THE RL PERCEPTRON: DYNAMICS OF POLICY LEARNING IN HIGH DIMENSIONS

Nishil Patel[1*], **Sebastian Lee**[1,2], **Stefano Sarao Mannelli**[1],
Sebastian Goldt[3] **& Andrew M. Saxe**[1,4*]
[1]Gatsby Computational Neuroscience Unit & Sainsbury Wellcome Centre, UCL
[2]Imperial College, London, UK
[3]International School of Advanced Studies (SISSA), Trieste, Italy
[4]CIFAR Azrieli Global Scholar, CIFAR

ABSTRACT

Reinforcement learning (RL) algorithms have proven transformative in a range of domains. To tackle real-world domains, these systems often use neural networks to learn policies directly from pixels or other high-dimensional sensory input. By contrast, much theory of RL has focused on discrete state spaces or worst case analyses, and fundamental questions remain about the dynamics of policy learning in high-dimensional settings. Here we propose a simple high-dimensional model of RL and derive its typical dynamics as a set of closed-form ODEs. We show that the model exhibits rich behavior including delayed learning under sparse rewards; a speed-accuracy trade-off depending on reward stringency; and a dependence of learning regime on reward baselines. These results offer a first step toward understanding policy gradient methods in high dimensional settings.

1 INTRODUCTION

Recent years have seen rapid progress in Reinforcement Learning (RL), both from an algorithmic and engineering standpoint, leading to super-human performance in a variety of domains including complex games (Silver et al., 2016; Mnih et al., 2015). Despite this, theory—particularly for high-dimensional problems requiring non-linear function approximation—is still limited (Yang et al., 2020). Many theoretical results to date are based on tabular RL, where the state and action spaces are small enough for value functions to be represented as tabular data (i.e. without function approximation). Because of the curse of dimensionality, these methods are limited to low-dimensional problems. Further, tabular methods cannot generalise to unseen states and across seen states. Consequently, much of this theoretical work provides limited insight to the RL practitioner, who increasingly relies on deep neural networks to approximate value functions, policies and associated building blocks of RL. Moreover, while RL theory has addressed 'worst-case' performance and convergence behavior (Du et al., 2020; 2021), fewer methods exist to characterise typical behaviour. On the other hand, there is a growing sub-field of deep learning theory dedicated to employing tools from statistical mechanics to analyse various learning paradigms in the *average-case*, see Seung et al. (1992); Engel & Van den Broeck (2001); Carleo et al. (2019); Bahri et al. (2020) for reviews. While this approach has recently been extended to curriculum learning (Saglietti et al., 2022), continual learning (Lee et al., 2021; 2022), few-shot learning (Sorscher et al., 2022) and transfer learning (Gerace et al., 2022), RL algorithms with function approximation are yet to be analysed under the framework of statistical mechanics—a gap we attempt to address here.

We build on the Teacher-Student framework widely used in the statistical physics community, in particular the classic works of Gardner & Derrida (1989), Biehl & Schwarze (1995), and Saad & Solla (1995). Our work adapts these methods to the RL setting. We present a novel application of the teacher-student framework to a sequential policy learning task with sparse delayed reward, which allows to characterize the typical learning curves of policy gradient RL agents with an exact set of Ordinary Differential Equations (ODEs). The ODEs permit efficient exploration of learning behaviour in a large number of scenarios, highlighting the advantages and disadvantages of different

*Correspondence to: {ucabnp2,a.saxe}@ucl.ac.uk

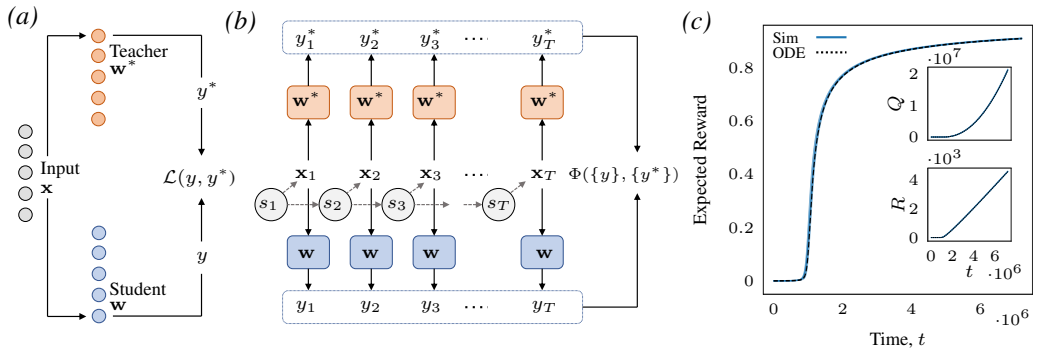

Figure 1: Overview of the RL-Perceptron. *(a)* Schematic of classic supervised teacher-student perceptron model. *(b)* Schematic of the RL-Perceptron setup, with sequential inputs $\mathbf{x}_i$ conditioned on underlying states s_i. *(c)* Comparison of simulation with ODE solution for evolution of expected reward, and order parameters R and Q (inset), for the case where all decisions must be correct in an episode of length 12 and $\eta_2 = 0$.

algorithmic choices. After introducing the model, we explore its rich behavior, including a delay in learning from sparse rewards; a speed-accuracy trade-off with varying reward stringency; and the effect of reward baselines on policy gradient learning.

2 Setup

Teacher-student models provide an analytically tractable model of a learning setting. In the classic supervised learning version, shown in Fig. 1a, a teacher perceptron defines a rule that a student perceptron must learn. Inputs are sampled from i.i.d. Gaussians $\mathbf{x}_t \sim \mathcal{N}(\mathbf{0}, \mathbb{1}_D)$, and labeled by the teacher $y_t^* = \text{sgn}(\mathbf{w}^{*\intercal}\mathbf{x}_t)$, with $\mathbf{w}^*$ uniformly drawn from a D-dimensional hypersphere of radius $\sqrt{D}$. The student learns from the dataset $\mathcal{D} = \{(\mathbf{x}_t, y_t)\}_t$ using a given loss and optimizer (e.g. MSE and stochastic gradient descent).

In RL, agents face a sequential decision-making task in which a sequence of correct intermediate choices is required to succeed in an episode. We schematise this process into the 'RL perceptron' depicted in Fig. 1b: the student $\mathbf{w}$ takes a sequence of T choices over an episode. The correct choices are governed by the same teacher network $\mathbf{w}^*$, i.e. the same underlying rule throughout every time-step of every episode. Crucially, unlike the supervised learning setting, the student does not observe the correct choice for each input; instead, it receives a reward which can depend on whether earlier decisions are correct. For instance, the student could receive reward only if all T choices are correct, and no reward otherwise—a considerably less informative learning signal.

Because the student in the RL perceptron makes choices in response to high dimensional inputs, it is analogous to a policy network. We therefore consider a policy gradient learning update inspired by the REINFORCE algorithm. At the end of the μth episode, $\mathbf{w}$ is updated according to the equation:

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

where $\mathbb{I}$ is an indicator function and Φ is a condition that depends on the student's decisions. For instance, Φ might be the condition that the student is correct at every timestep. The update is general in the sense that it prescribes a 'positive' update for the fulfillment of the condition, and a 'negative' update otherwise ($1 - \mathbb{I}(\Phi) = 1$). Note that in the case of $T = 1$, $\eta_2 = 0$, this update is equivalent to the perceptron learning rule for the classic supervised perceptron (Biehl & Riegler, 1994) (for suitably chosen Φ).

Using methods from statistical physics—details are reported in Appendix A—we can define and track the evolution of few order parameters, $R = \mathbf{w}^{*\intercal}\mathbf{w}/D$ and $Q = \mathbf{w}^{\intercal}\mathbf{w}/D$, that can be used to evaluate the expected reward at all times during training, as shown in Fig. 1c.

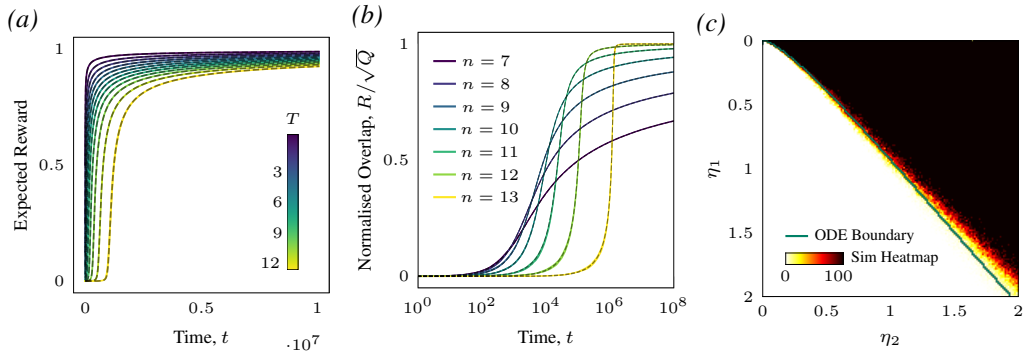

Figure 2: Learning dynamics. *(a)* Evolution of the expected reward for simulation (solid) and ODE solution (dashed) when all decisions in an episode of length T are required correct, and $\eta_2 = 0$. *(b)* Evolution of the normalised angle between student and teacher weights for simulation (solid) and ODE solution (dashed) for the case where n or more decisions in an episode of length 13 are required correct for an update with $\eta_2 = 0$. *(c)* Heat map of average time for learning to begin over a range of learning rates overlayed with region well predicted to learn by ODEs, for an initialisation of 1.57 radians of $\mathbf{w}$ relative to $\mathbf{w}^*$.

REINFORCE: This framework and update 1 can be grounded with *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 $\pi(y_t \mid \mathbf{x}_t)$, 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. 2. 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 $N(\mathbf{0}, \mathbb{1}_D)$ 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 $\mathbf{w}^*$ — ($N(\mathbf{0}, \mathbb{1}_D)$ 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 (unfulfillment) of Φ. Formally, the mapping to the RL setting can be stated by introducing the states and a deterministic policy $\pi_{\mathbf{w}}(y \mid \mathbf{x}) = 1/(1 + \exp -y\mathbf{w}^\intercal\mathbf{x}/\sqrt{D})$. The REINFORCE policy gradient update in this case is

$$\nabla_w J = \left\langle \sum_{t=0}^{T-1} \nabla_{\mathbf{w}} \log \pi_{\mathbf{w}}(y_t \mid \mathbf{x}_t) \left(\sum_{t'=t+1}^{T} r_{t'} \right) \right\rangle \approx \left\langle \sum_{t=0}^{T-1} y_t \mathbf{x}_t \left[\eta_1 \mathbb{I}(\Phi) - \eta_2 (1 - \mathbb{I}(\Phi)) \right] \right\rangle \quad (2)$$

$$\longrightarrow \Delta\mathbf{w} \propto \eta_1 \left\langle \sum_{t=0}^{T-1} y_t \mathbf{x}_t \mathbb{I}(\Phi) \right\rangle - \eta_2 \left\langle \sum_{t=0}^{T-1} y_t \mathbf{x}_t (1 - \mathbb{I}(\Phi)) \right\rangle \quad (3)$$

The approximation in eq. 2 holds in the early phases of learning – when $\mathbf{w}^\intercal\mathbf{x}/\sqrt{D}$ is small (Appendix B)– 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 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.

3 Results

We derive a set of dynamical equations that describe the dynamics of the student in the high-dimensional limit where the input dimension goes to infinity. We give explicit dynamics for different reward conditions Φ, namely requiring all decisions correct in an episode of length T; requiring n or more decisions correct in an episode of length T; and receiving reward for each

correct response. Due to the length of these expressions, we report the generic expression of the updates here and give the full forms and derivations in Appendix A.

$$\frac{dR}{d\alpha} = (\eta_1 + \eta_2)U_{R,1}(R, Q; \Phi) - \eta_2 U_{R,2}(R, Q), \tag{4}$$

$$\begin{aligned}\frac{dQ}{d\alpha} &= \frac{2\,(\eta_1 + \eta_2)}{T} U_{Q,1}(R, Q; \Phi) - \frac{2\eta_2}{T} U_{Q,2}(R, Q) \\ &\quad + \frac{\eta_1^2 - \eta_2^2}{T^2} U_{Q,3}(R, Q; \Phi) + \frac{\eta_2^2}{T^2} U_{Q,4}(R, Q);\end{aligned} \tag{5}$$

where α serves as a continuous time variable (not to be confused with t which counts episode steps). The validity of these equations for the case where all decisions must be correctly made for episodes of length $T = 12$ is shown in Fig. 1c. All simulations and corresponding numerical solutions of ODEs are done with $D = 900$. We now turn to several phenomena exhibited by this model.

Learning plateaus and sparse rewards. When rewarded for getting all T decisions correct, we observe a characteristic initial plateau in expected reward followed by a rapid jump (Fig. 2a). The length of this plateau increases with T, consistent with the notion that sparse rewards make exploration hard and slows learning (Bellemare et al., 2016). Plateaus during learning, which arise from saddle points in the loss landscape, have also been studied for (deep) neural networks in the supervised setting (Saad & Solla, 1995; Dauphin et al., 2014), but do not arise in the supervised perceptron. Hence the RL setting can qualitatively change the learning trajectory.

Speed-accuracy trade-off and reward stringency. Fig. 2b shows the evolution of the cosine angle between the student and teacher for simulation and ODE, in the case where n or more decisions must be correctly made in an episode of length 13 in order to receive a reward (and $\eta_2 = 0$). We observe a speed-accuracy trade-off, where decreasing n increases the initial speed of learning but leads to worse asymptotic performance. In essence, a lax reward function is probabilistically more achievable early in learning; but it rewards some fraction of incorrect decisions, leading to lower asymptotic accuracy. By contrast a stringent reward function slows learning but eventually produces a highly aligned student. For a given MDP, it is known that arbitrary shaping applied to the reward function will change the optimal policy (reduce asymptotic performance) (Ng et al., 1999). Empirically, reward shaping has been shown to speed up learning and help overcome difficult exploration problems (Gullapalli & Barto, 1992). Reconciling these results with the phenomena observed in our setting is an interesting avenue for future work.

Reward baselines and learning regimes. A common problem with REINFORCE is high variance gradient estimates leading to bad performance (Marbach & Tsitsiklis, 2003; Schulman et al., 2015). The reward (η_1) and punishment (η_2) magnitude alters the variance of the updates, potentially changing learning behavior. Indeed, we observe different regimes in the learning dynamics. Fig. 2c shows the time for learning to begin in simulations over a range of η_1, η_2, explained in Appendix C. In one regime, learning is stereotyped and well predicted by the ODEs (white regions); in a second fluctuation driven regime (coloured regions), learning still succeeds but depends on small fluctuations, making the exact learning time vary widely across runs. These regimes depend upon the initialisation of the student, reward and punishment magnitudes, and the condition for reward Φ; and the ODEs describe this boundary well (solid green line). This framework opens the possibility for studying phase transitions between learning regimes (Gamarnik et al., 2022).

4 Conclusion

Our proposed teacher-student model provides a framework to investigate the REINFORCE policy gradient method in RL for a range of plausible sparse reward structures. We derive closed ODEs that capture the *average-case* learning dynamics in high-dimensional settings and permit a reduction in cost of calculating learning behavior compared to simulation-based methods for large numbers of parameters. Our framework offers a starting point to explore additional settings that are closer to many real-world RL scenarios, such as those with conditional next states. Furthermore, the RL perceptron offers a means to study common training practices, including curricula; and more advanced algorithms, like actor-critic methods. We hope to extract more analytical insights from the ODEs, particularly on how initialization and learning rate influence an agent's learning regime. Our findings emphasize the intricate interplay of task, reward, architecture, and algorithm in modern RL systems.

ACKNOWLEDGMENTS

This work was supported by a Sir Henry Dale Fellowship from the Wellcome Trust and Royal Society (216386/Z/19/Z) to A.S., and the Sainsbury Wellcome Centre Core Grant from Wellcome (219627/Z/19/Z) and the Gatsby Charitable Foundation (GAT3755). A.S. is a CIFAR Azrieli Global Scholar in the Learning in Machines & Brains program.

REFERENCES

Yasaman Bahri, Jonathan Kadmon, Jeffrey Pennington, Sam S. Schoenholz, Jascha Sohl-Dickstein, and Surya Ganguli. Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11(1):501–528, 2020. doi: 10.1146/annurev-conmatphys-031119-050745. URL https://doi.org/10.1146/annurev-conmatphys-031119-050745.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. *Advances in neural information processing systems*, 29, 2016.

Michael Biehl and Peter Riegler. On-line learning with a perceptron. *Europhysics Letters*, 28(7): 525, 1994.

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

Giuseppe Carleo, Ignacio Cirac, Kyle Cranmer, Laurent Daudet, Maria Schuld, Naftali Tishby, Leslie Vogt-Maranto, and Lenka Zdeborová. Machine learning and the physical sciences. *Reviews of Modern Physics*, 91(4):045002, 2019.

Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 27, 2014.

Simon S. Du, Sham M. Kakade, Ruosong Wang, and Lin F. Yang. Is a good representation sufficient for sample efficient reinforcement learning?, 2020.

Simon S. Du, Sham M. Kakade, Jason D. Lee, Shachar Lovett, Gaurav Mahajan, Wen Sun, and Ruosong Wang. Bilinear classes: A structural framework for provable generalization in rl, 2021.

Andreas Engel and Christian Van den Broeck. *Statistical mechanics of learning*. Cambridge University Press, 2001.

David Gamarnik, Cristopher Moore, and Lenka Zdeborová . Disordered systems insights on computational hardness. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11): 114015, nov 2022. doi: 10.1088/1742-5468/ac9cc8. URL https://doi.org/10.1088%2F1742-5468%2Fac9cc8.

Elizabeth Gardner and Bernard Derrida. Three unfinished works on the optimal storage capacity of networks. *Journal of Physics A: Mathematical and General*, 22(12):1983, 1989.

Federica Gerace, Luca Saglietti, Stefano Sarao Mannelli, Andrew Saxe, and Lenka Zdeborová. Probing transfer learning with a model of synthetic correlated datasets. *Machine Learning: Science and Technology*, 3(1):015030, 2022.

Vijaykumar Gullapalli and Andrew G Barto. Shaping as a method for accelerating reinforcement learning. In *Proceedings of the 1992 IEEE international symposium on intelligent control*, pp. 554–559. IEEE, 1992.

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

Sebastian Lee, Stefano Sarao Mannelli, Claudia Clopath, Sebastian Goldt, and Andrew Saxe. Maslow's hammer for catastrophic forgetting: Node re-use vs node activation. *arXiv preprint arXiv:2205.09029*, 2022.

Peter Marbach and John N. Tsitsiklis. Approximate gradient methods in policy-space optimization of markov reward processes. *Discrete Event Dynamic Systems*, 13:111–148, 2003.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Andrew Y Ng, Daishi Harada, and Stuart Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Icml*, volume 99, pp. 278–287. Citeseer, 1999.

David Saad and Sara A Solla. On-line learning in soft committee machines. *Physical Review E*, 52 (4):4225, 1995.

Luca Saglietti, Stefano Sarao Mannelli, and Andrew Saxe. An analytical theory of curriculum learning in teacher–student networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2022(11):114014, 2022.

John Schulman, Philipp Moritz, Sergey Levine, Michael Jordan, and Pieter Abbeel. High-dimensional continuous control using generalized advantage estimation, 2015. URL https://arxiv.org/abs/1506.02438.

Hyunjune Sebastian Seung, Haim Sompolinsky, and Naftali Tishby. Statistical mechanics of learning from examples. *Physical review A*, 45(8):6056, 1992.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489, 2016.

Ben Sorscher, Surya Ganguli, and Haim Sompolinsky. Neural representational geometry underlies few-shot concept learning. *Proceedings of the National Academy of Sciences*, 119(43): e2200800119, 2022. doi: 10.1073/pnas.2200800119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2200800119.

R. S. Sutton, D. Mcallester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems 12*, volume 12, pp. 1057–1063. MIT Press, 2000.

Zhuoran Yang, Chi Jin, Zhaoran Wang, Mengdi Wang, and Michael I. Jordan. On function approximation in reinforcement learning: Optimism in the face of large state spaces, 2020.

A Appendix

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 $\mathbf{w}^*$.

$$\mathbf{w}^{\mu+1} = \mathbf{w}^{\mu} + \frac{\eta_1}{\sqrt{D}}\left(\frac{1}{T}\sum_{t=1}^{T} y_t \mathbf{x}_t \mathbb{I}(\Phi)\right)^{\mu} - \frac{\eta_2}{\sqrt{D}}\left(\frac{1}{T}\sum_{t=1}^{T} y_t \mathbf{x}_t (1-\mathbb{I}(\Phi))\right)^{\mu} \tag{6}$$

$$DR^{\mu+1} = DR^{\mu} + \frac{\eta_1}{\sqrt{D}}\left(\frac{1}{T}\sum_{t=1}^{T} y_t \mathbf{w}^{*\intercal} \mathbf{x}_t \mathbb{I}(\Phi)\right)^{\mu} - \frac{\eta_2}{\sqrt{D}}\left(\frac{1}{T}\sum_{t=1}^{T} y_t \mathbf{w}^{*\intercal} \mathbf{x}_t (1-\mathbb{I}(\Phi))\right)^{\mu} \tag{7}$$

we subtract DR^{μ} and sum over l episodes, the LHS is a telescopic sum, and 7 becomes

$$\begin{aligned}\frac{D(R^{\mu+l}-R^{\mu})}{l} &= \frac{\eta_1}{\sqrt{D}}\frac{1}{l}\sum_{i=0}^{l-1}\left(\frac{1}{T}\sum_{t=1}^{T} y_t \mathbf{w}^{*\intercal}\mathbf{x}_t \mathbb{I}(\Phi)\right)^{\mu+i} \\ &- \frac{\eta_2}{\sqrt{D}}\frac{1}{l}\sum_{i=0}^{l-1}\left(\frac{1}{T}\sum_{t=1}^{T} y_t \mathbf{w}^{*\intercal}\mathbf{x}_t (1-\mathbb{I}(\Phi))\right)^{\mu+i}\end{aligned} \tag{8}$$

$$\frac{dR}{d\alpha} = \frac{\eta_1+\eta_2}{\sqrt{D}}\left\langle \frac{1}{T}\sum_{t=1}^{T} y_t \mathbf{w}^{*\intercal}\mathbf{x}_t \mathbb{I}(\Phi)\right\rangle - \frac{\eta_2}{\sqrt{D}}\left\langle \frac{1}{T}\sum_{t=1}^{T} y_t \mathbf{w}^{*\intercal}\mathbf{x}_t\right\rangle \tag{9}$$

We go from eq. 8 to eq. 9 by taking the limit $D \to \infty$, $l \to \infty$ and $l/D = d\alpha \to 0$. The RHS of eq. 9 is a sum of a large number of random variables, and by the central limit theorem is self-averaging in the thermodynamic limit (under the assumption of weak correlations between episodes), consequently the LHS is self-averaging. A similar procedure can be followed for order parameter Q, but we instead take the square of eq. 1 and got to the limit described, obtaining:

$$\begin{aligned}\frac{dQ}{d\alpha} &= \frac{2(\eta_1+\eta_2)}{T\sqrt{D}}\left\langle \sum_{t=1}^{T} y_t \mathbf{w}^{\intercal}\mathbf{x}_t \mathbb{I}(\Phi)\right\rangle - \frac{2\eta_2}{T\sqrt{D}}\left\langle \sum_{t=1}^{T} y_t \mathbf{w}^{\intercal}\mathbf{x}_t\right\rangle \\ &+ \frac{\eta_1^2-\eta_2^2}{T^2 D}\left\langle \sum_{t,t'=1}^{T} y_t y_{t'} \mathbf{x}_t^{\intercal}\mathbf{x}_{t'} \mathbb{I}(\Phi)\right\rangle + \frac{\eta_2^2}{T^2 D}\left\langle \sum_{t,t'=1}^{T} y_t y_{t'} \mathbf{x}_t^{\intercal}\mathbf{x}_{t'}\right\rangle\end{aligned} \tag{10}$$

We introduce auxiliary variables (known in the literature as the aligning fields)

$$\nu = \frac{\mathbf{w}^{*\intercal}\mathbf{x}}{\sqrt{D}} \quad \text{and} \quad \lambda = \frac{\mathbf{w}^{\intercal}\mathbf{x}}{\sqrt{D}}, \tag{11}$$

which are sums of N independent terms and by the central limit theorem they obey a Gaussian distribution. One finds

$$\langle \nu \rangle = \langle \lambda \rangle = 0 \tag{12}$$

$$\langle \nu^2 \rangle = D \quad , \quad \langle \lambda^2 \rangle = DQ \tag{13}$$

$$\langle \nu\lambda \rangle = \mathbf{w}^{*\intercal}\mathbf{w} = DR \tag{14}$$

Substituting 11 into 9 and 10 we can rewrite as

$$\frac{dR}{d\alpha} = \frac{\eta_1+\eta_2}{T}\left\langle \sum_{t=1}^{T} \nu_t \mathrm{sgn}(\lambda_t)\mathbb{I}(\Phi)\right\rangle - \eta_2 \left\langle \nu \mathrm{sgn}(\lambda)\right\rangle \tag{15}$$

$$\begin{aligned}\frac{dQ}{d\alpha} &= \frac{2(\eta_1+\eta_2)}{T}\left\langle \sum_{t=1}^{T} \lambda_t \mathrm{sgn}(\lambda_t)\mathbb{I}(\Phi)\right\rangle - 2\eta_2 \left\langle \lambda \mathrm{sgn}(\lambda)\right\rangle \\ &+ \frac{\eta_1^2-\eta_2^2}{T^2 D}\left\langle \sum_{t,t'=1}^{T} y_t y_{t'} \mathbf{x}_t^{\intercal}\mathbf{x}_{t'} \mathbb{I}(\Phi)\right\rangle + \frac{\eta_2^2}{T^2 D}\left\langle \sum_{t,t'=1}^{T} y_t y_{t'} \mathbf{x}_t^{\intercal}\mathbf{x}_{t'}\right\rangle.\end{aligned} \tag{16}$$

Computing Averages: It remains to compute the expectations in eqs. 15 and 16. All expectations can be expressed in terms of the constituent expectations given below, which are trivially computed by considering the Gaussianity of ν and λ and $\mathbf{x}$:

$$\langle \nu \mathrm{sgn}(\lambda) \rangle = \sqrt{\frac{2}{\pi}} \frac{R}{\sqrt{Q}}, \quad \langle \lambda \mathrm{sgn}(\lambda) \rangle = \sqrt{\frac{2Q}{\pi}}, \quad \langle \nu \mathrm{sgn}(\nu) \rangle = \sqrt{\frac{2}{\pi}}, \quad \langle \lambda \mathrm{sgn}(\nu) \rangle = \sqrt{\frac{2}{\pi}} R \tag{17}$$

$$\frac{1}{D} \left\langle \sum_{t,t'=1}^{T} y_t y_{t'} \mathbf{x}_t^\intercal \mathbf{x}_{t'} \right\rangle = \frac{1}{D} \left\langle \left(\sum_{t=1}^{T} \mathbf{x}_t^\intercal \mathbf{x}_t + 2 \sum_{t=2}^{T} \sum_{t'=1}^{t-1} y_t y_{t'} \mathbf{x}_t^\intercal \mathbf{x}_{t'} \right) \right\rangle \tag{18}$$

$$= T + \mathcal{O}(1/D) \tag{19}$$

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

$$\mathrm{sgn}(\lambda)\theta(\nu\lambda) = \frac{1}{2}(\mathrm{sgn}(\lambda) + \mathrm{sgn}(\nu)) \quad \text{and} \quad \mathrm{sgn}(\lambda)\theta(-\nu\lambda) = \frac{1}{2}(\mathrm{sgn}(\lambda) - \mathrm{sgn}(\nu)) \tag{20}$$

We show as an example the case where Φ is the condition to get all decisions correct in an episode, $\mathbb{I}(\Phi) = \prod_{t=1}^{T} \theta(\nu_t \lambda_t)$, where $\theta(x)$ is the step function (1 for $x > 0$, 0 otherwise). The first term in Eq. 15 can be addressed:

$$\left\langle \frac{1}{T} \sum_{t=1}^{T} \nu_t \mathrm{sgn}(\lambda_t) \mathbb{I}(\Phi) \right\rangle \to \left\langle \frac{1}{T} \sum_{t=1}^{T} \nu_t \mathrm{sgn}(\lambda_t) \prod_{s=1}^{T} \theta(\nu_s \lambda_s) \right\rangle \tag{21}$$

$$= \langle \nu_t \mathrm{sgn}(\lambda_t) \theta(\nu_t \lambda_t) \rangle \left\langle \prod_{s \neq t}^{T} \theta(\nu_s \lambda_s) \right\rangle \tag{22}$$

$$= \frac{1}{2} \langle \nu_t (\mathrm{sgn}(\lambda_t) + \mathrm{sgn}(\nu_t)) \rangle P^{T-1} \tag{23}$$

$$= \frac{1}{\sqrt{2\pi}} \left(1 + \frac{R}{\sqrt{Q}} \right) P^{T-1} \tag{24}$$

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

$$P = \left(1 - \frac{\epsilon}{\pi} \right) = \left(1 - \frac{1}{\pi} \cos^{-1} \left(\frac{R}{\sqrt{Q}} \right) \right) \tag{25}$$

Similarly, the first term in Eq. 16 can be addressed:

$$\left\langle \frac{2}{T} \sum_{t=1}^{T} \lambda_t \mathrm{sgn}(\lambda_t) \prod_{s=1}^{T} \theta(\nu_s \lambda_s) \right\rangle = \langle \lambda_t (\mathrm{sgn}(\lambda_t) + \mathrm{sgn}(\nu_t)) \rangle P^{T-1} \tag{26}$$

$$= \sqrt{\frac{2Q}{\pi}} \left(1 + \frac{R}{\sqrt{Q}} \right) P^{T-1} \tag{27}$$

The cross terms in 16 can also be computed:

$$\frac{1}{D} \left\langle \sum_{t,t'=1}^{T} y_t y_{t'} \mathbf{x}_t^\intercal \mathbf{x}_{t'} \prod_{s=1}^{T} \theta(\nu_s \lambda_s) \right\rangle = \frac{1}{D} \left\langle \left(\sum_{t=1}^{T} \mathbf{x}_t^\intercal \mathbf{x}_{t'} + 2 \sum_{t=2}^{T} \sum_{t'=1}^{t-1} y_t y_{t'} \mathbf{x}_t^\intercal \mathbf{x}_{t'} \right) \prod_{s=1}^{T} \theta(\nu_s \lambda_s) \right\rangle \tag{28}$$

$$= T + \mathcal{O}(1/D) \tag{29}$$

where the 2nd term can be neglected in the high dimensional limit. Substituting these computed averages into equations 15 and 16, the ODEs for the order parameters can be written:

$$\frac{dR}{d\alpha} = \frac{\eta_1 + \eta_2}{\sqrt{2\pi}} \left(1 + \frac{R}{\sqrt{Q}}\right) P^{T-1} - \eta_2 R \sqrt{\frac{2}{\pi Q}} \tag{30}$$

$$\frac{dQ}{d\alpha} = (\eta_1 + \eta_2)\sqrt{\frac{2Q}{\pi}} \left(1 + \frac{R}{\sqrt{Q}}\right) P^{T-1} - 2\eta_2 \sqrt{\frac{2Q}{\pi}} + \frac{(\eta_1^2 - \eta_2^2)}{T} P^T + \frac{\eta_2^2}{T} \tag{31}$$

Equivalence of state formulation

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

$$\langle \cdot \rangle = \sum_{\tau} \mathrm{P}(\tau) \langle \cdot \mid \tau \rangle \tag{32}$$

By considering symmetry of the Gaussian and 'half-Gaussian' ($\mathcal{N}_\pm$) distributions, all expectations in 17 can be seen to be identical regardless of whether expectations are taken with respect to the full Gaussian or the half-Gaussian distributions, i.e.

$$\langle \cdot \rangle_{\mathcal{N}} = \langle \cdot \rangle_{\mathcal{N}_+} = \langle \cdot \rangle_{\mathcal{N}_-} \tag{33}$$

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

Other Reward structures

The expectations can be calculated in other conditions of Φ from considering combinatorical arguments. We state the ODEs for two reward conditions.

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

$$\frac{dR}{d\alpha} = \frac{\eta_1}{T\sqrt{2\pi}} \sum_{i=n}^{T} \binom{T}{i} \left[i \left(1 + \frac{R}{\sqrt{Q}}\right)(1-P) - (T-i)\left(1 - \frac{R}{\sqrt{Q}}\right) P \right] P^{i-1} (1-P)^{T-i-1} \tag{34}$$

$$\frac{dQ}{d\alpha} = \frac{\eta_1}{T}\sqrt{\frac{2Q}{\pi}} \sum_{i=n}^{T} \binom{T}{i} \left[i \left(1 + \frac{R}{\sqrt{Q}}\right)(1-P) - (T-i)\left(1 - \frac{R}{\sqrt{Q}}\right) P \right] P^{i-1} (1-P)^{T-i-1} + \frac{\eta_1^2}{T} \sum_{i=n}^{T} \binom{T}{i} P^i (1-P)^{T-i} \tag{35}$$

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

$$\frac{dR}{d\alpha} = \frac{1}{\sqrt{2\pi}}\left(1+\frac{R}{\sqrt{Q}}\right)\left(\eta_1 P^{T-1}+\beta\right)+\beta(T-1)\sqrt{\frac{2}{\pi}}\frac{R}{\sqrt{Q}}P \tag{36}$$

$$\begin{aligned}\frac{dQ}{d\alpha} &= \sqrt{\frac{2Q}{\pi}}\left(1+\frac{R}{\sqrt{Q}}\right)\left(\eta_1 P^{T-1}+\beta\right)+2\beta(T-1)\sqrt{\frac{2}{\pi}}P \\ &+\left(\frac{\eta_1^2}{T}+2\eta_1^2\beta^2\right)P^T+\frac{\beta^2}{T^2}\left((1+(T-1))T^2\right)P^2\end{aligned} \tag{37}$$

B ALTERNATIVE DERIVATION OF EQ.3

$$\nabla_w \log \pi(y|x) = \nabla_w \log \frac{1}{1+e^{-yw\cdot x}} \tag{38}$$

$$\approx -\nabla_w \log e^{-yw\cdot x} = yx \tag{39}$$

where the approximation is true in the early phases of learning, when the student is still trying to learn the rule ($w \cdot x \approx 0$). Later phases are not fully captured anymore by this approximation but our focus is mostly about the first part where an interesting dynamics occurs.

C TIME FOR LEARNING TO BEGIN

In Fig. 2c, the times ascertained for learning to begin are computed by specifying the the first timestep where accuracy ($R/\sqrt{Q}$) is monotonic increasing. The ODE boundary is computed by by finding the line separating the regions where ODE accuracy increases monotonic from initialisation and where ODE accuracy is unchanging.