$= \langle \nu_t \text{sgn}(\lambda_t) \theta(\nu_t \lambda_t) \rangle \left\langle \prod_{s \neq t}^T \theta(\nu_s \lambda_s) \right\rangle \quad (29)$$

$$= \frac{1}{2} \langle \nu_t (\text{sgn}(\lambda_t) + \text{sgn}(\nu_t)) \rangle P^{T-1} \quad (30)$$

$$= \frac{1}{\sqrt{2\pi}} \left(1 + \frac{R}{\sqrt{Q}} \right) P^{T-1} \quad (31)$$

577 where P is the probability of making a single correct decision, and can be calculated by considering
 578 that an incorrect decision is made if \mathbf{x} lies in the hypersectors defined by the intersection of $\mathcal{N}_\pm(\cdot|\mathbf{w}^*)$
 579 and $\mathcal{N}_\pm(\cdot|\mathbf{w})$, the angle ϵ subtended by these hypersectors is equal to the angle between \mathbf{w}^* and \mathbf{w} .

$$P = \left(1 - \frac{\epsilon}{\pi} \right) = \left(1 - \frac{1}{\pi} \cos^{-1} \left(\frac{R}{\sqrt{Q}} \right) \right) \quad (32)$$

580 Similarly, the first term in Eq. 23 can be addressed:

$$\left\langle \frac{2}{T} \sum_{t=1}^T \lambda_t \text{sgn}(\lambda_t) \prod_{s=1}^T \theta(\nu_s \lambda_s) \right\rangle = \langle \lambda_t (\text{sgn}(\lambda_t) + \text{sgn}(\nu_t)) \rangle P^{T-1} \quad (33)$$

$$= \sqrt{\frac{2Q}{\pi}} \left(1 + \frac{R}{\sqrt{Q}} \right) P^{T-1} \quad (34)$$

581 The cross terms in 23 can also be computed:

$$\frac{1}{D} \left\langle \sum_{t,t'=1}^T y_t y_{t'} \mathbf{x}_t^\top \mathbf{x}_{t'} \prod_{s=1}^T \theta(\nu_s \lambda_s) \right\rangle = \frac{1}{D} \left\langle \left(\sum_{t=1}^T \mathbf{x}_t^\top \mathbf{x}_{t'} + 2 \sum_{t=2}^T \sum_{t'=1}^{t-1} y_t y_{t'} \mathbf{x}_t^\top \mathbf{x}_{t'} \right) \prod_{s=1}^T \theta(\nu_s \lambda_s) \right\rangle \quad (35)$$

$$= T P^T + \mathcal{O}(1/D) \quad (36)$$

582 where the 2nd term can be neglected in the high dimensional limit. Substituting these computed
583 averages into equations 22 and 23, the ODEs for the order parameters can be written:

$$\frac{dR}{d\alpha} = \frac{\eta_1 + \eta_2}{\sqrt{2\pi}} \left(1 + \frac{R}{\sqrt{Q}} \right) P^{T-1} - \eta_2 R \sqrt{\frac{2}{\pi Q}} \quad (37)$$

$$\frac{dQ}{d\alpha} = (\eta_1 + \eta_2) \sqrt{\frac{2Q}{\pi}} \left(1 + \frac{R}{\sqrt{Q}} \right) P^{T-1} - 2\eta_2 \sqrt{\frac{2Q}{\pi}} + \frac{(\eta_1^2 - \eta_2^2)}{T} P^T + \frac{\eta_2^2}{T} \quad (38)$$

584 **Equivalence of state formulation**

585 The ODEs governing the dynamics of the order parameters in the previous section can be equivalently
586 calculated under the formulation involving the underlying states $\{s_+, s_-\}$ defined in section 1. The
587 underlying system can take a multitude of trajectories (τ) in state space, there are 2^T trajectories in
588 total (as the system can be in 2 possible states at each timestep), and expectations must now include
589 the averaging over all possible trajectories. All expectations will now be of the following form, where
590 the dot (\cdot) denotes some arbitrary term to be averaged over.

$$\langle \cdot \rangle = \sum_{\tau} P(\tau) \langle \cdot \mid \tau \rangle \quad (39)$$

591 By considering symmetry of the Gaussian and ‘half-Gaussian’ (\mathcal{N}_{\pm}) distributions, all expectations in
592 24 can be seen to be identical regardless of whether expectations are taken with respect to the full
593 Gaussian or the half-Gaussian distributions, i.e.

$$\langle \cdot \rangle_{\mathcal{N}} = \langle \cdot \rangle_{\mathcal{N}_+} = \langle \cdot \rangle_{\mathcal{N}_-} \quad (40)$$

594 this implies that all expectations are independent of the trajectory of the underlying system, hence
595 averaging over all trajectories leaves all expectations unchanged. This also allows the extension to
596 arbitrary transition probabilities between the underlying states $\{s_+, s_-\}$.

597 **Other Reward structures**

598 The expectations can be calculated in other conditions of Φ from considering combinatorial argu-
599 ments. We state the ODEs for two reward conditions.

600 **n or more:** the case where Φ is the requirement of getting n or more decisions in an episode of
601 length T correct. We give the ODEs below for the case of $\eta_2 = 0$

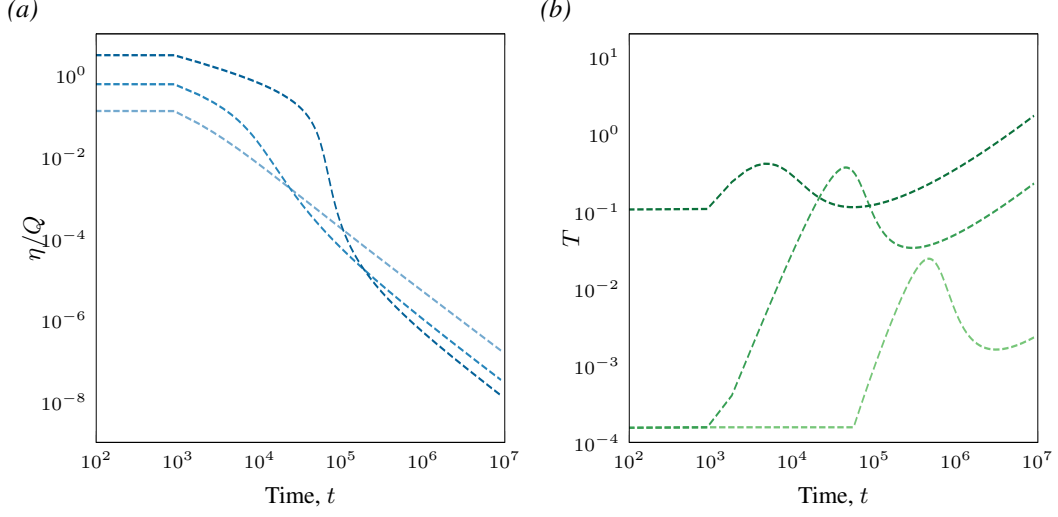


Figure 7: **Optimal Schedules for the unconstrained student.** (a) Evolution of optimal η (a) and T (b) over learning, while following the specified optimal schedule, over a range of rewards and episode lengths. *Parameters:* $D = 900$, $\eta_2 = 0$, (a) $T = 8$, (b) $\eta = 1$.

$$\frac{dR}{d\alpha} = \frac{\eta_1}{T\sqrt{2\pi}} \sum_{i=n}^T \binom{T}{i} \left[i \left(1 + \frac{R}{\sqrt{Q}} \right) (1-P) - (T-i) \left(1 - \frac{R}{\sqrt{Q}} \right) P \right] P^{i-1} (1-P)^{T-i-1} \quad (41)$$

$$\begin{aligned} \frac{dQ}{d\alpha} = & \frac{\eta_1}{T} \sqrt{\frac{2Q}{\pi}} \sum_{i=n}^T \binom{T}{i} \left[i \left(1 + \frac{R}{\sqrt{Q}} \right) (1-P) - (T-i) \left(1 - \frac{R}{\sqrt{Q}} \right) P \right] P^{i-1} (1-P)^{T-i-1} \\ & + \frac{\eta_1^2}{T} \sum_{i=n}^T \binom{T}{i} P^i (1-P)^{T-i} \end{aligned} \quad (42)$$

602 **Breadcrumb Trails** We also consider the case where a reward of size η_1 is received if all decisions in
 603 an episode are correct in addition to a smaller reward of size β for each individual decision correctly
 604 made in an episode:

$$\frac{dR}{d\alpha} = \frac{1}{\sqrt{2\pi}} \left(1 + \frac{R}{\sqrt{Q}} \right) (\eta_1 P^{T-1} + \beta) + \beta(T-1) \sqrt{\frac{2}{\pi}} \frac{R}{\sqrt{Q}} P \quad (43)$$

$$\begin{aligned} \frac{dQ}{d\alpha} = & \sqrt{\frac{2Q}{\pi}} \left(1 + \frac{R}{\sqrt{Q}} \right) (\eta_1 P^{T-1} + \beta) + 2\beta(T-1) \sqrt{\frac{2Q}{\pi}} P \\ & + \left(\frac{\eta_1^2}{T} + 2\eta_1\beta \right) P^T + \beta^2 (1 + (T-1)P) P \end{aligned} \quad (44)$$

605 D Theory extension

606 **Optimal scheduling (Unconstrained)** The optimal schedules for learning rate and episode length
 607 (eq. (8)) hold in the unconstrained case too (where $Q(\alpha)$ isn't restricted to the surface of a sphere);
 608 this is because the parameters were derived from the general requirement of extremising the update
 609 of ρ from any point in the (ρ, Q) plane. The evolution of T_{opt} and η_{opt} over time (while following

their respective scheduling) is shown in fig. 7. In the unconstrained case the magnitude of the student grows quadratically, an increase Q acts as a decrease in effective learning rate. Hence contrary to the spherical case a decaying learning rate is not optimal, and optimal T grows much slower, as shown in fig. 7b; the plots for T_{opt} do not show a clear trend and require further investigation. The evolution of $\eta_{\text{opt}}/\sqrt{Q}$ is plotted in fig. 7a, this value is the effective learning rate and we observe a polynomial decay in the value as with the spherical case presented in section 2.3.

Phases (Unconstrained) The phases observed in fig. 4 are not an artifact of the spherical case. When $Q(\alpha)$ is not constrained we also observe regimes where a ‘bad’ fixed point of ρ may be attained. Figure 8 shows flow diagrams in the (ρ, Q) plane for various parameter instantiations in the case where a reward of $\eta_1 = 1$ is received if all decisions in an episode of length $T = 8$ are correctly made, and a penalty of η_2 otherwise. Figure 8a is the flow diagram for $\eta_2 = 0$, in this regime the agent can always perfectly align with the teacher from any initialisation (the student flows to $\rho = 1$ at $Q = \infty$). This is analogous to the student being in the easy phase at the *bottom of the plot* in fig. 4b, as with probability 1 the algorithm naturally converges to the optimal $\rho = 1$. Figure 8b shows the flow for $\eta_2 = 0.05$; in this regime we observe the flow to some suboptimal ρ at $Q = \infty$; this is analogous to the student being in the easy phase at the *top of the plot* in fig. 4b, as with probability 1 the algorithm converges to a value of ρ from any initialisation. However, this value of ρ is suboptimal. Figure 8c shows the flow for $\eta_2 = 0.045$, we see that depending on the initial ρ , the agent will flow to one of two fixed point in ρ at $Q = \infty$; this is analogous to the agent being in the *hybrid-hard* phase in fig. 4b, where with high probability the agent converges to the worse ρ . The ‘good easy phase’, characterising the behaviour seen in fig. 8a, is indicated by the green region in fig. 8d.

Critical Slowing down With the addition of a penalty term we observe initial speed up in learning as shown in fig. 9. Towards the end of learning, however, we observe a critical slowing down, and see how in many instances a non-zero η_2 can instead give an overall slowing to learning. This is most easily seen in the spherical case for the rule where all decisions in an episode of length T must be correct for a reward: fig. 9a shows the times to reach 0.99 of the fixed point starting from an initial $\rho = 0$ for $T = 13$ and $\eta_1 = 1$. We observe that increasing η_2 (up to η_{crit} , at which point the algorithm enters the hybrid-hard phase detailed in section 2.4) increases the time taken to reach the fixed point. This is similarly seen for $T = 20$ in fig. 9b. This slowing is not present over the entire range of η_2 ; it is true that for small values of η_2 there is actually a small speed up in the reaching of the fixed point, showing that the criticality severely reduces the range of η_2 that improves convergence speed. We plot the distance of η_2 away from the critical penalty value ($|\eta_2 - \eta_{\text{crit}}|$) against time for convergence in fig. 9a, for $T = 20$ (top) and $T = 13$ (bottom). We observe a polynomial scaling of the convergence time with distance away from criticality.

E Experiments

Pong speed accuracy

For another verification of the speed accuracy tradeoff introduced in section 2.5 we train agents from pixels on the ALE [Machado et al., 2018] game ‘Pong’. The notion of lives (or requiring n or more correct decisions in an episode for a reward) is essentially a way to control the difficulty of a task, whereby higher n (fewer lives) is a more stringent condition i.e. a more difficult task. We examine a corresponding setup in pong, where task difficulty is varied in order to study generalisation performance of agents. The Pong task difficulty is varied by changing the episode length after which the agent receives a reward, where intuitively longer episode length is a more difficult task. On each timestep the agent has a binary choice of moving left/right and aims to return the ball. If the ball manages to get past the agent the episode ends without reward, if the agent survives until the end of the episode, it receives a reward. The decisions of the agent are sampled from the outputs of a deep policy network (detailed in code). Pong is deterministic, so in order to introduce stochasticity, we employ ‘Sticky actions’, where instead of taking the action sampled from the policy, the agent instead with probability 0.2 repeats the previous action. The weights of the policy network are trained using the policy gradient update eq. (9) by running 20 agents in parallel. To study the speed-accuracy trade-off, we train agents that require a different number of timesteps before receiving reward. We trained from 7 different random seeds. The results are shown in fig. 10, we observe a speed-accuracy tradeoff, but in this setup the tradeoff is mediated by episode length T (instead of lives). We see that agents trained on shorter episode lengths initially learn much faster, but reach a lower asymptotic accuracy.

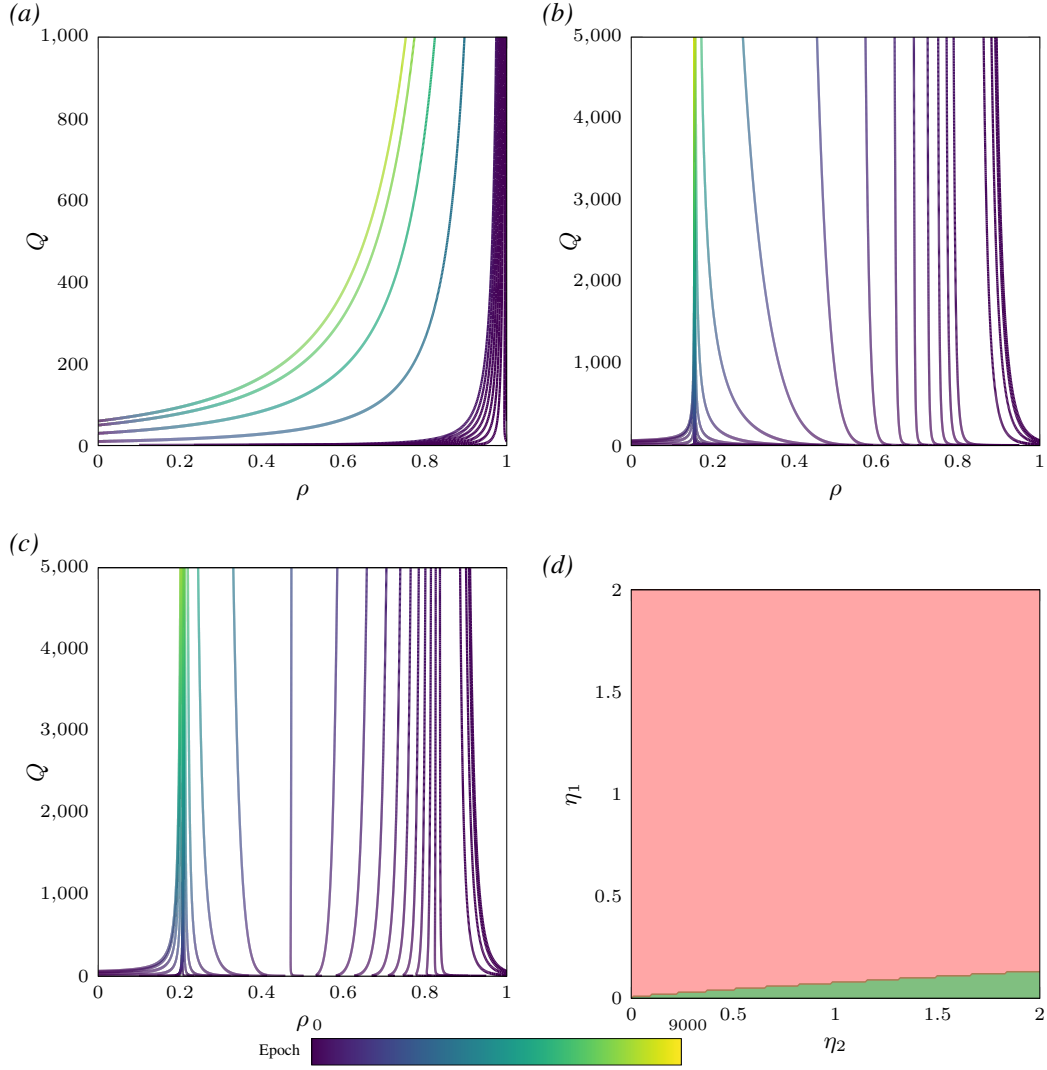


Figure 8: **Flow and phase plots.** Flow in the (ρ, Q) plane for the case where all decisions in an episode are required correct for a reward of $\eta_1 = 1$ and a penalty other wise of $\eta_2 = 0$ (a), 0.05 (b), and 0.045 (c). (d) Phase plot showing the region where learning failed (red) and succeeded (green) over the η_1, η_2 plane, for the same learning rule. *Parameters: Initialised from $\rho = 0$ and $Q = 1$*

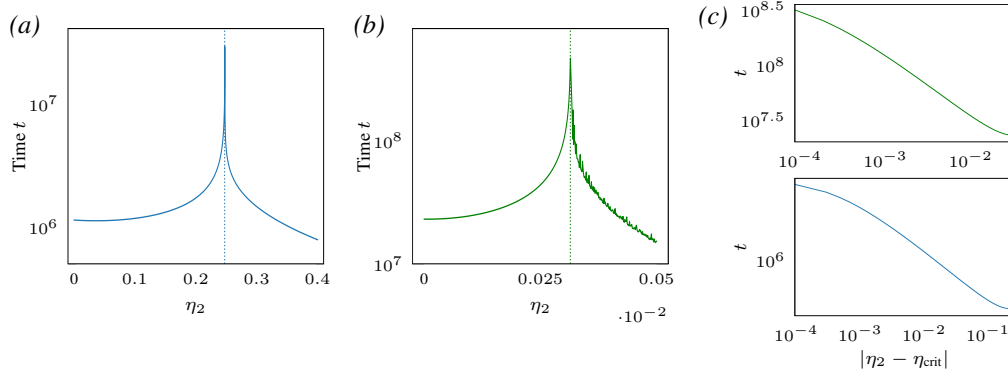


Figure 9: **Critical slowing down for $\eta_2 \neq 0$.** The above plots are for the spherical case where the agent must get every decision correct in order to receive a reward of $\eta_1 = 1$, and a penalty of η_2 otherwise. (a) The time for convergence to the fixed point for $T = 13$. (b) The time for convergence to the fixed point for $T = 20$ (c) Time for convergence plotted against distance of η_2 away from the critical penalty for $T = 13$ (bottom) and $T = 20$ (top) Parameters: $D = 900$, $Q = 1$, $\eta_1 = 1$

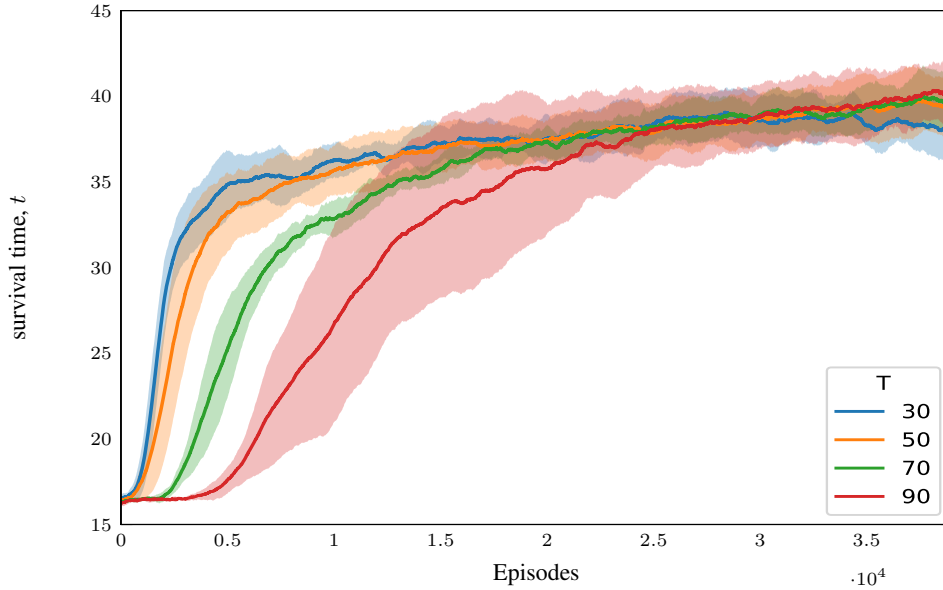


Figure 10: **Speed-accuracy tradeoff for Pong.** The mean survival time over the course of training for agents required to survive up to completion of an episode of length T in order to receive reward. Parameters: $\eta_1 = 2$, $\eta_2 = 0$