# A dynamical systems theory of reward-modulated learning in linear recurrent networks

**Dimitra Maoutsa***

dimitra.maoutsa@gmail.com

## Abstract

We study how recurrence shapes sparse-reward learning in linear recurrent networks trained with a linearised REINFORCE rule in a high-dimensional teacher–student setting. We show that learning is governed by the temporal responses of the student and teacher, which determine the joint Gaussian statistics of their trajectories. Recurrence introduces temporal correlations, making the probability of receiving a reward a Gaussian persistence problem. We characterise its exponential decay with episode length through the dominant eigenvalue of a transfer operator, and show that recurrence can either increase or decrease reward availability depending on the alignment of temporal responses. These changes in trajectory statistics can accelerate or delay escape from sparse-reward plateaus, while in a reward-limited regime plateau-escape times grow exponentially with episode length. Our results provide a quantitative theory of how recurrent connectivity and sparse reward interact during learning.

## 1 Introduction

Recurrent policies are the standard solution when actions must depend on memory, like, in partially observable environments, a recurrent state lets an agent integrate evidence across time [White et al., 2004, Ganguli et al., 2008, Bakker, 2001, Wierstra et al., 2007]. This suggests a simple intuition that recurrence should help reinforcement learning since a temporal kernel gives the policy a head start on the task-relevant memory trace. Here, we show that this intuition is not always correct. A recurrent kernel aligned with the task's temporal structure can shorten the sparse-reward plateau exponentially in the episode length, while a kernel not aligned with the task-required direction can make learning exponentially slower.

The difficulty is that recurrence changes the object that controls reward. In a feedforward policy, decisions at different episode times are conditionally independent given the weights, and the probability of a fully correct episode factorizes into one-step agreement probabilities. In a recurrent policy, the network dynamics correlate pre-activations across time, so the reward probability is no longer a product of independent events but a trajectory-level persistence probability [Slepian, 1962, Newell and Rosenblatt, 1962, Bray et al., 2013]: the probability that a correlated Gaussian process remains in the student–teacher agreement orthants for the whole episode. This makes sparse-reward recurrent learning a problem of **temporal-filter alignment**, extending the vector-alignment perspective of its feedforward counterpart [Patel et al., 2025].

Statistical-mechanics analyses of online learning provide macroscopic dynamics for high-dimensional supervised learners [Heskes and Kappen, 1991, Engel, 2001, Saad and Solla, 1995, Goldt et al., 2019, Bös and Opper, 1996, Kinzel and Opper, 1991], and recent work has extended this program to feedforward reinforcement learning, showing that sparse rewards generate exponentially long plateaus in high-dimensional policy learning [Patel et al., 2025]. In parallel, analyses of supervised linear and recurrent networks show how learning decomposes into low-dimensional mode-wise dynamics,

*https://dimitra-maoutsa.github.io

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

revealing how initialisation, singular modes, and recurrent eigenvalues shape learning [Saxe et al., 2013, 2019, Hardt et al., 2018, Li et al., 2022, Proca et al., 2025, Bordelon et al., 2025, Braun et al., 2022, J Dominé et al., 2023, Kunin et al., 2024]. The learning rule itself reshapes these dynamics and the resulting representations [Bordelon and Pehlevan, 2023, Williams et al., 2026], and in low-rank recurrent networks learning unfolds in a few overlaps, some of them invisible to the loss [Ger and Barak, 2026]. Recent sequence-modeling work further emphasises that linear or diagonal recurrent modes are not merely analytical conveniences, since carefully designed linear recurrent blocks can compete with state-space models on long-range tasks, and linear recurrences followed by nonlinear projections are highly expressive sequence models [Orvieto et al., 2023b,a]. Yet, these theories largely consider supervised objectives, or memoryless policies, and we still lack a quantitative theory of how sparse reward shapes the acquisition of recurrent temporal modes.

Here we develop a theory of reward-modulated learning in linear recurrent networks using a teacher–student framework. Both networks receive the same independent Gaussian observations, but integrate past inputs through potentially different temporal responses. The student produces binary decisions and receives a reward only when its entire sequence of decisions matches that of the teacher. Our approach combines a characterisation of network activity with an approximate description of how connectivity changes during learning, allowing us to distinguish the probability of receiving a reward from the learning progress made when a reward occurs. We summarise our main contributions below.

**Exact trajectory statistics and learning moments.** We describe the student and teacher outputs in terms of their (impulse) response vectors, whose overlaps determine the joint Gaussian statistics of their trajectories. This allows us to express the probability of receiving a reward as the probability of agreement over an entire trajectory. By also accounting for correlations between network outputs and their derivatives with respect to the weights, we obtain the population learning field from Gaussian trajectory moments.

**Projected master equations with a shared left subspace.** We show that the updates of the input and recurrent weights share a set of left directions determined by the readout and recurrent dynamics. This observation motivates a reduced connectivity representation with evolving diagonal coefficients and a shared adaptive left subspace that captures collective interactions. By projecting the population learning field onto the weight changes admissible within this representation, we obtain a closed system of ordinary differential equations for the reduced variables. We compare these reduced dynamics (the **projected shared-subspace model**) with full network learning to assess their accuracy.

**Persistence and escape from sparse-reward plateaus.** We characterise the persistence of agreement between student and teacher through a killed transition Markov operator, which determines the long-horizon decay of the probability of success under suitable spectral conditions. In the weak-recurrence regime, the leading correction to persistence can be either positive or negative, depending on the temporal structure of the networks. For our considered setting, matched temporal responses increase persistence, while opposed responses decrease it. We then relate these differences in reward frequency to the time required to escape learning plateaus, accounting separately for the updates produced by successful episodes and retaining the dependence on configuration-specific prefactors. Finally, we show how a recurrent teacher can induce a nonzero update of recurrent connectivity even when the student is initially feedforward.

## 2 Setting

We study learning in recurrent neural networks performing sequential decision-making tasks informed by episode-level rewards. Within each learning episode of length $T$, the network makes $T$ binary choices. At every timestep $t \in \{1, \dots, T\}$, it receives a **high-dimensional observation** $\mathbf{x}_t \in \mathbb{R}^{N_x}$, drawn independently across time and episodes from $\mathcal{N}(0, \mathrm{I}_{N_x})$, and provides a **binary output (action)** $y_t \in \{-1, +1\}$. The distribution of observations does not depend on the network's actions.

A fixed recurrent teacher network provides the target decisions (actions) at each timestep, and a second student network learns the task without having direct access to the teacher's actions. Instead, it learns from a scalar reward delivered at the end of each episode depending on whether their actions within the episode agreed with those of the teacher network.

### 2.1 The RL-$\ell$RNN model in a teacher-student setup

**Student.** We consider $\ell$inear RNNs ($\ell$RNNs) construed as **policy networks** in a reinforcement learning setting within a **teacher-student setup**. We parametrise the **student network** in terms

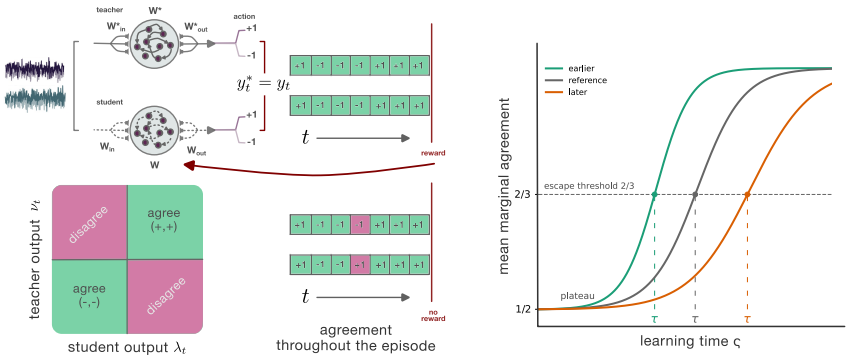



Figure 1: **Sparse reward turns recurrent learning into a trajectory-agreement problem.** The student and teacher receive the same input stream, but the student is rewarded only if their actions agree at every step of the episode. Temporal correlations induced by recurrent connections change the probability of obtaining reward even at comparable marginal accuracy. During learning, this modifies the duration of the sparse-reward plateau with connectivity structure that favours persistent agreement leading to earlier escape, while unfavorable correlations delaying it.

of three weight matrices: input $W_x \in \mathbb{R}^{N_h \times N_x}$, recurrent $W_h \in \mathbb{R}^{N_h \times N_h}$, and output weights $W_y \in \mathbb{R}^{N_y \times N_h}$. Its recurrent state $\mathbf{h}_t$ evolves within each training episode $\mu$ according to

**student**
$$\mathbf{h}_{t+1}^{(\mu)} = W_h \mathbf{h}_t^{(\mu)} + W_x \mathbf{x}_t^{(\mu)} = \sum_{s=1}^{t} W_h^{t-s} W_x \mathbf{x}_s^{(\mu)}, \qquad t \geq 1, \qquad \mathbf{h}_1^{(\mu)} = 0, \tag{1}$$

where in the last equation we have unrolled the state $\mathbf{h}_t$ in time. The student network produces a **binary action** at every timestep $t$ denoted by

$$y_t^{(\mu)} = \mathrm{sgn}(\lambda_t^{(\mu)}) = \mathrm{sgn}(W_y \, \mathbf{h}_{t+1}^{(\mu)}) = \mathrm{sgn}\left( W_y \sum_{s=1}^{t} W_h^{t-s} W_x \mathbf{x}_s^{(\mu)} \right) \in \{-1, +1\}, \tag{2}$$

where $\lambda_t^{(\mu)}$ indicates the **student output pre-activation** at episode $\mu$. For simplicity, we focus here on the single-output case $N_y = 1$, so that $W_y \in \mathbb{R}^{1 \times N_h}$ is a row vector, and view the extension to $N_y > 1$ as straightforward. Thus, the state dynamics is linear, and the action results from thresholding the output pre-activation.

**Teacher.** The **teacher network** follows the same dynamics as the student (Eq. 1) with fixed coupling matrices $W_x^*, W_h^*, W_y^*$ throughout learning, that represent a "target policy" the student has to learn. The teacher produces, thus, outputs

**teacher**
$$y_t^{*(\mu)} = \mathrm{sgn}(\nu_t^{(\mu)}) = \mathrm{sgn}(W_y^* \, \mathbf{h}_{t+1}^{*(\mu)}) = \mathrm{sgn}\left( W_y^* \sum_{s=1}^{t} {W_h^*}^{t-s} W_x^* \mathbf{x}_s^{(\mu)} \right), \tag{3}$$

where $\nu_t$ denotes the **teacher output pre-activation**.
Both teacher and student policies are therefore deterministic functions of the observation history $\mathbf{x}_{1:t}$.

**Rewards.** The student does **not** have access to the outputs $y_t^*$ directly. Instead, at the end of each training episode, it receives a **reward** $G$ reflecting whether its output agrees with teacher's outputs $\{y_{1:T} = y_{1:T}^*\}$, $G^{(\mu)} = r_1 \mathcal{I}\big(\Phi(y_{1:T}, y_{1:T}^*)\big) - r_2 \Big(1 - \mathcal{I}\big(\Phi(y_{1:T}, y_{1:T}^*)\big)\Big)$, where $r_1 \geq 0$ stands for the reward obtained for choosing matching decisions over all time steps $[1, T]$, while $r_2 \geq 0$ is the penalty for non-matching ones, and $\mathcal{I}$ denotes the indicator function.

**Learning.** Learning here amounts to modifying the student's weights so that the student reproduces the teacher's choices from the observations $\mathbf{x}_t$ alone. The student defines a deterministic hard-sign policy $y_t = \pi_\theta(\mathbf{x}_{1:t}) = \mathrm{sgn}(\lambda_t)$, or, equivalently, a degenerate trajectory-level policy $\pi_\theta(y_{1:T} \mid \mathbf{x}_{1:T}) = \prod_{t=1}^{T} \mathcal{I}[y_t = \mathrm{sgn}(\lambda_t)]$. We use a modified form of REINFORCE learning [Williams,

1992]. For a deterministic hard-sign policy, exact differentiation of the action with respect to the parameters is singular, because $\partial_w \operatorname{sgn}(\lambda_t) = 2\delta(\lambda_t)\,\partial_w \lambda_t$, where $\delta(\cdot)$ is the Dirac delta function. Thus, exact gradients involve boundary-supported terms instead of ordinary Gaussian trajectory averages. Following the linearized REINFORCE approximation of Patel et al. [2025], we use a surrogate derivative for the nondifferentiable hard-sign unit, analogous in spirit to surrogate-gradient approaches for discrete nonlinearities [Bengio et al., 2013, Neftci et al., 2019]. Thus we replace this derivative by the derivative of the pre-activation $\lambda_t$ itself. Hence, for any student parameter $w$, we use the surrogate update equation

$$\Delta w = \frac{\eta}{T} G \sum_{t=1}^{T} y_t \frac{\partial \lambda_t}{\partial w}. \tag{4}$$

We use pre-update weights for all derivatives and the same learning rate $\eta$. We average over learning episodes at fixed weights and take the small-step limit with learning time $\varsigma = \mu\eta$ results into the population field [Benaim, 1996]

$$F(\theta) = (F_x, F_h, \mathbf{f}_c) = \frac{1}{T} \left\langle G \sum_{t=1}^{T} y_t \nabla_\theta \lambda_t \;\middle|\; \theta \right\rangle. \tag{5}$$

**Assumptions.** For the theory that follows we adopt the following assumptions
**Assumption 1 (whitened i.i.d. inputs).** All inputs are uncorrelated and whitened across all timesteps, episodes, and dimensions $\mathbf{x}_t^{(\mu)} \sim \mathcal{N}(0, \mathrm{I}_{N_x}), \qquad \langle \mathbf{x}_t^{(\mu)} \mathbf{x}_{t'}^{(\mu')\top} \rangle = \delta_{tt'} \delta_{\mu\mu'} \mathrm{I}_{N_x}$.

**Assumption 2 (linear state dynamics).** The state dynamics for the student and the teacher should follow linear evolution equaitons akin to the ones provided in Eqs. 1-3.

### 2.2 Temporal responses and Gaussian trajectories

We first characterise the episode statistics for the full system weight matrices. An observation influences later decisions through the **lag-$k$ response vectors**

$$\boldsymbol{\varphi}_k = \mathbf{c}^\top W_h^k W_x, \qquad \boldsymbol{\varphi}_k^* = (\mathbf{c}^*)^\top (W_h^*)^k W_x^*, \in \mathbb{R}^{1\times N_x} \quad k \geq 0. \tag{6}$$

By unrolling the recurrent equation results into an expression for the output pre-activation in terms of the response vectors $\lambda_t = \sum_{j=1}^{t} \boldsymbol{\varphi}_{t-j} \mathbf{x}_j$ and similarly for $\nu_t$. Thus recurrence changes which input directions affect the output and how their contributions depend on time distance between input and related output.

For fixed weight matrices, the joint teacher-student output trajectory $Z = (\lambda_1, \ldots, \lambda_T, \nu_1, \ldots, \nu_T)^\top$ is a centered Gaussian vector. Its covariance $\Sigma = \left(\begin{smallmatrix} \Lambda & M \\ M^\top & N \end{smallmatrix}\right)$ is characterised by the response overlaps

$$\Lambda_{tt'} = \sum_{j=1}^{\min(t,t')} \boldsymbol{\varphi}_{t-j} \boldsymbol{\varphi}_{t'-j}^\top, \qquad M_{tt'} = \sum_{j=1}^{\min(t,t')} \boldsymbol{\varphi}_{t-j} (\boldsymbol{\varphi}_{t'-j}^*)^\top, \qquad N_{tt'} = \sum_{j=1}^{\min(t,t')} \boldsymbol{\varphi}_{t-j}^* (\boldsymbol{\varphi}_{t'-j}^*)^\top, \tag{7}$$

with entries the student, student-teacher, and teacher covariances. The outputs at different times generally correlate because they depend on overlapping input histories, e.g., $\operatorname{Cov}(\lambda_1, \lambda_2) = \boldsymbol{\varphi}_0 \boldsymbol{\varphi}_1^\top$.

### 2.3 Structured connectivity and projected learning

**Structure of the weight updates.** The learning rule suggests a shared representation of the input and recurrent interactions. We define the **backward readout vectors** $\mathbf{d}_k = (W_h^\top)^k \mathbf{c}$. Differentiating the output gives

$$\nabla_{W_x} \lambda_t = \sum_{j=1}^{t} \mathbf{d}_{t-j} \mathbf{x}_j^\top, \qquad \nabla_{W_h} \lambda_t = \sum_{j=1}^{t} \mathbf{d}_{t-j} \mathbf{h}_j^\top. \tag{8}$$

The two gradient fields share the same left vectors $\mathbf{d}_k$, and since they depend only on the current weights, reward weighting and averaging preserve the inclusion [Antoulas, 2005]

$$\operatorname{col}[F_x \; F_h] \subseteq \mathcal{K}_T := \operatorname{span}\{\mathbf{c}, W_h^\top \mathbf{c}, \ldots, (W_h^\top)^{T-1} \mathbf{c}\}, \quad \dim \mathcal{K}_T \leq \min(T, N_h). \tag{9}$$

While this observation holds for the instantnaneous gradients, we denote that the subspace $\mathcal{K}_T$ can and does evolve over learning.

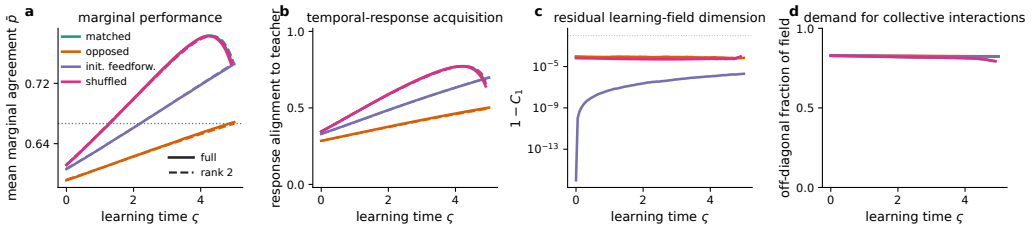



Figure 2: **Learning is low-dimensional and predominantly collective.** Learning dynamics for a fixed teacher and four student initializations: matched, opposed, initially feedforward, and shuffled recurrence with the same eigenvalues. Colors indicate initial weight configurations. **a,** Mean marginal agreement for unrestricted (solid) and rank-two projected (dashed) dynamics. The dotted line marks the escape threshold $\bar{p} = 2/3$. **b,** Finite-horizon input–output response alignment with the teacher. **c,** Fraction of matrix-field squared norm outside the leading singular direction, $1 - C_1$, where $C_1 = \sigma_1^2 / \sum_j \sigma_j^2$ for $\mathcal{F} = [F_x; F_h]$. **d,** Fraction of field squared norm in off-diagonal entries.

**Reduced connectivity.** Motivated by this observation of shared left vectors, we allow the input and recurrent interactions to share an adaptive left subspace, while keeping recurrent self-couplings and selected direct input coefficients separately adjustable. For fixed input and recurrent coordinates, we denote with $r = \min(N_h, N_x)$ the smallest input-recurrent dimension and introduce the diagonal operator $\mathcal{D}_x(\mathbf{a})$ that has entries $a_i$ at $(i, i)$ for $i \leq r$, and zero elsewhere. The orthogonal projection operators $\mathcal{P}_x$ and $\mathcal{P}_h$ (see Eq. 71) set these selected input entries and the recurrent diagonal, respectively, to zero, leaving all other entries unchanged. We use

$$W_x(S_L) = \mathcal{D}_x(\mathbf{a}) + \mathcal{P}_x(UA^\top), \qquad W_h(S_L) = \mathrm{Diag}(\mathbf{b}) + \mathcal{P}_h(UB^\top), \qquad W_y = \mathbf{c}^\top, \tag{10}$$

with $U^\top U = I_L$, $U, B \in \mathbb{R}^{N_h \times L}$, $A \in \mathbb{R}^{N_x \times L}$, $\mathbf{a} \in \mathbb{R}^r$, and $\mathbf{b}, \mathbf{c} \in \mathbb{R}^{N_h}$. The reduced state thus is $S_L = (\mathbf{a}, \mathbf{b}, \mathbf{c}, U, A, B)$, and we denote by $\Psi(S_L) = (W_x(S_L), W_h(S_L), \mathbf{c})$ the reconstructed weight matrices from this reduction, where $U$ is an orthonormal basis for the shared left subspace. We refer to this model as the **projected shared-subspace model.**

The diagonal terms of the reduced connectivity remain free to evolve during learning, while the selected input entries simply define a coordinate convention. Notice that although the unmasked products have rank at most $L$, the orthogonal projection can increase their rank, so that both the projected contributions and the full weight matrices may have higher rank. When $L \ll N_h$, this formalism reduces the number of variables to $O(N_h + L(N_h + N_x))$.

**Projected master ODEs.** We determine the reduced velocity by minimising its Frobenius discrepancy from the non-reduced gradient field evaluated at $\theta = \Psi(S_L)$, following a tangent-projection principle of dynamical model reduction [Feppon and Lermusiaux, 2018]. We obtain a deterministic description of learning by averaging the episode-wise update and taking the continuous-time limit $\eta \to 0$ with $\varsigma = \mu\eta$ [Benaim, 1996]. For the structured representation, the evolving variables are $S_L = (\mathbf{a}, \mathbf{b}, \mathbf{c}, U, A, B)$. Evaluating the population field at the reconstructed weights $\theta = \Psi(S_L)$ and projecting it onto the admissible matrix velocities gives

$$\begin{aligned} \dot{\mathbf{a}} &= \mathrm{diag}_x\, F_x(\Psi(S_L)), & \dot{\mathbf{b}} &= \mathrm{diag}\, F_h(\Psi(S_L)), \\ \dot{\mathbf{c}} &= \mathbf{f}_c(\Psi(S_L)), & (\dot{U}, \dot{A}, \dot{B}) &= \mathcal{J}_{S_L}^{\dagger}\left(\mathcal{P}_x F_x(\Psi(S_L)), \mathcal{P}_h F_h(\Psi(S_L))\right). \end{aligned} \tag{11}$$

Here dots denote derivatives with respect to learning time $\varsigma$. The linear map $\mathcal{J}_{S_L}$ converts shared-factor velocities into the associated off-diagonal matrix velocities, subject to $U^\top \dot{U} = 0$. Its pseudoinverse selects the minimum-norm solution of the coupled least-squares projection (see App. B.4.3).

These equations form an autonomous system, and the reduced state determines both the episode statistics and the next learning velocity. They retain the diagonal and readout components of the population field at the represented network, while approximating the remaining matrix updates through the adaptive factors $U$. Notice that at $L = 0$, the equations become diagonal similar to [Saxe et al., 2013, Williams et al., 2026], however the linearised update does not restrict the dynamics in this diagonal form. The population-field components separate into explicit Gaussian penalty terms and reward-weighted trajectory moments. Although the current reduced state fully specifies the network, evaluating these fields still requires averaging the reward-modulated updates over the distribution

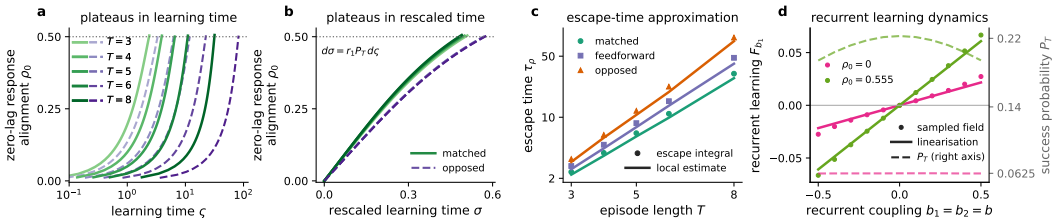



Figure 3: **Reward availability and conditional learning dynamics determine plateau escape. a,b,** Zero-lag alignment $\rho_0$ versus learning time $\varsigma$ (a) and reward-rescaled time $\sigma$ (b), for different episode lengths $T$. Solid and dashed lines indicate matched and opposed recurrence. Time rescaling reduces the horizon dependence of plateau duration. **c,** Escape times from numerical integration (symbols) and the local approximation (lines). **d,** Recurrent learning field $F_{b_1}$ near the feedforward state, along $b_1 = b_2 = b$, for two initial response alignments. Markers indicate Monte Carlo estimates, solid lines the theoretical linearisation, and dashed lines $P_T$ (right axis).

of within-episode input trajectories. In the linear Gaussian setting, we express these expectations in terms of the response vectors of the reconstructed network and the resulting joint trajectory statistics.

**Learning observables.** We evaluate the performance of the network from the response vectors. The **mean decision accuracy** and **zero-lag response vector alignment** are

$$\bar{p} = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{1}{2} + \frac{1}{\pi}\arcsin\left(\frac{M_{tt}}{\sqrt{\Lambda_{tt}N_{tt}}}\right)\right], \qquad \rho_0 = \frac{\boldsymbol{\varphi}_0(\boldsymbol{\varphi}_0^*)^\top}{\|\boldsymbol{\varphi}_0\|\,\|\boldsymbol{\varphi}_0^*\|}. \tag{12}$$

The subscript in $\rho_0$ denotes time-lag within an epsiode. We quantify the **plateau escape** by the first learning time at which the mean decision accuracy reaches the fixed threshold $\bar{p}_{\rm th} = 2/3$, i.e. $\tau_{\rm esc} = \inf\left\{\varsigma \geq 0 : \bar{p}(\varsigma) \geq \frac{2}{3}\right\}$. We apply this criterion both to the full network and reduced learning trajectories. We further keep track of the **episode-success probability** $P_T = \langle \mathcal{I}_T \mid \theta \rangle$ as the expectation of episode success indicator $\mathcal{I}_T$ (Eq. 54), which measures the probability of agreement at every step. Unlike the mean decision accuracy, $P_T$ depends on correlations between decisions at different times. Its evaluation, together with the reward-weighted moments entering the learning field, therefore requires the joint teacher-student Gaussian trajectory statistics.

### 2.4 Trajectory orthant integrals

The projected equations defined above describe how the population learning field evolves the reduced system state. To evaluate this field at the current network state, we must account for the entire output network trajectory, since sparse reward requires agreement throughout the entire episode. In the joint teacher-student output trajectory space this event is $\mathcal{A}_T = \{z \in \mathbb{R}^{2T} : z_t z_{T+t} > 0 \text{ for every } t\}$. It is a union of $2^T$ orthants, one for each common action sequence. Since the joint output trajectory is Gaussian, the probability mass of this region gives the episode-success probability. We denote by $\gamma_\Sigma$ the centered Gaussian measure with covariance as in Eq. 7, and we define [Tallis, 1961]

$$P_T = \int_{\mathcal{A}_T} \mathrm{d}\gamma_\Sigma(z), \qquad \mathbf{v}_t(\Sigma) = \int_{\mathcal{A}_T} z\ \mathrm{sgn}(z_t)\,\mathrm{d}\gamma_\Sigma(z). \tag{13}$$

The first integral quantifies the **reward availability**, while the second is an unnormalised, action-weighted first moment over successful trajectories. The factor $\mathrm{sgn}(z_t)$ comes directly from the learning rule and accounts for the sign of the student's decision at time $t$. This prevents contributions from successful trajectories related by a simultaneous teacher-student sign reversal from cancelling, as they would in an ordinary first moment. This quantity therefore captures how successful episodes contribute to the population learning field.

For a scalar student weight $w$, we define the **output pre-activation derivative** $e_{t,w} = \partial_w \lambda_t$ and the covariance between the pre-activation-derivative and joint teacher-student trajectory $C_{t,w} = \mathrm{Cov}(e_{t,w}, Z \mid \theta)$. The latter quantifies how changes in the current output induced by perturbing the weight $w$ correlate with the full teacher-student trajectory. We determine both from the current weights before projection. Joint Gaussianity implies $\langle e_{t,w} \mid Z, \theta \rangle = C_{t,w}\Sigma^\dagger Z$ [Marsaglia, 1964] and hence the learning field becomes

$$F_w = \frac{1}{T}\sum_{t=1}^{T}\left[(r_1 + r_2)C_{t,w}\Sigma^\dagger \mathbf{v}_t(\Sigma) - r_2\sqrt{\frac{2}{\pi}}\frac{(C_{t,w})_t}{\sqrt{\Lambda_{tt}}}\right]. \tag{14}$$

The pseudoinverse allows for singular trajectory covariances to occur. Thus we obtain every population-field component from $2T$-dimensional trajectory moments and the cross-covariances $C_{t,w}$

(see App. B.4.4). While the covariance alone suffices to determine reward statistics, the derivatives are required to determine the weight updates.

When $W_h = W_h^* = 0$, the output pairs are uncorrelated across time and $P_T = \mathfrak{p}^T$, with $\mathfrak{p} = 1/2 + \arcsin(\rho_0)/\pi$. With recurrent connections, the shared input histories generally prevent this factorisation, and therefore we study in the following the ensuing persistence of agreement of the output pre-activation.

### 2.5 Persistence of agreement

The Gaussian trajectory formulation distinguishes the accuracy of individual choices from the probability of a fully correct episode. Recurrence affects both, since it changes the teacher-student correlation at each time and introduces dependence between agreement events at different times. Eq. 13 provides the success probability for a fixed episode length $T$, but it does not expose how this probability changes as longer sequences of correct decisions are required, i.e. for increasing $T$. To isolate this dependence, we fix the network weights and examine how the statistics of teacher-student agreement determine the probability of success over increasingly long episodes. Persistence in correlated Gaussian processes has a long history in probability theory and statistical physics [Slepian, 1962, Newell and Rosenblatt, 1962, Bray et al., 2013]. In our setting, the Gaussian process arises from the recurrent dynamics of the student and teacher, and persistence corresponds to maintaining agreement between their output signs over time.

**A transfer-operator description.** The joint teacher-student recurrent state is a Gaussian Markov process. Restricting this process to states with matching output signs defines a killed transition operator $\mathcal{T}$ [Nardo, 2005, Azzimonti and Ginsbourger, 2018]. With $\chi$ the agreement indicator and $\pi_1$ the joint state law after the first input, the success probability is $P_T = \langle \pi_1, \mathcal{T}^{T-1}\chi \rangle$.

We formulate the operator on the reachable state space because the shared inputs can make the Gaussian transition covariance singular. Under stability and the spectral conditions specified in Appendix C, an isolated simple dominant eigenvalue $\mu$ gives

$$P_T = C\mu^{T-1}[1 + o(1)], \qquad T \to \infty. \tag{15}$$

Here $C$ depends on the network configuration. Under these conditions, the normalised distribution of surviving trajectories approaches a stable long-horizon profile, while their total probability mass decreases by a factor $\mu$ with each additional time step. Thus, $\mu$ is the asymptotic conditional probability of maintaining agreement for one additional step. In the feedforward limit, $\mu = \mathfrak{p}$ and $C = \mathfrak{p}$, recovering $P_T = \mathfrak{p}^T$. At finite episode length, however, the ratio $P_T/P_{T-1}$ is a conditional agreement probability and should not be conflated with the asymptotic rate $\mu$.

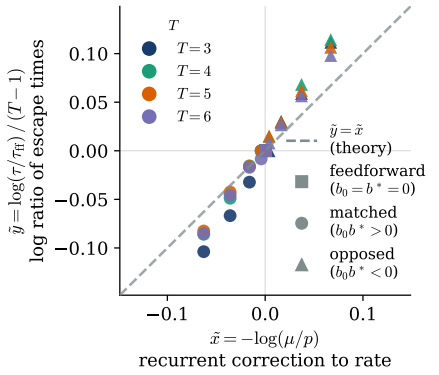



Figure 4: **Ratio-normalised escape times collapse onto the persistence-rate prediction.** Each point is an escape time for a different episode length, recurrent magnitude, and matched/opposed configuration. The horizontal coordinate is the theoretical recurrent rate correction, the vertical coordinate is the measured normalized escape-time change.

**Weak recurrence.** To isolate the first effect of recurrent connectivity, we set $W_h = \epsilon \overline{W}_h$ and $W_h^* = \epsilon \overline{W}_h^*$, keeping input weights and readouts fixed, with nonzero lag-zero responses and $|\rho_0| < 1$. Simultaneously reversing both recurrent matrices leaves every finite-horizon agreement probability unchanged, so odd powers of $\epsilon$ vanish. The leading correction to the success probability is

$$P_T = \mathfrak{p}^T + (T-1)\mathfrak{p}^{T-2}\mathcal{J} + O_T(\epsilon^4), \qquad \mathfrak{p} = \frac{1}{2} + \frac{\arcsin \rho_0}{\pi}, \tag{16}$$

where we determine $\mathcal{J} = O(\epsilon^2)$ by the zero-lag and lag-one response overlaps (Appendix B.7). It includes both the change in same-time student-teacher correlation and the dependence between adjacent agreement events. The remainder in Eq. (16) refers to a fixed episode length $T$. When the dominant rate has a regular weak-recurrence expansion, its quadratic term is $\mu = \mathfrak{p} + \mathcal{J}/\mathfrak{p} + O(\epsilon^4)$.

**The sign of the correction is determined by the temporal responses.** The correction isn't always positive, and recurrent eigenvalues alone do not determine its sign. A special case is $W_h = \mathbf{b} I_{N_h}$, $W_h^* = \mathbf{b}^* I_{N_h}$, and $\rho_0 = 0$, for which

$$\mathcal{J} = \frac{\mathbf{b}\mathbf{b}^*}{\pi^2}, \qquad \mu = \frac{1}{2} + \frac{2\mathbf{b}\mathbf{b}^*}{\pi^2} + O\big((|\mathbf{b}| + |\mathbf{b}^*|)^4\big)\,. \tag{17}$$

Small same-sign recurrence increases agreement persistence in this case,

## 3 Consequences for learning - Phenomenological predictions

The persistence calculation determines how often a given network receives a reward. To relate this probability to the time needed to escape a learning plateau, we must also account for the changes in connectivity produced by successful episodes and how these changes evolve during learning.

**Reward availability and conditional progress.** For a pure success reward setting ($r_2 = 0$), we define the expected update conditional on success $V_T(\theta) = \frac{1}{T}\left\langle \sum_{t=1}^{T} y_t \nabla_\theta \lambda_t \,\middle|\, \mathcal{I}_T = 1, \theta \right\rangle$. The population learning field then factorises as $F(\theta) = r_1 P_T(\theta) V_T(\theta)$ whenever $P_T > 0$. At a regular system state, the corresponding projected weight velocity is $r_1 P_T \Pi_{S_L} V_T$, where $\Pi_{S_L}$ denotes the orthogonal projection onto admissible weight velocities. Thus, the probability of success sets the frequency of learning updates, without determinining their direction or the progress they produce. When failure is also penalised, an additional term not multiplied by $P_T$ enters the learning field, and this factorisation no longer holds.

**Separating reward scarcity from conditional learning.** For pure success reward setting and $P_T > 0$, we introduce the **rescaled learning time** $\mathrm{d}\sigma = r_1 P_T(\theta)\,\mathrm{d}\varsigma$. The population dynamics then become $\mathrm{d}\theta/\mathrm{d}\sigma = V_T(\theta)$, with $V_T$ replaced by $\Pi_{S_L} V_T$ for the projected model. At fixed episode length, this transformation preserves the learning path while removing the multiplicative effect of reward availability. If the system reaches the escape threshold at rescaled time $\sigma_e$, the original escape time is $\tau_{\mathrm{esc}} = \frac{1}{r_1}\int_0^{\sigma_e} \frac{\mathrm{d}\sigma}{P_T(\theta(\sigma))}$. Thus persistence determines the rate at which the path is traversed, whereas the conditional dynamics determine the path and whether the threshold is reached at the end.

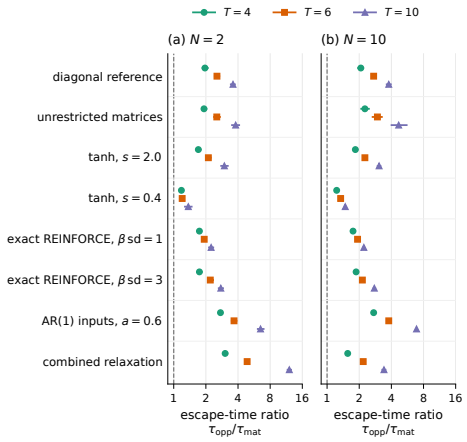



Figure 5: **Robustness of the matched–opposed escape-time asymmetry.** Ratio of opposed to matched escape times, $\tau_{\mathrm{opp}}/\tau_{\mathrm{mat}}$, for episode lengths $T = 4, 6, 10$ and network sizes $N_h = 2, 10$ across successive relaxations of the model assumptions. The matched–opposed asymmetry persists for all considered relaxations. The dashed line indicates equal escape times. The effect size varies across conditions, but remains consistently above unity and is amplified with episode length.

**Reward-limited plateau scaling.** We call learning **reward-limited** when the scarcity of successful episodes is the dominant source of dependence of the escape time on episode length. We assume that the relevant temporal responses and conditional learning progress change slowly over the plateau, and that the latter permits escape without introducing a competing exponential dependence on episode length (see App. D.2). In this setting, the persistence rate suggests the approximation $\tau_T \simeq A_T \mu_{\mathrm{pl}}^{-(T-1)}$, Here, we evaluate $\mu_{\mathrm{pl}}$ at the approximately fixed plateau configuration, while $A_T$ accounts for the boundary factor $C$, the conditional learning progress, and the criterion used to define plateau escape. Importantly, $A_T$ is different from $C$ in Eq. 15.

In the weak-recurrence setting, $-\log(\mu_{\mathrm{pl}}/\mathfrak{p}) = -\mathcal{J}/\mathfrak{p}^2 + O(\epsilon^4)$. Differences in response vectors can therefore have an increasingly large effect on escape times as the episode length grows, provided learning remains reward-limited.

**Learning recurrence from a feedforward state.** Recurrent connectivity can begin to develop even when the student is initially feedforward. Indeed, at $W_h = 0$, the derivatives satisfy $\nabla_{W_h}\lambda_t = \mathbf{c}(W_x \mathbf{x}_{t-1})^\top$ for $t \geq 2$, resulting in $F_h\big|_{W_h=0} = \frac{1}{T}\sum_{t=2}^{T} \mathbf{c}\left\langle G y_t (W_x \mathbf{x}_{t-1})^\top \mid \theta \right\rangle_{W_h=0}$.

When the teacher is also feedforward, this field vanishes because successive time blocks are independent and the relevant expectations cancel by sign symmetry. For a recurrent teacher, however, reward can depend on previous inputs, so this cancellation does not necessarily occur. The resulting update can introduce recurrent connectivity into the student, with the reduced model retaining its projection onto the admissible directions. This provides a mechanism through which temporal dependence can emerge during learning, although it does not guarantee an increase in reward or convergence towards the teacher's recurrent connectivity.

**Response alignment can accelerate or delay escape.** The effect of recurrent connectivity on persistence depends on the temporal responses of the student and teacher. In particular, the quadratic correction $\mathcal{J}$ can be positive or negative. Within the weak-recurrence regime, its dominant contribution to the exponential learning-time rate is $-\log(\mu_{\mathrm{pl}}/\mathfrak{p}) = -\mathcal{J}/\mathfrak{p}^2 + o(\epsilon^2)$, where $\mathfrak{p}$ is the feedforward agreement probability at the same lag-zero response alignment.

This dependence can be seen explicitly in the homogeneous reference family $W_h = \mathbf{b} I_{N_h}$, $W_h^* = \mathbf{b}^* I_{N_h}$, with $\rho_0 = 0$, for which $\mathcal{J} = \mathbf{b}\mathbf{b}^*/\pi^2$. When recurrent connectivity is weak, matching signs increase persistence, whereas opposite signs decrease it. Under the reward-limited approximation, these changes favour shorter and longer learning plateaus, respectively.

# 4 Results

To assess how many interaction directions are needed to capture learning within the projected shared-subspace model, we compared projected models of increasing rank $L$ with the unrestricted full population dynamics. The diagonal model corresponds to $L = 0$, while $L = 1, 2$ introduce additional directions for off-diagonal interactions. We assessed these approximations by measuring the learning field left outside the projected tangent space, the error in the input–output impulse responses, and the ratio of plateau-escape times between opposed and matched recurrent connectivities. In the $N_h = 4$ and $N_h = 8$ networks considered here, the diagonal model cannot capture most of the learning field outside its admissible directions and fails to escape the plateau from the selected initial conditions. Remarkably, introducing a single shared factor is already sufficient to capture most of the correction relevant for learning, reproducing the escape-time ratio of the full network within approximately $1\%$. Adding a second factor further improves the approximation of both the learning field and the impulse responses, but has little effect on the predicted escape-time ratio (Fig. S1). These results suggest that, in the regimes studied here, the main deviations from diagonal learning are concentrated along a small number of collective directions, although the number of directions required may differ in other settings and more realistic learning scenarios.

We next asked whether the accuracy of the projected models reflects a low-dimensional structure of the learning dynamics themselves, rather than simply providing a good approximation of the learned weights. To isolate the role of temporal alignment, we kept the teacher fixed and considered four different student initialisations: matched recurrence, sign-reversed recurrence, no initial recurrence, and a shuffled assignment of the same recurrent eigenvalues across recurrent coordinates (see details in App. A.2). The last condition preserves the recurrent spectrum but changes how its modes align with task-relevant directions. Across all four conditions, the rank-two projected dynamics closely reproduced the unrestricted network in terms of marginal agreement (Fig. 2a.), episode-success probability (Fig. S2a.), and finite-horizon input–output response alignment (Fig. 2b.). To understand why such a low-dimensional representation is sufficient, we examined the structure of the population learning field. Writing the joint matrix field as $\mathcal{F} = [F_x \; F_h]$, we characterised the fraction of the squared Frobenius norm of the instantaneous learning field captured by its first $L$ singular directions, $C_L = \frac{\sum_{\ell=1}^{L} \sigma_\ell^2(\mathcal{F})}{\sum_j \sigma_j^2(\mathcal{F})}$, and found that a single singular direction accounts for nearly all of its squared norm throughout learning, across all four initialisations (Fig. 2c.). Yet, approximately $80\%$ of this squared norm lies in off-diagonal entries (Fig. 2d.). The failure of the diagonal approximation therefore does not reflect a need for *many* independent learning directions. Instead, learning is dominated by a collective update that lies largely outside the diagonal subspace.

Consistent with this picture, the rank-two projection leaves only a small residual, and the reduced and unrestricted networks follow closely similar functional trajectories. This remains true even when the student starts without recurrence or with sign-reversed or shuffled recurrent connectivity. Together, these results suggest that learning beyond the diagonal approximation is largely low-dimensional, but fundamentally collective: a few shared directions capture changes in connectivity that cannot be described by independently adjusting diagonal coefficients.

Learning plateaus can arise either because rewards are rare or because successful episodes produce only small changes towards the task. To determine which mechanism dominates as episode length increases, we studied a two-dimensional diagonal system using the rescaled learning time $\sigma$ (App. D.1). For pure success reward, this rescaling removes the multiplicative effect of reward frequency, while preserving the dynamics conditioned on success. For each configuration, the strong dependence of zero-lag response alignment on episode length largely disappears in the rescaled time (Fig.3a,b). Thus, longer episodes slow learning primarily by making successful updates rarer, rather than by reducing the progress made following each success. However, matched and opposed recurrence still reach the same alignment threshold at different rescaled times, showing that the initial configuration also influences the conditional learning progress. Together, these results support a reward-limited interpretation of the plateaus in this example, while also showing that the initial configuration also affects the progress produced by successful episodes.

We next examined (Fig. 5) whether the difference in escape times between matched and opposed recurrence persists when the assumptions of the theory are relaxed. We considered full network learning trajectories recurrent connectivity, networks with nonlinear activation functions (tanh), with sampled logistic actions, networks learning exact REINFORCE updates, and networks with temporally correlated inputs. In all tested settings, escape times remained longer for opposed than for matched recurrence, although the magnitude of the difference depended on the setting and task horizon. These results suggest that the qualitative ordering extends beyond the assumptions of the analytical model, at least in the configurations we examined. For more details on the examined settings and additional results see App. A.5.

## 5 Discussion

We developed a theory of sparse-reward learning for linear recurrent networks evolving under a linearised variant of REINFORCE updates, by combining a description of network activity with a reduced description of how connectivity evolves during learning. By expressing the student and teacher trajectories in terms of their response vectors, we obtained their joint Gaussian statistics. The structure of the weight updates suggested a connectivity representation in which input and recurrent interactions share a set of adaptive factors. Projecting the learning dynamics onto this reduced representation then yielded a closed system of equations for the reduced variables, whose agreement with the full learning dynamics we tested explicitly. We called this reduced model the **projected shared-subspace model**.

Within this studied framework, our main insights are that recurrence influences reward in two ways, by shaping the agreement between student and teacher at a given time, and by introducing correlations between decisions across time steps. We characterised the persistence of these correlations through a transfer operator at fixed network connectivity, while the learning dynamics also depended on how the outputs changed with the network weights. The considered quadratic correction can either increase or decrease persistence, depending on the network configuration. We established the relative persistence of matched and opposed configurations analytically for specific families of responses. Relating this persistence to escape from learning plateaus required further assumptions, in particular that learning is reward-limited and that the conditional updates and non-exponential contributions remain controlled.

**Limitations.** Our analysis relied on linear recurrent dynamics driven by independent Gaussian observations, with no interaction from actions to subsequent inputs. This isolates the effect of recurrent temporal filtering on sparse reward, but excludes correlated sensory streams and genuine closed-loop RL. Actions are binary and generated by a deterministic hard-sign policy, for which we analyze a linearized reward-modulated surrogate rather than exact stochastic-policy REINFORCE, and the reduced connectivity provides an approximation to the full learning dynamics rather than an invariant description of them. The stationary analysis of persistence further requires stability of the reachable dynamics and suitable spectral conditions. Finally, the weak-recurrence expansions are valid only within their perturbative regime and do not necessarily remain accurate at long time horizons or when output correlations become degenerate. These theoretical limitations are different from pure numerical finite-step and sampling errors that may arise in the numerical implementation of the framework, but point to natural extensions suitable for future work. For instance, correlated observations could be incorporated through structured Gaussian input processes or Gaussian-equivalence arguments [Goldt et al., 2022].

## Acknowledgements

The author is grateful to the anonymous reviewers and area chair for their thoughtful and constructive feedback, which helped strengthen the theoretical analysis and improved the presentation of the manuscript.

Implementations for reproducing the numerical experiments of the paper can be found here: https://github.com/dimitra-maoutsa/RL-RNN.

## References

Athanasios C Antoulas. *Approximation of large-scale dynamical systems*. SIAM, 2005.

Dario Azzimonti and David Ginsbourger. Estimating orthant probabilities of high-dimensional Gaussian vectors with an application to set estimation. *Journal of Computational and Graphical Statistics*, 27(2):255–267, 2018.

Bram Bakker. Reinforcement learning with long short-term memory. *Advances in neural information processing systems*, 14, 2001.

Michel Benaim. A dynamical system approach to stochastic approximations. *SIAM Journal on Control and Optimization*, 34(2):437–472, 1996.

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv preprint arXiv:1308.3432*, 2013.

Blake Bordelon and Cengiz Pehlevan. The influence of learning rule on representation dynamics in wide neural networks. *The Eleventh International Conference on Learning Representations*, 2023.

Blake Bordelon, Jordan Cotler, Cengiz Pehlevan, and Jacob A Zavatone-Veth. Dynamics of learning to integrate in linear recurrent neural networks. *arXiv preprint arXiv:2503.18754*, 2025.

Siegfried Bös and Manfred Opper. Dynamics of training. *Advances in Neural Information Processing Systems*, 9, 1996.

Lukas Braun, Clémentine Dominé, James Fitzgerald, and Andrew Saxe. Exact learning dynamics of deep linear networks with prior knowledge. *Advances in Neural Information Processing Systems*, 35:6615–6629, 2022.

Alan J Bray, Satya N Majumdar, and Grégory Schehr. Persistence and first-passage properties in nonequilibrium systems. *Advances in Physics*, 62(3):225–361, 2013.

Nicolas Champagnat and Denis Villemonais. Exponential convergence to quasi-stationary distribution and Q-process. *Probability Theory and Related Fields*, 164(1):243–283, 2016.

Pierre Collet, Servet Martínez, and Jaime San Martín. *Quasi-stationary distributions: Markov chains, diffusions and dynamical systems*. Springer Science & Business Media, 2012.

Andreas Engel. *Statistical mechanics of learning*. Cambridge University Press, 2001.

Florian Feppon and Pierre FJ Lermusiaux. A geometric approach to dynamical model order reduction. *SIAM Journal on Matrix Analysis and Applications*, 39(1):510–538, 2018.

Surya Ganguli, Dongsung Huh, and Haim Sompolinsky. Memory traces in dynamical systems. *Proceedings of the national academy of sciences*, 105(48):18970–18975, 2008.

Yoav Ger and Omri Barak. Learning reveals invisible structure in low-rank RNNs. *arXiv preprint arXiv:2605.04115*, 2026.

Sebastian Goldt, Madhu S Advani, Andrew M Saxe, Florent Krzakala, and Lenka Zdeborová. Generalisation dynamics of online learning in over-parameterised neural networks. *arXiv preprint arXiv:1901.09085*, 2019.

Sebastian Goldt, Bruno Loureiro, Galen Reeves, Florent Krzakala, Marc Mézard, and Lenka Zdeborová. The Gaussian equivalence of generative models for learning with shallow neural networks. In *Mathematical and Scientific Machine Learning*, pages 426–471. PMLR, 2022.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

Tom M Heskes and Bert Kappen. Learning processes in neural networks. *Physical Review A*, 44(4): 2718, 1991.

Clémentine C J Dominé, Lukas Braun, James E Fitzgerald, and Andrew M Saxe. Exact learning dynamics of deep linear networks with prior knowledge. *Journal of Statistical Mechanics: Theory and Experiment*, 2023(11):114004, nov 2023. doi: 10.1088/1742-5468/ad01b8.

Rudolf Emil Kalman. Mathematical description of linear dynamical systems. *Journal of the Society for Industrial and Applied Mathematics, Series A: Control*, 1(2):152–192, 1963.

Wolfgang Kinzel and Manfred Opper. Dynamics of learning. In *Models of neural networks*, pages 149–171. Springer, 1991.

Daniel Kunin, Allan Raventós, Clémentine Dominé, Feng Chen, David Klindt, Andrew Saxe, and Surya Ganguli. Get rich quick: exact solutions reveal how unbalanced initializations promote rapid feature learning. *Advances in Neural Information Processing Systems*, 37:81157–81203, 2024.

Zhong Li, Jiequn Han, Qianxiao Li, et al. Approximation and optimization theory for linear continuous-time recurrent neural networks. *Journal of Machine Learning Research*, 23(42):1–85, 2022.

Dimitra Maoutsa and Manfred Opper. Deterministic particle flows for constraining SDEs. *Machine Learning and the Physical Sciences, Workshop at the 35th Conference on Neural Information Processing Systems (NeurIPS), arXiv preprint arXiv:2110.13020*, 2021.

Dimitra Maoutsa and Manfred Opper. Deterministic Particle Flows for Constraining Stochastic Nonlinear Systems. *Phys. Rev. Research*, 4:043035, Oct 2022.

George Marsaglia. Conditional means and covariances of normal variables with singular covariance matrix. *Journal of the American Statistical Association*, 59(308):1203–1204, 1964.

E Di Nardo. On the connection between orthant probabilities and the first passage time problem. *Journal of Statistical Computation and Simulation*, 75(6):437–445, 2005.

Emre O Neftci, Hesham Mostafa, and Friedemann Zenke. Surrogate gradient learning in spiking neural networks: Bringing the power of gradient-based optimization to spiking neural networks. *IEEE Signal Processing Magazine*, 36(6):51–63, 2019.

Gordon Frank Newell and Murray Rosenblatt. Zero crossing probabilities for Gaussian stationary processes. *The Annals of Mathematical Statistics*, 33(4):1306–1313, 1962.

Antonio Orvieto, Soham De, Caglar Gulcehre, Razvan Pascanu, and Samuel L Smith. Universality of linear recurrences followed by non-linear projections: finite-width guarantees and benefits of complex eigenvalues. *arXiv preprint arXiv:2307.11888*, 2023a.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International conference on machine learning*, pages 26670–26698. PMLR, 2023b.

Nishil Patel, Sebastian Lee, Stefano Sarao Mannelli, Sebastian Goldt, and Andrew Saxe. RL perceptron: Generalization dynamics of policy learning in high dimensions. *Physical Review X*, 15 (2):021051, 2025.

Alexandra Maria Proca, Clémentine Carla Juliette Dominé, Murray Shanahan, and Pedro AM Mediano. Learning dynamics in linear recurrent neural networks. In *Forty-second International Conference on Machine Learning*, 2025.

David Saad and Sara A Solla. On-line learning in soft committee machines. *Physical Review E*, 52 (4):4225–4243, 1995.

Andrew M Saxe, James L McClelland, and Surya Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*, 2013.

Andrew M Saxe, James L McClelland, and Surya Ganguli. A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences*, 116(23): 11537–11546, 2019.

David Slepian. The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal*, 41 (2):463–501, 1962.

Georges M Tallis. The moment generating function of the truncated multi-normal distribution. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 23(1):223–229, 1961.

Olivia L White, Daniel D Lee, and Haim Sompolinsky. Short-term memory in orthogonal neural networks. *Physical review letters*, 92(14):148102, 2004.

Daan Wierstra, Alexander Foerster, Jan Peters, and Juergen Schmidhuber. Solving deep memory pomdps with recurrent policy gradients. In *International conference on artificial neural networks*, pages 697–706. Springer, 2007.

Ezekiel Williams, Alexandre Payeur, and Guillaume Lajoie. Dynamics and Representation Structure of Local Approximations to Gradient-Based Learning in Linear Recurrent Neural Networks. In *Forty-third International Conference on Machine Learning*, 2026.

Ronald J Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256, 1992.

# Supplementary Information


**A Additional numerical experiments** **16**

A.1 Accuracy of the reduced connectivity description . . . . . 16

A.2 Functional and structural validation of the projected shared-subspace dynamics . . . . 18

A.3 Escape-time ratios in the full and projected shared-subspace models . . . . . 22

A.4 Initial teacher-student response-vector alignment . . . . . 23

A.5 Robustness of the temporal-alignment effect . . . . . 23

A.6 A delayed-cue task requiring memory . . . . . 25

A.7 Numerical validation of the transfer operator . . . . . 26

A.8 Numerical tests of the conditional learning dynamics . . . . . 28

**B Order parameters and the reduction to low-dimensional dynamics** **35**

B.1 The general setting . . . . . 36

B.1.1 Output pre-activations as Gaussian functionals . . . . . 36

B.1.2 Covariance kernels of output pre-activation as order parameters . . . . . 37

B.1.3 Closure of the dynamics in the general formulation . . . . . 37

B.2 Model assumptions and learning rule . . . . . 38

B.2.1 Linearised reward-modulated update rule . . . . . 38

B.3 Response representation and Gaussian trajectory statistics . . . . . 39

B.4 Learning dynamics . . . . . 40

B.4.1 Weight derivatives and the population learning dynamics . . . . . 40

B.4.2 Structured connectivity and reduced variables . . . . . 41

B.4.3 Variational projection and master equations . . . . . 42

B.4.4 Gaussian evaluation of the population learning field . . . . . 43

B.4.5 Learning observables and consistency checks . . . . . 45

B.5 Invariant balances and the diagonal sign constraint . . . . . 48

B.6 Response-equivalent networks and a matrix-state transformation . . . . . 48

B.7 The quadratic correction in response coordinates . . . . . 49

**C The transfer operator construction for the persistence rate** **49**

C.1 Joint recurrent-state process and survival representation . . . . . 50

C.2 Killed transfer operator . . . . . 50

C.3 Reachable state space and stationary dynamics . . . . . 51

C.4 Persistence rate . . . . . 51

C.5 Weak recurrence and connection to the quadratic correction . . . . . 52

C.6 Numerical evaluation . . . . . 53

C.7 Non-perturbative use and relation to learning . . . . . 53

**D Dynamical structure of reward-modulated learning** **54**

D.1 Conditional learning paths and sucess-dependent time rescaling . . . . . 54

D.2 Escape as an integral of the rescaled time . . . . . 55
D.3 The linearised REINFORCE update increases the rewarded pre-activation magnitude . 56
D.4 Local recurrent learning at the feedforward state . . . . . 56
D.5 A single-unit example: alignment, memory, and gain . . . . . 58

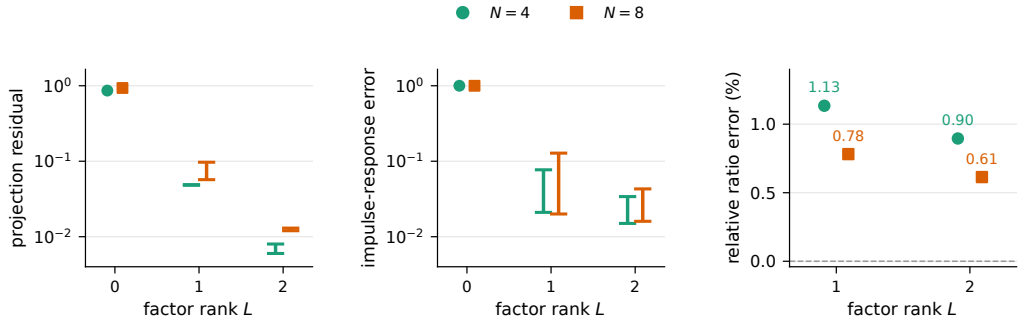



Figure S1: **A small number of collective directions captures the full learning dynamics.** Accuracy of the projected shared-subspace model for $T = 4$, initial alignment $\rho_0 = 0.35$, and network sizes $N_h = 4$ and $N_h = 8$. The diagonal model ($L = 0$) misses much of the population learning field and fails to escape the plateau from the selected initial conditions. Adding a single shared factor substantially reduces the projection residual and response vector error, while reproducing the unrestricted opposed-to-matched escape-time ratio within approximately $1\%$. Rank two further improves the learning-field and temporal-response approximations, with little change in the escape-time ratio. Reported error ranges span matched and opposed conditions and are not confidence intervals. These results establish the accuracy of the reduced hierarchy in the tested configurations.

# A Additional numerical experiments

We provide additional comparisons of the projected shared-subspace and unrestricted learning dynamics, tests of the temporal-alignment effect under changes to the model, and a delayed-cue task requiring memory. We also evaluate the Gaussian support and finite-horizon survival statistics that enter the transfer-operator description. Throughout this section, $N = N_h$ denotes the size of the recurrent network and $T$ the episode length. We write $R_\tau = \tau_{\text{opp}}/\tau_{\text{mat}}$ for the opposed-to-matched escape-time ratio.

## A.1 Accuracy of the reduced connectivity description

**Dependence on the retained rank.** Table 1 compares the projected shared-subspace model with one- and two-rank interactions against the full network dynamics at $T = 4$ and initial zero-lag alignment $\rho_0 = 0.35$, for $N_h = 4$ and $N_h = 8$.

The projection residual $\epsilon_{\text{proj}}$ quantifies the fraction of the full learning vector field that lies outside the tangent space of the rank-$L$ manifold, while the response vector error quantifies the discrepancy in the task-relevant input-output dynamics. They are computed as

$$\epsilon_{\text{proj}} = \frac{\left\| \dot{\theta} - \Pi_{T_\theta} \dot{W} \right\|_2}{\left\| \dot{W} \right\|_2}, \qquad \epsilon_{\text{VR}} = \left( \frac{\sum_{k=0}^{T-1} |\boldsymbol{\varphi}_k - \hat{\boldsymbol{\varphi}}_k|^2}{\sum_{k=0}^{T-1} |\boldsymbol{\varphi}_k|^2} \right)^{1/2}, \tag{18}$$

where the norms are Frobenius norms and $\boldsymbol{\varphi}_k = W_y W_h^k W_x$ is the response vector, and $\Pi_{T_\theta}$ denotes the projection operator. This analysis complements the comparison of escape times by testing the learning field and temporal response directly.

In the considered settings, the zero-rank projected shared-subspace (diagonal) model has no observed escape, while the one- and two-rank models reproduce the full network escape-time ratio very closely. The relative deviations are $|R_\tau^{(L)}/R_\tau^{\text{full}} - 1|$ are $1.13\%$ and $0.90\%$ for $L = 1, 2$ at $N_h = 4$, and $0.78\%$ and $0.61\%$ at $N_h = 8$. Increasing the retained rank reduces also the projection residual and the largest response vector error for each considered network.

**Agreement of behavioral trajectories.** We compare three observables along learning. At learning time $\varsigma$, $A_t = \{y_t = y_t^*\}$ denotes teacher-student agreement at episode time $t$, with the weights $\theta(\varsigma)$

| $N_h$ | model | $\epsilon_{\mathrm{proj}}$ | $\epsilon_{\mathrm{RV}}$ | $\tau_{\mathrm{mat}}$ | $\tau_{\mathrm{opp}}$ | $\tau_{\mathrm{opp}}/\tau_{\mathrm{mat}}$ |
|---|---|---|---|---|---|---|
| 4 | $L = 0$ | 0.863 | 1.000 | *n.r.* | *n.r.* | – |
| 4 | $L = 1$ | 0.048–0.049 | 0.021–0.077 | 0.378 | 0.641 | 1.694 |
| 4 | $L = 2$ | 0.006–0.008 | 0.015–0.034 | 0.378 | 0.640 | 1.690 |
| 4 | full | – | – | 0.382 | 0.639 | 1.675 |
| 8 | $L = 0$ | 0.934 | 0.999 | *n.r.* | *n.r.* | – |
| 8 | $L = 1$ | 0.057–0.097 | 0.020–0.128 | 0.250 | 0.451 | 1.805 |
| 8 | $L = 2$ | 0.012–0.013 | 0.016–0.043 | 0.250 | 0.451 | 1.802 |
| 8 | full | – | – | 0.250 | 0.447 | 1.791 |

Table 1: Accuracy of the projected shared-subspace model for different ranks $L$ at $T = 4$ and $\rho_0 = 0.35$. The projection residual $\epsilon_{\mathrm{proj}}$ quantifies the component of the full population learning field that cannot be represented by the admissible reduced velocity of each considered model. The response vector error $\epsilon_{\mathrm{RV}}$ measures the discrepancy between the induced finite-horizon input-output response between the full model and each considered version of the projected shared-subspace model. The ranges for $\epsilon_{\mathrm{proj}}$ and $\epsilon_{\mathrm{RV}}$ span the matched and opposed conditions, and should not be interpreted as confidence intervals. The $L = 0$ entries marked as *n.r.* indicate no recorded plateau escape.

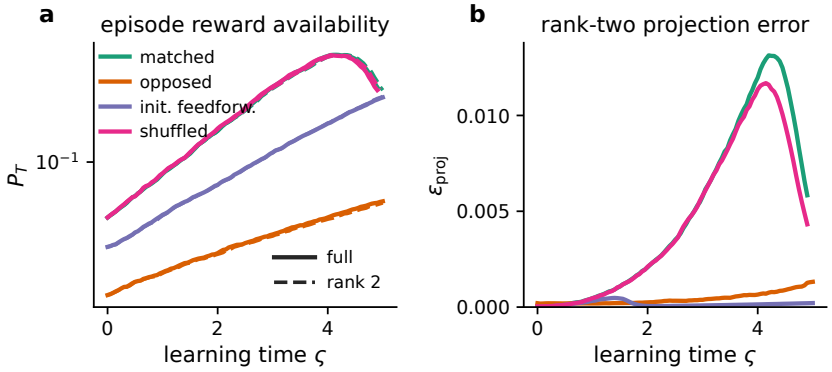



Figure S2: **Validation of projected dynamics across initializations.** The teacher and four student initializations are as in Fig. 2, with colours indicating matched, opposed, initially feedforward, and shuffled recurrence. **a,** Episode-success probability $P_T$ during learning for unrestricted (solid) and rank-two projected (dashed) dynamics, showing close agreement across initializations. **b,** Relative projection residual $\epsilon_{\mathrm{proj}}$ of the population learning field onto the rank-two projected shared-subspace model's tangent space, evaluated along the projected trajectories. The small residual indicates that the reduced representation captures the dominant instantaneous learning directions across the tested conditions.

held fixed during the episode. The **mean marginal agreement** and **episode-success probability** are

$$\bar{p}(\varsigma) = \frac{1}{T}\sum_{t=1}^{T} \Pr(A_t \mid \theta(\varsigma)), \qquad P_T(\varsigma) = \Pr\left(\bigcap_{t=1}^{T} A_t \,\middle|\, \theta(\varsigma)\right). \tag{19}$$

We further additionally evaluate the finite-horizon conditional survival probability

$$\mu_T^{\mathrm{cond}}(\varsigma) = \frac{P_T(\varsigma)}{P_{T-1}(\varsigma)}, \qquad T \geq 2, \quad P_{T-1}(\varsigma) > 0. \tag{20}$$

Both probabilities in this ratio use the same weights and episode initialization, $P_{T-1}$ identifies agreement over the first $T-1$ steps. Thereby $\mu_T^{\mathrm{cond}}$ measures the survival of the final step *conditional* on agreement at all preceding steps. This is a finite-horizon observable and should not be conflated with the asymptotic persistence eigenvalue.

Table 2 presents the trajectory root-mean-square errors (RMSEs) between the $L = 2$ projected shared-subspace model and full network learning at initial $\rho_0 = 0.4$, using eight seeds per setting. The largest episode-success RMSE among the $T = 4$ comparisons is $0.016$. The error is larger in several $T = 6$ comparisons, most clearly for the $N_h = 4$ FF condition, where the $P_T$ RMSE is $0.32$ with a $95\%$ confidence interval of $[0.12, 0.678]$.

Table 2: Trajectory RMSEs for the $L = 2$ projected shared-subspace model versus unrestricted learning, with $\rho_0 = 0.4$ and eight seeds. Brackets indicate 95% confidence intervals.

| $N_h$ | $T$ | condition | $\bar{p}$: RMSE [95% CI] | $P_T$: RMSE [95% CI] | $\mu_T^{\mathrm{cond}}$: RMSE [95% CI] |
|---|---|---|---|---|---|
| 4 | 4 | matched | 0.00111 [0.0006, 0.00169] | 0.00196 [0.00129, 0.00271] | 0.000431 [0.000294, 0.000589] |
| 4 | 4 | FF | 0.000923 [0.000511, 0.00138] | 0.00161 [0.000973, 0.00234] | 0.000325 [0.000194, 0.000497] |
| 4 | 4 | opposed | 0.00219 [0.00104, 0.00357] | 0.016 [0.005, 0.0295] | 0.00185 [0.000708, 0.00326] |
| 4 | 6 | matched | 0.00638 [0.00368, 0.00938] | 0.0641 [0.0301, 0.101] | 0.00663 [0.00363, 0.01] |
| 4 | 6 | FF | 0.0276 [0.00996, 0.0588] | 0.32 [0.12, 0.678] | 0.0272 [0.00962, 0.0578] |
| 4 | 6 | opposed | 0.00942 [0.00506, 0.0146] | 0.115 [0.0568, 0.183] | 0.0102 [0.00544, 0.0163] |
| 8 | 4 | matched | 0.00104 [0.000669, 0.0014] | 0.0014 [0.000783, 0.00204] | 0.000309 [0.000175, 0.000459] |
| 8 | 4 | FF | 0.00111 [0.000686, 0.00155] | 0.00148 [0.000723, 0.00241] | 0.000292 [0.000159, 0.000459] |
| 8 | 4 | opposed | 0.00149 [0.00078, 0.00241] | 0.00697 [0.0021, 0.0154] | 0.00111 [0.00041, 0.00229] |
| 8 | 6 | matched | 0.00122 [0.000523, 0.00229] | 0.00914 [0.00255, 0.0209] | 0.00126 [0.000366, 0.00278] |
| 8 | 6 | FF | 0.000858 [0.000536, 0.00119] | 0.0047 [0.00314, 0.0064] | 0.000809 [0.000579, 0.00112] |
| 8 | 6 | opposed | 0.00564 [0.00413, 0.00779] | 0.0685 [0.0495, 0.0963] | 0.00722 [0.0053, 0.00916] |

The finite-horizon conditional survival quantity should not be identified with the asymptotic persistence eigenvalue.

Table 3: Pearson correlations between the same full network and projected shared-subspace model behavioural trajectories.

| $N_h$ | $T$ | condition | $\bar{p}$ | $P_T$ | $\mu_T^{\mathrm{cond}}$ |
|---|---|---|---|---|---|
| 4 | 4 | matched | 1 | 1 | 1 |
| 4 | 4 | FF | 1 | 1 | 1 |
| 4 | 4 | opposed | 0.996 | 0.995 | 0.996 |
| 4 | 6 | matched | 0.977 | 0.972 | 0.972 |
| 4 | 6 | FF | 0.842 | 0.827 | 0.842 |
| 4 | 6 | opposed | 0.94 | 0.935 | 0.933 |
| 8 | 4 | matched | 1 | 1 | 1 |
| 8 | 4 | FF | 1 | 1 | 1 |
| 8 | 4 | opposed | 0.998 | 0.997 | 0.996 |
| 8 | 6 | matched | 0.998 | 0.998 | 0.996 |
| 8 | 6 | FF | 1 | 1 | 0.999 |
| 8 | 6 | opposed | 0.967 | 0.961 | 0.939 |

Table 3 provides the Pearson correlations for the same trajectory comparisons. Correlation quantifies the similarity in temporal variation and thereby complements the RMSE, which is sensitive to differences in absolute values. The weakest correlation, $0.827$, occurs for $P_T$ in the $N_h = 4$, $T = 6$ FF condition, consistent with the larger error in Table 2.

### A.2 Functional and structural validation of the projected shared-subspace dynamics

We compared the projected and unrestricted dynamics to assess both how accurately the reduced model captures learning and why a low-dimensional representation is sufficient. We first examined whether the two systems develop similar behavioural performance and finite-horizon input–output responses. We then studied the structure of the population learning field, asking whether it is dominated by a few collective matrix directions and whether these directions lie outside the diagonal subspace.

The four functional and structural quantities presented in Fig. 2 and Fig. S2 allow us to distinguish how closely the networks behave from how accurately the reduced model captures the underlying learning field.

**Numerical setting and temporal initializations.** For the comparisons in Figs. 2 and 2, we kept the recurrent teacher fixed and vary the initial recurrent connectivity of the student with the following characteristics

$$N_h = N_x = 8, \qquad T = 6, \qquad L = 2, \qquad \rho_0(0) = 0.35.$$

We considered pure success reward, with $r_1 = 1$ and $r_2 = 0$. Here, $\rho_0(0)$ denotes the initial zero-lag response alignment.

We considered four student initializations:

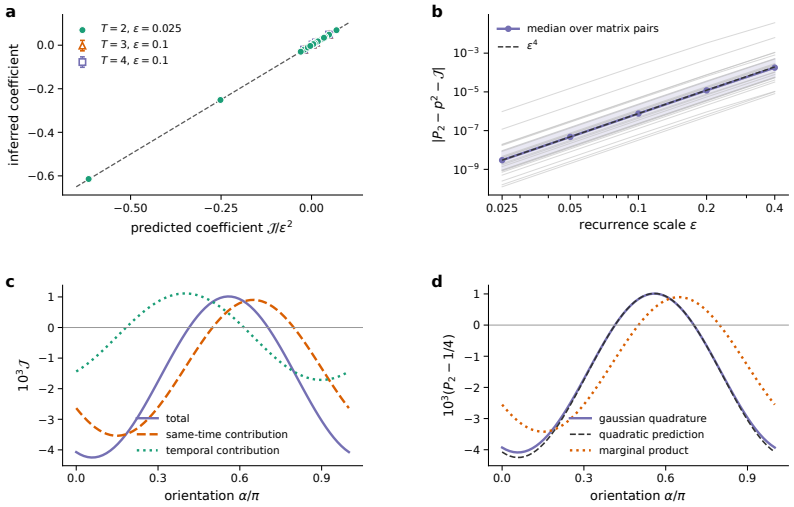



Figure S3: **The quadratic correction. a**, Predicted coefficient $\mathcal{J}/\epsilon^2$ against the inferred coefficient $(P_T - \mathfrak{p}^T)/[(T-1)\mathfrak{p}^{T-2}\epsilon^2]$. **b**, The two-step residual scales as $\epsilon^4$ in the weak-recurrence setting. Thin lines are individual matrix pairs, while the thick line and shading indicate the median and interquartile range across pairs. **c**, Decomposition of $\mathcal{J}$ into same-time and temporal contributions as a nonsymmetric recurrence is rotated relative to fixed input and readout maps. The eigenvalues, singular values, norms, and commutator nonnormality remain fixed. **d**, Exact two-step survival correction, its quadratic prediction, and the correction predicted from marginal accuracies alone.

1. **matched**: the student and teacher recurrent coefficients initially coincide;
2. **opposed**: the signs of the student recurrent coefficients are reversed relative to the teacher;
3. **initially feedforward**: the student recurrent matrix is initialized at zero while the teacher remains recurrent;
4. **shuffled**: the student retains the same recurrent eigenvalue multiset as in the matched condition, but the coefficients are permuted across recurrent coordinates.

The shuffled condition preserves the recurrent eigenvalues, but changes their alignment with task-relevant input and readout directions. Across conditions, the teacher, input and readout initialization, and initial zero-lag response alignment remain fixed.

For each initialisation, we evolved the unrestricted network and the projected $L = 2$ model independently, starting from the same connectivity. The projected dynamics were therefore not fitted to the unrestricted trajectory during learning. Colors identify the four initialisations, while solid and dashed lines show the unrestricted and projected dynamics, respectively.

**Mean marginal agreement.** At a fixed learning time $\varsigma$, the weights are held constant during an episode. We denote agreement between student and teacher at episode time $t$ by

$$A_t = \{y_t = y_t^*\}.$$

The student and teacher pre-activations are jointly Gaussian, with same-time correlation

$$\rho_t(\varsigma) = \frac{M_{tt}(\varsigma)}{\sqrt{\Lambda_{tt}(\varsigma) N_{tt}(\varsigma)}}.$$

The marginal probability of sign agreement is therefore

$$\mathfrak{p}_t(\varsigma) = \Pr(A_t \mid \theta(\varsigma)) = \frac{1}{2} + \frac{1}{\pi} \arcsin \rho_t(\varsigma), \tag{21}$$

and the mean marginal agreement is

$$\bar{p}(\varsigma) = \frac{1}{T} \sum_{t=1}^{T} p_t(\varsigma). \tag{22}$$

This observable quantifies average decision accuracy within an episode and defines the plateau-escape criterion used in the main text,

$$\tau_{\mathrm{esc}} = \inf\left\{\varsigma : \bar{p}(\varsigma) \geq \frac{2}{3}\right\}.$$

Agreement of the unrestricted and projected trajectories in $\bar{p}$ therefore tests whether the reduction reproduces the behavioral quantity used to define escape from the sparse-reward plateau.

For the final figure, we evaluate Eq. 22 from the Gaussian response covariance for both models. A direct Monte Carlo estimate from an independent evaluation batch gives an additional consistency check for the unrestricted simulation.

**Episode-success probability.** Mean marginal agreement does not determine the probability of obtaining a reward, because recurrent dynamics correlate agreement events at different episode times. The episode-success probability is

$$P_T(\varsigma) = \Pr\left(\bigcap_{t=1}^{T} A_t \,\middle|\, \theta(\varsigma)\right) = \int_{\mathcal{A}_T} d\gamma_{\Sigma(\varsigma)}(z), \tag{23}$$

where

$$\mathcal{A}_T = \left\{z \in \mathbb{R}^{2T} : z_t z_{T+t} > 0, \quad t = 1, \ldots, T\right\}$$

is the union of the student–teacher agreement orthants, and $\gamma_{\Sigma(\varsigma)}$ is the centered Gaussian measure determined by the current response kernels.

Numerically, $P_T$ is evaluated from the $2T$-dimensional Gaussian trajectory distribution using a fixed set of antithetically paired Sobol points. The same quasi-random points are reused across learning times within each run, reducing numerical fluctuations when comparing consecutive network states.

While $\bar{p}$ measures the accuracy of individual decisions, $P_T$ measures trajectory-level reward availability. Agreement between the unrestricted and projected curves in $P_T$ therefore tests whether the reduced dynamics reproduce the temporal correlations that govern sparse reward, rather than only marginal performance.

**Finite-horizon temporal-response alignment.** Behavioral agreement alone does not imply that two networks implement the same temporal input–output transformation. We therefore compare the full finite-horizon response kernels. For the student and teacher,

$$\boldsymbol{\varphi}_k = \mathbf{c}^\top W_h^k W_x, \qquad \boldsymbol{\varphi}_k^* = (\mathbf{c}^*)^\top (W_h^*)^k W_x^*, \qquad 0 \leq k < T.$$

We collect these response vectors into the weighted matrices

$$\mathcal{R}_T = \begin{pmatrix} \sqrt{T}\,\boldsymbol{\varphi}_0 \\ \sqrt{T-1}\,\boldsymbol{\varphi}_1 \\ \vdots \\ \boldsymbol{\varphi}_{T-1} \end{pmatrix}, \qquad \mathcal{R}_T^* = \begin{pmatrix} \sqrt{T}\,\boldsymbol{\varphi}_0^* \\ \sqrt{T-1}\,\boldsymbol{\varphi}_1^* \\ \vdots \\ \boldsymbol{\varphi}_{T-1}^* \end{pmatrix}. \tag{24}$$

The factor $T - k$ accounts for the number of input–output pairs at lag $k$ within an episode of length $T$. In particular,

$$\|\mathcal{R}_T\|_F^2 = \sum_{k=0}^{T-1} (T-k)\|\boldsymbol{\varphi}_k\|^2$$

is the total finite-horizon response energy under independent unit-variance inputs.

We define the temporal-response alignment by the normalized Frobenius overlap

$$\rho_{\mathrm{resp}}^{(T)}(\varsigma) = \frac{\langle \mathcal{R}_T(\varsigma), \mathcal{R}_T^* \rangle_F}{\|\mathcal{R}_T(\varsigma)\|_F \|\mathcal{R}_T^*\|_F}. \tag{25}$$

This quantity lies in $[-1, 1]$ and is invariant under a common positive rescaling of the student response. It measures similarity of the finite-horizon input–output transformation, rather than similarity of the recurrent matrices or their eigenvalues. Increasing $\rho_{\mathrm{resp}}^{(T)}$ therefore indicates acquisition of the teacher's task-relevant temporal response. Agreement of the unrestricted and projected curves shows that the reduced dynamics reproduce this functional change, rather than merely matching marginal accuracy.

**Dimension of the instantaneous matrix learning field.** To determine whether the success of the projected model reflects a low-dimensional organization of learning itself, we evaluate the unrestricted population field at the current projected state $\theta = \Psi(S_L)$. We horizontally concatenate its input and recurrent matrix components,

$$\mathcal{F}(\varsigma) = [F_x(\Psi(S_L(\varsigma))) \quad F_h(\Psi(S_L(\varsigma)))]\,. \tag{26}$$

The horizontal concatenation is essential: its left singular vectors describe receiving directions shared by the input and recurrent updates, which are the directions represented by the adaptive basis $U$.

Let

$$\sigma_1(\varsigma) \geq \sigma_2(\varsigma) \geq \cdots$$

be the singular values of $\mathcal{F}(\varsigma)$. We define the cumulative field energy captured by the first $L$ singular directions as

$$C_L(\varsigma) = \frac{\sum_{\ell=1}^{L} \sigma_\ell^2(\varsigma)}{\sum_j \sigma_j^2(\varsigma)}. \tag{27}$$

Since

$$\sum_j \sigma_j^2 = \|\mathcal{F}\|_F^2,$$

$C_L$ is the fraction of the squared matrix-field norm captured by its best rank-$L$ approximation.

The main-text figure shows

$$1 - C_1 = \frac{\sum_{j\geq 2} \sigma_j^2}{\sum_j \sigma_j^2},$$

the fraction of matrix-field energy not accounted for by the leading shared direction. This is a property of the instantaneous learning field. It is not the rank of the connectivity matrix, the rank of the accumulated weight change, or the dimension of the neural activity.

If $C_1 \simeq 1$, the two matrix fields have the approximate form

$$F_x \simeq u_1 a^\top, \qquad F_h \simeq u_1 b^\top,$$

and therefore share a dominant left direction $u_1$. The small values of $1 - C_1$ observed across the four temporal initializations show that the unrestricted matrix learning field is strongly concentrated along a collective receiving direction.

**Fraction of the learning field outside the diagonal sector.** Low dimensionality does not by itself explain why the diagonal model fails, since in principle a rank-one field could still lie entirely inside the diagonal subspace. However, this would refer to a degenerate case where the rank-one component would have only one nonzero entry. SInce the left factor in our setting is the readout vector $\mathbf{c} = W_y$, this would correspond to a network configuration where the output is connected through the readout only to a single recurrent neuron. To probe further the failure of the purely diagonal connectivty model, we measure how much of the instantaneous matrix field belongs to the interaction sector.

Let $\mathcal{P}_x$ remove the selected direct input entries and let $\mathcal{P}_h$ remove the recurrent diagonal, as in the structured connectivity representation. We define

$$\chi_{\text{off}}(\varsigma) = \frac{\|\mathcal{P}_x F_x(\Psi(S_L))\|_F^2 + \|\mathcal{P}_h F_h(\Psi(S_L))\|_F^2}{\|F_x(\Psi(S_L))\|_F^2 + \|F_h(\Psi(S_L))\|_F^2}. \tag{28}$$

For the square input matrices used in these experiments, $\mathcal{P}_x$ and $\mathcal{P}_h$ simply remove the corresponding matrix diagonals.

The quantity $\chi_{\text{off}}$ lies between zero and one, and $\chi_{\text{off}} = 0$ corresponds to a purely diagonal matrix update, while large values indicate that learning predominantly requires interactions between different coordinates. It specifically quantifies the structure of the **instantaneous population field**. This should not be conflated with the fraction of the learned connectivity that is off diagonal.

The two field validation tests answer complementary questions, $C_1$ reflects how many shared directions are required to learn the task imposed by the teacher, and $\chi_{\text{off}}$ identifies whether the dominant population field lies inside the diagonal subspace.

Table 4: Full and projected shared-subspace model escape-time ratios from the separate $T = 4$ comparison. Brackets give the 95% confidence intervals.

| $N_h$ | $T$ | $\rho_0$ | full [95% CI] | extended [95% CI] | extended/full |
|---|---|---|---|---|---|
| 4 | 4 | 0.3 | 1.45 [1.18, 1.96] | 1.46 [1.2, 2.01] | 1.01 |
| 4 | 4 | 0.4 | 2.05 [1.54, 3.65] | 2.08 [1.59, 3.81] | 1.01 |
| 8 | 4 | 0.3 | 1.55 [1.4, 1.75] | 1.54 [1.42, 1.75] | 0.998 |
| 8 | 4 | 0.4 | 2.24 [2.16, 3.06] | 2.27 [2.17, 3.04] | 1.02 |

The combination of the two

$$C_1 \simeq 1, \qquad \chi_{\text{off}} \gg 0$$

therefore indicates that learning is low-dimensional and not mode-wise. The diagonal approximation fails not because learning involves many independent directions, but because the dominant update is collective and lies largely outside the diagonal subspace.

**Tangent-space projection residual.** A low-dimensional learning field does not necessarily imply that it can be captured by the proposed diagonal-plus-factorized representation. To test this, we measure the residual obtained by projecting the population field onto the tangent space of the $L = 2$ model.

At a reduced state $S_L = (a, b, c, U, A, B)$, the admissible matrix velocities are

$$\dot{W}_x^{(L)} = D_x(\dot{a}) + \mathcal{P}_x \left( \dot{U} A^\top + U \dot{A}^\top \right),$$

$$\dot{W}_h^{(L)} = \mathrm{Diag}(\dot{b}) + \mathcal{P}_h \left( \dot{U} B^\top + U \dot{B}^\top \right).$$

The factor velocities are chosen by the coupled least-squares projection. We define the matrix residuals

$$E_x = F_x - \dot{W}_x^{(L)}, \qquad E_h = F_h - \dot{W}_h^{(L)}, \tag{29}$$

evaluated at $\theta = \Psi(S_L)$, and plot the relative matrix-field residual

$$\epsilon_{\text{proj}} = \frac{\left( \|E_x\|_F^2 + \|E_h\|_F^2 \right)^{1/2}}{\left( \|F_x\|_F^2 + \|F_h\|_F^2 \right)^{1/2}}. \tag{30}$$

The readout velocity is not included in this normalisation because $\dot{\mathbf{c}} = \mathbf{f}_c$ is represented exactly and contributes no projection residual.

These two measures capture different aspects of the learning field. While $C_1$ measures how much of the unrestricted matrix field is concentrated along a single direction, independently of any parameterization, $\epsilon_{\text{proj}}$ measures how well the tangent space of the proposed shared-subspace model captures that field. A small projection residual, together with the close agreement between independently evolved trajectories, indicates that the reduced representation captures the dominant learning directions in the configurations studied.

### A.3 Escape-time ratios in the full and projected shared-subspace models

Table 4 demonstrates a separate comparison at $T = 4$ and initial zero-lag teacher-student response alignment $\rho_0 \in \{0.3, 0.4\}$. The full and projected shared-subspace models give point estimates of the opposed-to-matched escape-time ratio in all four settings. The ratio of the projected shared-subspace estimate to the full estimate (which should optimally be 1) ranges from $0.998$ to $1.02$, which indicates that the two models result in nearly equal escape times for all settings.

The finite escape-time summaries must be read together with the observed crossing fractions. All matched runs crossed the threshold. For the opposed initial condition, the crossing fractions were $5/8$, $5/8$, $5/8$, and $7/8$ for $(N_h, \rho_0) = (4, 0.3), (4, 0.4), (8, 0.3), (8, 0.4)$, respectively, in both models.

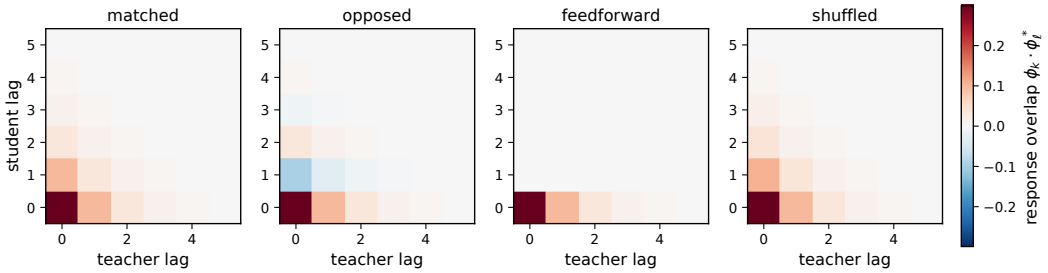



Figure S4: **Initial response vector alignment for different initial recurrent configurations.** Teacher-student response overlaps $Q_{k\ell} = \varphi_k \cdot \varphi_\ell^*$ for matched, opposed, initially feedforward, and shuffled initial recurrent connectivity. Matched and shuffled configurations show similar positive overlaps, while opposed recurrence reverses the sign of odd-lag responses. The initially feedforward student has vanishing recurrent responses ($\varphi_k = 0$ for $k \geq 1$). Color scale is same for all panels.

### A.4 Initial teacher-student response-vector alignment

To characterise how recurrent initialisation shapes the temporal computation, we evaluate the initial teacher-student response-overlap matrix $Q_{k\ell} = \boldsymbol{\varphi} k \cdot \boldsymbol{\varphi} \ell^*$. Matched and shuffled recurrence produce similar, predominantly positive overlaps, while opposed recurrence reverses the sign of the odd-lag student responses (Fig. S4). In the initially feedforward condition, all recurrent response coefficients vanish, leaving only the zero-lag row. The overlap magnitude decreases rapidly with lag, indicating that the differences between temporal initializations are concentrated in the first few recurrent responses. These results show that the relevant distinction between recurrent configurations lies in their observable temporal-response geometry rather than in recurrence strength alone. The response overlaps contribute to the joint Gaussian trajectory covariance and thereby influence both reward availability and the reward-modulated learning field.

### A.5 Robustness of the temporal-alignment effect

We next examined whether the ordering of learning times persists when assumptions of the analytical model are progressively relaxed. To identify the escape time in these more relaxed settings, in these experiments we used the first crossing of a common behavioural criterion,

$$\tau_{\bar{p}} = \inf\{\varsigma \geq 0 : \bar{p}(\varsigma) > 2/3\}. \tag{31}$$

The displayed settings use recurrent population size $N_h \in \{2, 10\}$, episode length $T \in \{4, 6, 10\}$, and eight random seeds per setting. The FF label identifies the feedforward reference condition used in each comparison.

Starting from a diagonal reference model, we tested the full population models , networks with a nonlinear activation function $f_s(z) = s \tanh(z/s)$ for slope parameters $s = 2.0$ and $s = 0.4$, linear networks with stochastic logistic actions instead of deterministic ones, linear networks learning with a score-function REINFORCE updates, and linear networks with correlated AR(1) observations with lag-one correlation parameter $0.6$. The logistic-policy conditions retain the surrogate update, while the REINFORCE conditions use the policy score. The combined condition considers nonlinear activation, stochastic logistic actions, exact REINFORCE updates, correlated AR(1) observations. The observation process remains open loop in these experiments.

Table 5 summarises the escape-time ratio for a selection of these conditions. Every displayed ratio exceeds one, and the separation increases with episode length within each considered condition. **The exact ratio magnitude nevertheless depends on the considered model and learning rule, with stronger saturation attenuating the observed escape time separation, and exact REINFORCE preserving the ordering across the displayed horizons.**

For $N_h = 2$, Table 6 provides all fourteen conditions. It reports the matched, FF, and opposed escape-time summaries together with $R_\tau$ and its confidence interval. Keeping the absolute time summaries makes visible changes in learning time that a ratio alone does not describe.

Table 5: Opposed-to-matched escape-time ratios for selected robustness conditions. Each setting uses eight seeds.

| condition | $N_h = 2, T = 4$ | $N_h = 2, T = 6$ | $N_h = 2, T = 10$ | $N_h = 10, T = 4$ | $N_h = 10, T = 6$ | $N_h = 10, T = 10$ |
|---|---|---|---|---|---|---|
| diagonal reference | 1.96 | 2.54 | 3.59 | 2.07 | 2.73 | 3.78 |
| full matrices | 1.92 | 2.53 | 3.80 | 2.26 | 2.96 | 4.68 |
| tanh, $s = 2.0$ | 1.70 | 2.11 | 2.98 | 1.84 | 2.26 | 3.07 |
| tanh, $s = 0.4$ | 1.18 | 1.20 | 1.37 | 1.23 | 1.34 | 1.48 |
| exact REINFORCE, $\beta\,\mathrm{sd} = 1$ | 1.74 | 1.93 | 2.24 | 1.75 | 1.93 | 2.21 |
| exact REINFORCE, $\beta\,\mathrm{sd} = 3$ | 1.74 | 2.20 | 2.76 | 1.87 | 2.14 | 2.77 |
| AR(1) inputs, $a = 0.6$ | 2.74 | 3.67 | 6.50 | 2.73 | 3.78 | 6.88 |
| combined relaxation | 3.03 | 4.87 | 12.09 | 1.56 | 2.18 | 3.41 |

Table 6: Escape-time robustness results for $N_h = 2$.

| condition | $T$ | $\tau_{\mathrm{mat}}$ | $\tau_{\mathrm{FF}}$ | $\tau_{\mathrm{opp}}$ | ratio [95% CI] |
|---|---|---|---|---|---|
| diagonal reference | 4 | 4.02 | 5.29 | 7.89 | 1.96 [1.83, 2.12] |
| diagonal reference | 6 | 10.97 | 16.15 | 27.84 | 2.54 [2.38, 2.72] |
| diagonal reference | 10 | 92.36 | 177.49 | 331.65 | 3.59 [3.36, 3.81] |
| full matrices | 4 | 3.25 | 4.48 | 6.23 | 1.92 [1.80, 2.06] |
| full matrices | 6 | 8.58 | 12.74 | 21.68 | 2.53 [2.34, 2.76] |
| full matrices | 10 | 70.85 | 138.84 | 269.11 | 3.80 [3.45, 4.18] |
| tanh, $s = 2.0$ | 4 | 4.56 | 5.66 | 7.76 | 1.70 [1.60, 1.82] |
| tanh, $s = 2.0$ | 6 | 12.78 | 17.07 | 27.01 | 2.11 [1.98, 2.26] |
| tanh, $s = 2.0$ | 10 | 111.28 | 184.66 | 331.65 | 2.98 [2.73, 3.27] |
| tanh, $s = 0.4$ | 4 | 8.07 | 8.65 | 9.53 | 1.18 [1.12, 1.24] |
| tanh, $s = 0.4$ | 6 | 24.41 | 26.20 | 29.32 | 1.20 [1.16, 1.25] |
| tanh, $s = 0.4$ | 10 | 217.55 | 255.37 | 298.22 | 1.37 [1.25, 1.50] |
| logistic, $\beta\,\mathrm{sd} = 1$ | 4 | 5.36 | 6.75 | 9.33 | 1.74 [1.65, 1.84] |
| logistic, $\beta\,\mathrm{sd} = 1$ | 6 | 17.45 | 23.15 | 33.73 | 1.93 [1.82, 2.05] |
| logistic, $\beta\,\mathrm{sd} = 1$ | 10 | 177.69 | 264.01 | 391.00 | 2.20 [2.07, 2.35] |
| logistic, $\beta\,\mathrm{sd} = 3$ | 4 | 4.18 | 5.68 | 8.22 | 1.97 [1.88, 2.08] |
| logistic, $\beta\,\mathrm{sd} = 3$ | 6 | 12.00 | 17.45 | 29.76 | 2.48 [2.31, 2.67] |
| logistic, $\beta\,\mathrm{sd} = 3$ | 10 | 107.76 | 196.80 | 365.73 | 3.39 [3.13, 3.70] |
| exact REINFORCE, $\beta\,\mathrm{sd} = 1$ | 4 | 3.91 | 4.65 | 6.80 | 1.74 [1.65, 1.84] |
| exact REINFORCE, $\beta\,\mathrm{sd} = 1$ | 6 | 12.78 | 16.19 | 24.67 | 1.93 [1.83, 2.03] |
| exact REINFORCE, $\beta\,\mathrm{sd} = 1$ | 10 | 130.91 | 185.40 | 293.25 | 2.24 [2.10, 2.39] |
| exact REINFORCE, $\beta\,\mathrm{sd} = 3$ | 4 | 2.21 | 2.55 | 3.85 | 1.74 [1.63, 1.86] |
| exact REINFORCE, $\beta\,\mathrm{sd} = 3$ | 6 | 6.07 | 8.00 | 13.39 | 2.20 [2.05, 2.38] |
| exact REINFORCE, $\beta\,\mathrm{sd} = 3$ | 10 | 55.71 | 82.95 | 153.91 | 2.76 [2.59, 2.95] |
| AR(1) inputs, $a = 0.6$ | 4 | 1.62 | 2.56 | 4.43 | 2.74 [2.64, 2.85] |
| AR(1) inputs, $a = 0.6$ | 6 | 2.90 | 5.30 | 10.64 | 3.67 [3.43, 3.94] |
| AR(1) inputs, $a = 0.6$ | 10 | 9.61 | 24.04 | 62.52 | 6.50 [6.01, 7.08] |
| combined relaxation | 4 | 0.78 | 1.20 | 2.36 | 3.03 [2.85, 3.24] |
| combined relaxation | 6 | 1.82 | 3.39 | 8.87 | 4.87 [4.62, 5.12] |
| combined relaxation | 10 | 9.30 | 29.63 | 112.40 | 12.09 [11.43, 12.79] |

Table 7 presents the corresponding results at $N_h = 10$. Across both widths, the listed escape-time summaries satisfy $\tau_{\mathrm{mat}} < \tau_{\mathrm{FF}} < \tau_{\mathrm{opp}}$ in every tested condition. These comparisons support the numerical robustness of the ordering over the displayed settings.

Table 7: Escape-time robustness results for $N_h = 10$. Intervals indicate 95% confidence level.

| condition | $T$ | $\tau_{\mathrm{mat}}$ | $\tau_{\mathrm{FF}}$ | $\tau_{\mathrm{opp}}$ | ratio [95% CI] |
|---|---|---|---|---|---|
| diagonal reference | 4 | 8.68 | 11.92 | 17.95 | 2.07 [1.92, 2.22] |
| diagonal reference | 6 | 23.51 | 36.74 | 64.16 | 2.73 [2.59, 2.86] |
| diagonal reference | 10 | 203.72 | 401.86 | 770.83 | 3.78 [3.55, 4.03] |
| full matrices | 4 | 2.01 | 2.94 | 4.55 | 2.26 [2.05, 2.51] |
| full matrices | 6 | 5.73 | 9.22 | 16.95 | 2.96 [2.62, 3.32] |
| full matrices | 10 | 42.17 | 91.92 | 197.30 | 4.68 [3.94, 5.67] |
| tanh, $s = 2.0$ | 4 | 10.51 | 13.82 | 19.36 | 1.84 [1.74, 1.96] |
| tanh, $s = 2.0$ | 6 | 29.52 | 42.04 | 66.85 | 2.26 [2.17, 2.36] |

Table 8: Delayed-cue memory experiment. Frozen-policy accuracies are distinguished from the peak accuracy $p_{\max}$ attained during learning. Learned runs start at $b_0 = 0$.

| $T$ | reward at timestep | frozen $b = 0$ | frozen $b = 0.95$ | learned $p_{\max}$ | final $b$ | scalar ceiling |
|---|---|---|---|---|---|---|
| 4 | all | 0.500 | 0.699 | 0.687 | 0.917 | 0.718 |
| 6 | all | 0.500 | 0.649 | 0.631 | 0.999 | 0.673 |
| 4 | final | 0.500 | 0.699 | 0.611 | 0.999 | 0.718 |
| 6 | final | 0.500 | 0.649 | 0.630 | 0.999 | 0.673 |

| condition | $T$ | $\tau_{\mathrm{mat}}$ | $\tau_{\mathrm{FF}}$ | $\tau_{\mathrm{opp}}$ | ratio [95% CI] |
|---|---|---|---|---|---|
| tanh, $s = 2.0$ | 10 | 256.41 | 446.75 | 786.27 | 3.07 [2.92, 3.22] |
| tanh, $s = 0.4$ | 4 | 33.44 | 36.60 | 41.14 | 1.23 [1.20, 1.26] |
| tanh, $s = 0.4$ | 6 | 97.20 | 110.47 | 130.60 | 1.34 [1.31, 1.38] |
| tanh, $s = 0.4$ | 10 | 911.35 | 1097.40 | 1347.90 | 1.48 [1.43, 1.52] |
| logistic, $\beta\,\mathrm{sd} = 1$ | 4 | 11.81 | 14.76 | 20.43 | 1.73 [1.66, 1.80] |
| logistic, $\beta\,\mathrm{sd} = 1$ | 6 | 35.51 | 49.40 | 72.41 | 2.04 [1.97, 2.12] |
| logistic, $\beta\,\mathrm{sd} = 1$ | 10 | 380.01 | 579.12 | 897.14 | 2.36 [2.22, 2.53] |
| logistic, $\beta\,\mathrm{sd} = 3$ | 4 | 9.27 | 12.24 | 18.35 | 1.98 [1.88, 2.08] |
| logistic, $\beta\,\mathrm{sd} = 3$ | 6 | 26.38 | 38.67 | 62.86 | 2.38 [2.25, 2.53] |
| logistic, $\beta\,\mathrm{sd} = 3$ | 10 | 229.49 | 409.13 | 1015.08 | 4.42 [3.30, 6.47] |
| exact REINFORCE, $\beta\,\mathrm{sd} = 1$ | 4 | 18.95 | 22.80 | 33.20 | 1.75 [1.70, 1.81] |
| exact REINFORCE, $\beta\,\mathrm{sd} = 1$ | 6 | 60.94 | 77.50 | 117.58 | 1.93 [1.88, 1.98] |
| exact REINFORCE, $\beta\,\mathrm{sd} = 1$ | 10 | 648.15 | 896.30 | 1429.29 | 2.21 [2.14, 2.27] |
| exact REINFORCE, $\beta\,\mathrm{sd} = 3$ | 4 | 10.26 | 12.38 | 19.19 | 1.87 [1.81, 1.93] |
| exact REINFORCE, $\beta\,\mathrm{sd} = 3$ | 6 | 29.74 | 38.26 | 63.70 | 2.14 [2.10, 2.19] |
| exact REINFORCE, $\beta\,\mathrm{sd} = 3$ | 10 | 266.09 | 389.15 | 737.85 | 2.77 [2.67, 2.88] |
| AR(1) inputs, $a = 0.6$ | 4 | 3.61 | 5.68 | 9.85 | 2.73 [2.63, 2.83] |
| AR(1) inputs, $a = 0.6$ | 6 | 6.27 | 11.66 | 23.72 | 3.78 [3.69, 3.88] |
| AR(1) inputs, $a = 0.6$ | 10 | 20.22 | 50.99 | 139.14 | 6.88 [6.69, 7.07] |
| combined relaxation | 4 | 1.43 | 1.64 | 2.24 | 1.56 [1.50, 1.64] |
| combined relaxation | 6 | 3.58 | 4.80 | 7.81 | 2.18 [2.09, 2.27] |
| combined relaxation | 10 | 24.51 | 41.47 | 83.51 | 3.41 [3.23, 3.59] |

### A.6 A delayed-cue task requiring memory

We further considered a delayed-cue task in which information needed for the decision is absent from the current observations. A binary cue is presented at $t = 1$, followed by noise until the terminal decision at $t = T$. Reward then depends only on whether the terminal action matches the cue. Performance is therefore measured by the terminal-decision accuracy, not by sustained decision accuracy as in the settings considered in the rest of the paper.

Table 8 compares fixed-recurrence control runs with runs in which recurrence is learned from $b_0 = 0$. The "all" credit condition applies the terminal reward to eligibility contributions from all action times;, while the "final" condition uses only the terminal-action contribution. For the scalar reference construction, the ceiling used in this comparison is

$$p_{\mathrm{ceil}}(T) = \Phi\left(\frac{1}{\sqrt{T-1}}\right), \tag{32}$$

where $\Phi$ is the standard normal cummulative distribution function. This value applies to the tested noisy scalar integrator with its recurrent constraint.

At fixed $b = 0$, the terminal response contains no information about the earlier cue and accuracy is $1/2$. With $b = 0.95$, accuracy reaches $0.699$ at $T = 4$ and $0.649$ at $T = 6$. The learning runs also attain above-chance accuracy. Their maximum accuracy $p_{\max}$ and final recurrent coefficient are provided separately because they do not necessarily refer to the same learning time.

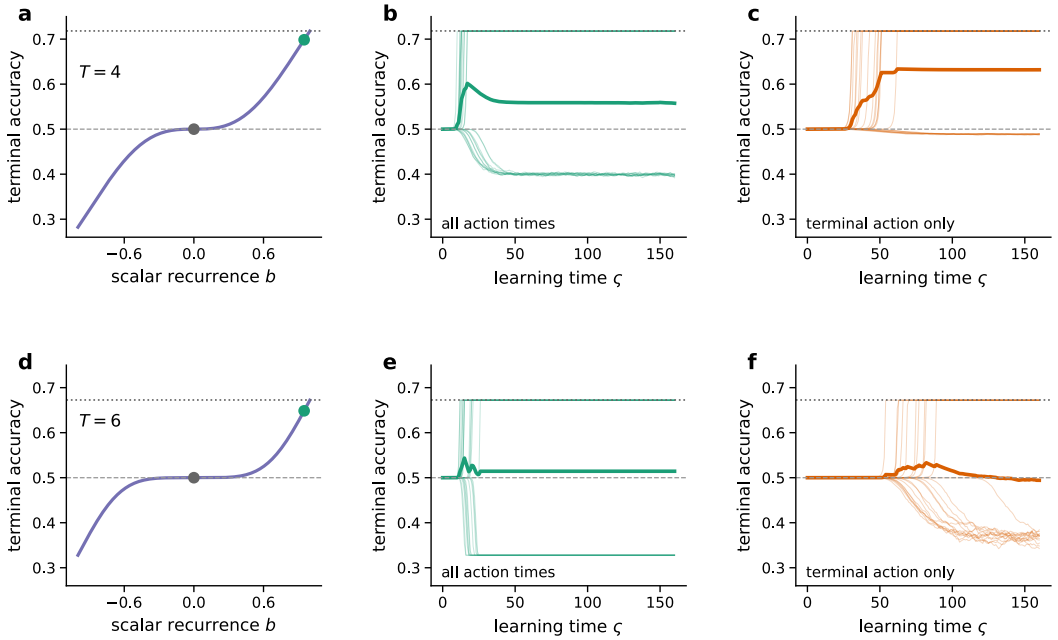



Figure S5: **Memory acquisition from zero recurrence is driven by stochastic fluctuations.** The network receives a balanced binary cue followed by Gaussian noise and earns a reward only for its terminal decision. We fix the input and readout weiights and constrain scalar recurrence to $[-0.999, 0.999]$, starting from zero. **a,d,** Terminal accuracy at fixed weights for horizons four and six, with controls at $b = 0$ and $b = 0.95$. **b,e,** Stochastic learning using all action times. **c,f,** Learning using only the terminal-action term, with the same episode normalisation. Thin curves show individual seeds, thick curves show ensemble means, and dotted lines mark the scalar bound-one ceiling. At zero recurrence, the population update vanishes, so finite fluctuations drive the initial changes. Both positive and negative recurrent branches emerge under the symmetric constraint.

### A.7 Numerical validation of the transfer operator

**Reachable support and finite-step smoothing.** At fixed parameters, the joint teacher-student state follows

$$\boldsymbol{\xi}_{t+1} = \mathsf{A}\boldsymbol{\xi}_t + \Gamma \mathbf{x}_{t+1}, \qquad \mathbf{x}_{t+1} \sim \mathcal{N}(0, I). \tag{33}$$

For these numerical validation tests, we denote its reachable subspace by

$$\mathcal{H} = \mathrm{span}_{k\geq 0}\, \mathrm{im}(\mathsf{A}^k\Gamma), \qquad W_\ell = \sum_{k=0}^{\ell-1} \mathsf{A}^k \Gamma\Gamma^\top (\mathsf{A}^\top)^k. \tag{34}$$

The **controllability index** $\ell_c$ is the smallest positive integer for which $\mathrm{range}(W_{\ell_c}) = \mathcal{H}$. We evaluate all matrices, inverses, and eigenvalues in the following validation tests evaluated in orthonormal coordinates on $\mathcal{H}$.

For stable dynamics, $\rho(\mathsf{A}) < 1$, we let $\Sigma_\infty$ to be the stationary covariance on this support and define

$$C_\ell = \Sigma_\infty^{-1/2} \mathsf{A}^\ell \Sigma_\infty^{1/2}. \tag{35}$$

Its singular values are the canonical correlations between stationary states separated by $\ell$ transitions. The covariance identity

$$I - C_\ell C_\ell^\top = \Sigma_\infty^{-1/2} W_\ell \Sigma_\infty^{-1/2} \tag{36}$$

relates these correlations to the covariance of the accumulated innovations.

Table 9 presents the support dimension, controllability index, smallest Gramian eigenvalue, largest canonical correlation, and relative numerical residual of this identity. Every presented example has

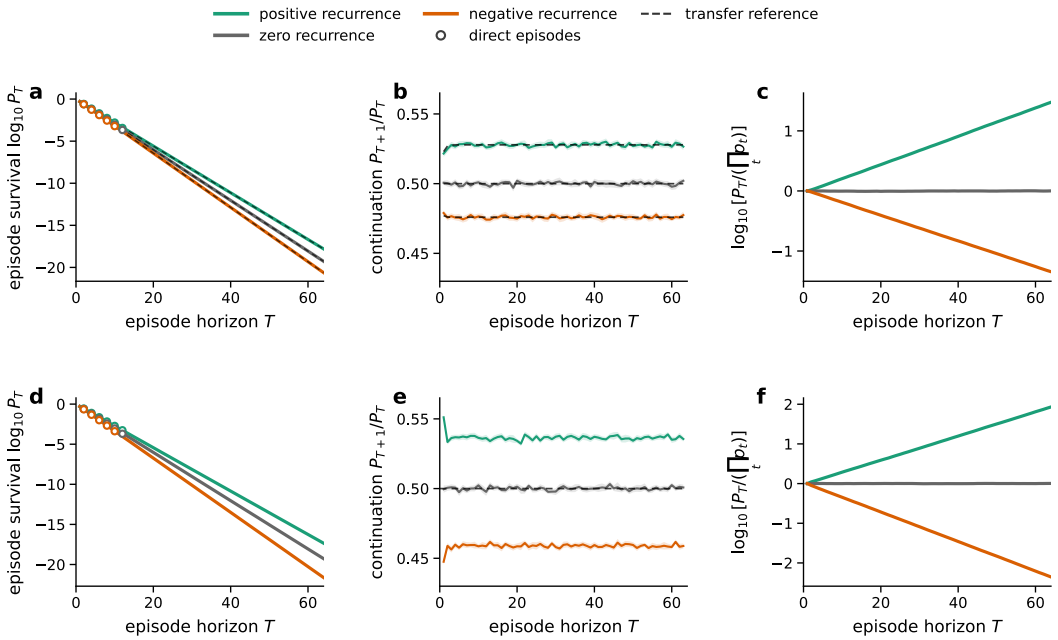



Figure S6: **Temporal dependence changes reward availability even at identical marginal accuracy.** We compare (top row) a homogeneous teacher with $b_* = 0.35$ with students initialised at $b = 0.35, 0, -0.35$, (bottom row) a fixed nilpotent teacher with nilpotent students of positive, zero, and negative coupling, where every recurrent eigenvalue is zero. In both rows, every single-time agreement probability is exactly $1/2$. **a,d**, Episode survival as a function of horizon. Open markers are independent simulation estimates, while curves use a killed-particle calculation. **b,e**, The finite-horizon continuation probability $P_{T+1}/P_T$. Dashed references in the top row come from an independently discretized transfer operator. **c,f**, Survival relative to the product of the exact marginal agreement probabilities. Temporal dependence changes the probability of receiving reward by orders of magnitude over the displayed horizons. Shading indicates 95% numerical confidence intervals over independent particle populations.

Table 9: Finite-step Gaussian support evaluation for the heterogeneous eigenvalue settings. Gramian eigenvalues and canonical correlations are evaluated on the reachable support $\mathcal{H}$.

| model | $\dim \mathcal{H}$ | $\ell_c$ | $\lambda_{\min}(W_{\ell_c})$ | $s_{\max}(C_{\ell_c})$ | identity error |
|---|---|---|---|---|---|
| diagonal, matched | 4 | 1 | $2.00 \times 10^{0}$ | 0.614 | $1.25 \times 10^{-16}$ |
| diagonal, partially matched | 6 | 2 | $6.05 \times 10^{-3}$ | 0.377 | $5.62 \times 10^{-16}$ |
| diagonal, mismatched | 8 | 2 | $6.05 \times 10^{-3}$ | 0.420 | $2.06 \times 10^{-15}$ |
| diagonal + rank-one (A) | 8 | 2 | $4.82 \times 10^{-4}$ | 0.931 | $5.48 \times 10^{-14}$ |
| diagonal + rank-one (B) | 8 | 2 | $6.41 \times 10^{-3}$ | 0.616 | $1.35 \times 10^{-14}$ |
| diagonal + rank-one (C) | 8 | 2 | $1.69 \times 10^{-2}$ | 0.592 | $5.31 \times 10^{-15}$ |

$\lambda_{\min}(W_{\ell_c}) > 0$ and $s_{\max}(C_{\ell_c}) < 1$, with relative identity residual below $6 \times 10^{-14}$. These tests support finite-step Gaussian smoothing on the reachable support. The accessibility and aperiodicity conditions used for the killed process are separate from these free-process diagnostics.

**Finite-horizon survival statistics.** For the same fixed-parameter systems, we estimate the probability of teacher-student agreement at every step through horizon $T$. Table 10 gives the Monte Carlo estimate $\widehat{P}_T$, its number of surviving sampled trajectories, and estimates of

$$r_T = \frac{P_{T+1}}{P_T}, \qquad \kappa_T = \frac{P_{T+2}P_T}{P_{T+1}^2} = \frac{r_{T+1}}{r_T}. \tag{37}$$

Table 10: Finite-horizon survival summaries. Each row is a single horizon snapshot

| model | $T$ | $\widehat{P}_T$ | survivors | $\widehat{P}_{T+1}/\widehat{P}_T$ | $\widehat{P}_{T+2}\widehat{P}_T/\widehat{P}_{T+1}^2$ |
|---|---|---|---|---|---|
| diagonal, matched | 11 | $4.98 \times 10^{-4}$ | 249 | 0.514 | 1.079 |
| diagonal, mismatched | 9 | $8.94 \times 10^{-4}$ | 447 | 0.445 | 1.106 |
| diagonal, partially matched | 11 | $5.74 \times 10^{-4}$ | 287 | 0.544 | 0.873 |
| diagonal + rank-one (A) | 12 | $4.10 \times 10^{-4}$ | 205 | 0.522 | 1.021 |
| diagonal + rank-one (B) | 12 | $6.97 \times 10^{-2}$ | 34,833 | 0.796 | 1.003 |
| diagonal + rank-one (C) | 12 | $1.32 \times 10^{-3}$ | 660 | 0.588 | 0.978 |

Here $r_T = \mu_{T+1}^{\mathrm{cond}}$ is the probability of surviving one additional step, conditional on survival through $T$. The second ratio compares successive conditional survival probabilities, $\log \kappa_T$ is the second finite difference of $\log P_T$. An exponential law $P_T = C\mu^{T-1}$ results in $r_T = \mu$ and $\kappa_T = 1$. Under the corresponding asymptotic law, these limits are approached as $T$ increases.

Each table row contains one selected horizon, with different horizons used for different systems. The entries therefore provide local finite-horizon validation tests.

We further compared the projected models with the full network model over learning, rather than only fit reduced weights to full-model snapshots. In the rank comparison at $T = 4$, $N_h \in \{4, 8\}$, and $\rho_0(0) = 0.35$, the rank-zero model did not reach the escape threshold within the simulated interval. Ranks one and two of the projected shared-subspace model reproduced the opposed-to-matched escape-time ratio to approximately one percent (Table 11). Increasing the rank also reduced the projection and response vector errors. These comparisons establish accuracy in the tested configurations.

Table 11: Opposed-to-matched escape-time ratios in the rank study ($T = 4$, $\rho_0(0) = 0.35$). The rank-zero (purelly diagonal) model did not escape in either configuration over the simulated interval.

| $N_h$ | Unrestricted | $L = 1$ | $L = 2$ |
|---|---|---|---|
| 4 | 1.675 | 1.694 | 1.690 |
| 8 | 1.791 | 1.805 | 1.802 |

### A.8 Numerical tests of the conditional learning dynamics

We use three complementary numerical experiments to examine the dynamical mechanisms outlined in Appendix D. The first separates the effect of rare rewards from progress along a conditional learning path. The second measures the local recurrent field near a feedforward student. The third follows single-unit trajectories and accumulated recurrent changes in an unrestricted network. All use independent Gaussian observations, zero episode resets, and pure success reward $r_1 = 1$, $r_2 = 0$.

**Time-rescaling analysis.** The student has size $N_h = N_x = 2$ initiated within the $L = 0$ rank hierarchy of our model, i.e. it is diagonal,

$$W_x = \mathrm{Diag}(\mathbf{a}), \qquad W_h = \mathrm{Diag}(\mathbf{b}), \qquad W_y = \mathbf{c}^\top.$$

The initial input and readout coefficients are balanced and positive, $a_m(0) = c_m(0) = Q_0^{1/4}$ with $Q_0 = 0.025$. The teacher has weight configuations $W_x^* = I$, $\mathbf{c}^* = (1, -1)^\top$, and $W_h^* = \mathbf{b}^* I$. Thus $\rho_0(0) = 0$. We compare the configurations

$$(b_1(0), b_2(0), \mathbf{b}^*) = \Big\{(0.35, 0.35, 0.35), \quad (0, 0, 0), \quad (0.35, 0.35, -0.35)\Big\},$$

referred to as *matched, feedforward, and opposed*, respectively. The teacher is fixed *within* each run but is different across these three configurations. The input, readout, and diagonal recurrent coefficients are all plastic.

For each $T \in \{3, 4, 5, 6, 8\}$, we estimate the episode moments with a fixed batch of Gaussian observation sequences. The generator uses $400{,}000$ episodes for $T \leq 6$ and $1{,}200{,}000$ episodes for

$T = 8$. With $\widehat{F}$ and $\widehat{P}_T$ denoting the resulting estimates, we advance the diagonal state with Euler integration of timestep $\Delta\sigma = 0.01$ in the rescaled learning time

$$\theta_{n+1} = \theta_n + \Delta\sigma \, \frac{\widehat{F}(\theta_n)}{\widehat{P}_T(\theta_n)}, \qquad \varsigma_{n+1} = \varsigma_n + \frac{\Delta\sigma}{\widehat{P}_T(\theta_n)}. \tag{38}$$

We retain only the diagonal matrix components in this experiment. These are finite-sample approximations to the population ODE. The feedforward population field has zero recurrent drift, small recurrent changes in its sampled implementation are numerical effects.

We stop the siulation at the first crossing of $\rho_0 = 0.5$, using linear interpolation between the two adjacent saved states to obtain both $\tau_\rho$ and $\sigma_e$. This threshold is below the $1/\sqrt{2}$ derived in Appendix B.5, and it is not generally equivalent to $\bar{p} = 2/3$ in a recurrent network. In Fig. 3a,b, we present the same alignment values against $\varsigma$ and $\sigma$. Zero learning time is omitted on the logarithmic axis.

**Local approximation to escape time.** For each clock trajectory, the numerical estimate of the initial log-probability slope is the early secant

$$\widehat{k}_T = \frac{\log \widehat{P}_T(0.05) - \log \widehat{P}_T(0)}{0.05}, \tag{39}$$

where the arguments refer to $\sigma$. We insert this slope, the initial success probability, and the measured $\sigma_e$ into Eq. 176. Figure 3c compares this estimate with the escape-time integral. The comparison therefore evaluates a local approximation along an already computed learning path. The examples show substantially less variation of the alignment crossing time in $\sigma$ than in $\varsigma$.

**Recurrent-field with a feedforward teacher.** To test Eq. 184, we used a $T = 4$, $N_h = N_x = 2$ initially diagonal network with teacher gain ($g = \mathbf{a}^\top \mathbf{c}$), $\mathbf{g}^* = (1, -1)^\top$ and $b^* = 0$. The student gain vectors were $\mathbf{g} = (1, 1)^\top$ and $\mathbf{g} = (1, 0.2)^\top$, with zero-lag response alignments 0 and approximately $0.554700$, respectively. At each of eleven points of sampled $b \in [-0.5, 0.5]$, we set $b_1 = b_2 = b$ and evaluated the recurrent population field using $1{,}500{,}000$ Gaussian episodes.

Fig. 3d demonstrates the first coordinate $F_{b_1}$ of the sampled field, the first coordinate of the linear approximation $K(\mathbf{b}, \mathbf{b})^\top$, and the sampled success probability on the same fixed-parameter slice. The linear approximation is local about $b = 0$, and values at larger $|b|$ illustrate departures from its regime of validity. In the zero-alignment configuration, the student responses are all proportional to a feature orthogonal to the teacher's, so the exact success probability is $2^{-4}$ throughout the slice.

**Single-unit trajectories.** We studied learning in a student network with a single recurrent unit and two input coordinates. The student has an input weight vector $\mathbf{w}_x \in \mathbb{R}^2$, a scalar readout $c$, and recurrent coefficient $b$. The teacher has recurrence $\beta = 0.4$, input vector $(1, 0)$, and unit readout. We consider episodes of length $T = 4$ with pure success reward.

To understand how recurrence interacts with response learning, we studied the conditional learning field in the plane of input–output alignment $\rho$ and student recurrence $b$. We evaluated the field over $\rho \in [-0.3, 0.95]$ and $b \in [-0.9, 0.9]$, fixing $c = |\mathbf{w}_x| = 1$. We then integrated the learning dynamics from $\rho(0) = 0.02$, $c(0) = |\mathbf{w}_x(0)| = 0.4$, and $b(0) \in 0.35, 0, -0.35$, using a rescaled-time step $\Delta\sigma = 0.01$. All three sets of parameters evolve jointly according to Eq. (195).

The resulting trajectories show how the initial recurrence shapes the acquisition of input–output alignment. The conditional learning field contains regions where recurrence increases or decreases, separated by a recurrent nullcline. Because the input norm and readout evolve alongside recurrence, the trajectories need not follow the two-dimensional field evaluated at fixed gain.

**Dependence on initial response gain.** We next asked how the magnitude of the input–output response affects learning. We introduced the response gain $g = c|\mathbf{w}_x|$ and compared initial values $g_0 \in 0.16, 1, 4$, with balanced input and readout magnitudes, $c_0 = |\mathbf{w}_{x,0}|$. All networks start with the same input direction and recurrent coefficients. Changing the positive gain therefore leaves the initial policy and reward probability unchanged, while altering the subsequent learning dynamics.

At low gain, learning primarily increases response vector alignment, with relatively little change in recurrence (Fig. S7a,c,e). As the gain increases, recurrent changes becomes more pronounced, and

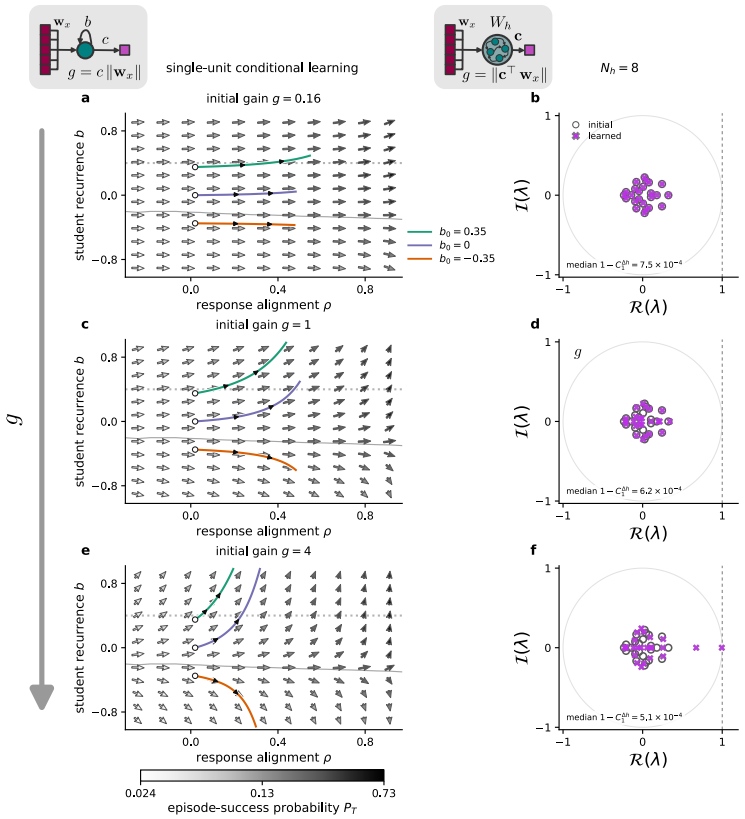



Figure S7: **Initial output gain changes the relative rates of response alignment.** Rows compare initial gains $g = 0.16, 1, 4$ at episode length $T = 4$ and fixed teacher recurrence $\beta = 0.4$. **Left,** Instantaneous single-unit learning fields in the $(\rho, b)$ plane, evaluated at the initial gain of each row on the positive balanced branch. Vector field arrow colour identifies episode-success probability $P_T$ using a logarithmic scale. Curves are independently integrated trajectories from $b_0 = 0.35, 0, -0.35$ (mathced, feedforward, opposed), with initial response alignment $\rho_0 = 0.02$. The grey contour is the nullcline and the dotted line identifies the teacher recurrence value $b^*$. **Right,** Recurrent eigenvalues at initialisation (*hollow grey*) and at a rescaled time $\sigma = 0.22$ (*purple*) in multi-unit networks with $N_h = 8$, $N_x = 24$. Among these three realisations only the initial input-readout scale (gain) changes between rows. In light grey is the unit circle for reference. Left and right columns refer to separate single-unit and multi-unit experiments, not two projections of the same network trajectory.

trajectories reach the imposed recurrent stability boundary before attaining the learning threshold. Thus, networks with identical initial policies can follow different learning trajectories, as the response gain changes the relative evolution of response alignment and recurrence.

We also examined recurrent connectivity changes in unrestricted multi-unit networks with $N_h = 8$ and $N_x = 24$, comparing the initial and learned recurrent spectra at the same rescaled learning time (Fig. S7b,d,f). At higher initial gain, individual recurrent eigenvalues undergo larger displacements, while the accumulated recurrent update remains strongly concentrated along its leading singular direction.

**Dependence on input–readout imbalance.** We then studied whether learning depends only on the overall response gain or also on how this gain is distributed between input and readout weights. We defined the **input-readout imbalance** $\gamma = c/|\mathbf{w}_x|$ and varied $\gamma \in 0.25, 1, 4$ while keeping the initial gain value constant $g(0) = 1$. These initialisations produce identical input–output responses and therefore have the same initial reward statistics, i.e. the networks at initialisaiton have the same response vectors, same pre-activations, produce the same actions, and thereby have also the same reward probabilities.

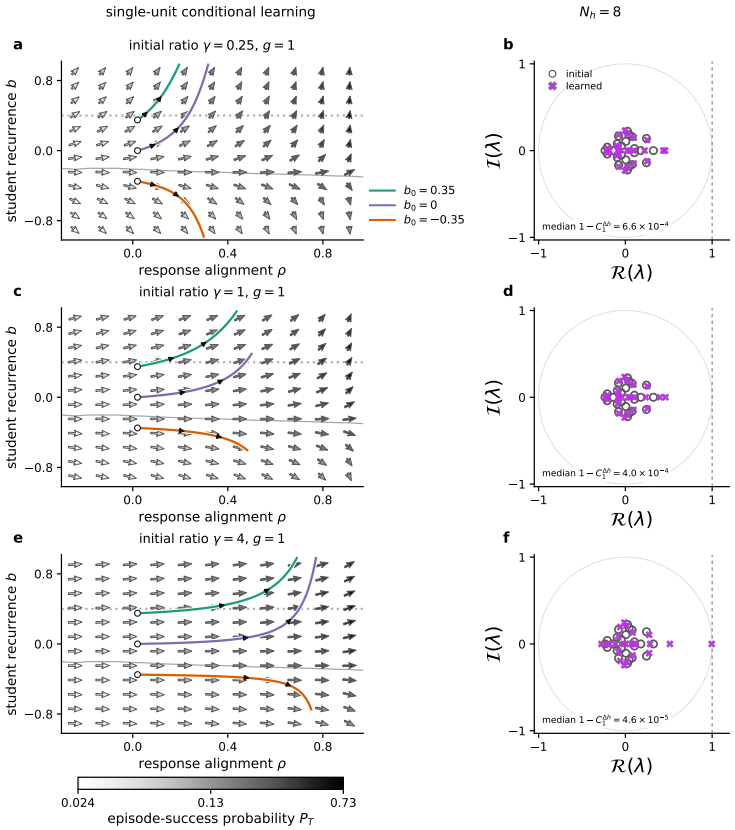



Figure S8: **Input–readout imbalance changes learning despite identical initial response vectors.** Rows compare initial norm ratios $\gamma = 0.25, 1, 4$ at fixed initial response gain $g = 1$, with $T = 4$, teacher recurrence $\beta = 0.4$, and pure success reward. **(a,c,e,)** Single-unit conditional learning fields and trajectories, with $\gamma = c/\|\mathbf{w}_x\|$ and $g = c\|\mathbf{w}_x\|$. Reciprocal changes in the initial input norm and readout preserve the student's temporal responses but change the relative rates of response alignment and recurrent adaptation. Coloured arrows show the instantaneous field at the initial $(g, \gamma)$ of each row; their common logarithmic colour scale indicates $P_T$, and their lengths are normalised. Black trajectories evolve the input weights, readout, and recurrence jointly, so neither gain nor imbalance is held fixed. We follow same colouring and marker conventions as in Fig. S7. The probability colours are unchanged across rows, whereas the field directions and subsequent learning paths differ. **(b,d,f,)** Recurrent spectra from three paired initialisations of a multi-unit recurrent network with $N_h = 8$ and $N_x = 24$. Here we vary $\gamma = \|c\|/\|\mathbf{w}_x\|_F$ by reciprocal input-readout scaling, while the gain $\|c^\top \mathbf{w}_x\| = 1$ remains fixed. Within each pair, this transformation preserves the initial response vector. Hollow grey circles and purple crosses denote initial and learned eigenvalues, respectively, at the rescaled time $\sigma \simeq 0.435$. Together, these comparisons distinguish the parameters determining current behaviour from those determining its evolution during learning.

Despite this equivalence, their learning trajectories differ (Fig. S8). In the single recurrent unit networks, smaller initial imbalance favoured changes in recurrence relative to response vector alignment, while larger imbalance ($\gamma = 4$) allowed alignment to increase further before comparable recurrent changes developed.

Changing the input-output imbalance leaves the initial response vectors unchanged. Indeed, reciprocal rescaling of the input weights and readout preserves their product, and therefore the initial output pre-activations and reward statistics in this setting. The different learning trajectories arise from the sensitivity of the output to changes in the weights. In particular, $\partial_{\mathbf{w}_x} \lambda_t = c\, \mathbf{z}_t$, with $\mathbf{z}_t = \sum_{k=0}^{t-1} b^k \mathbf{x}_{t-k}$. A larger readout therefore amplifies input-weight updates, while a smaller input norm makes these updates produce larger changes in the represented input direction.

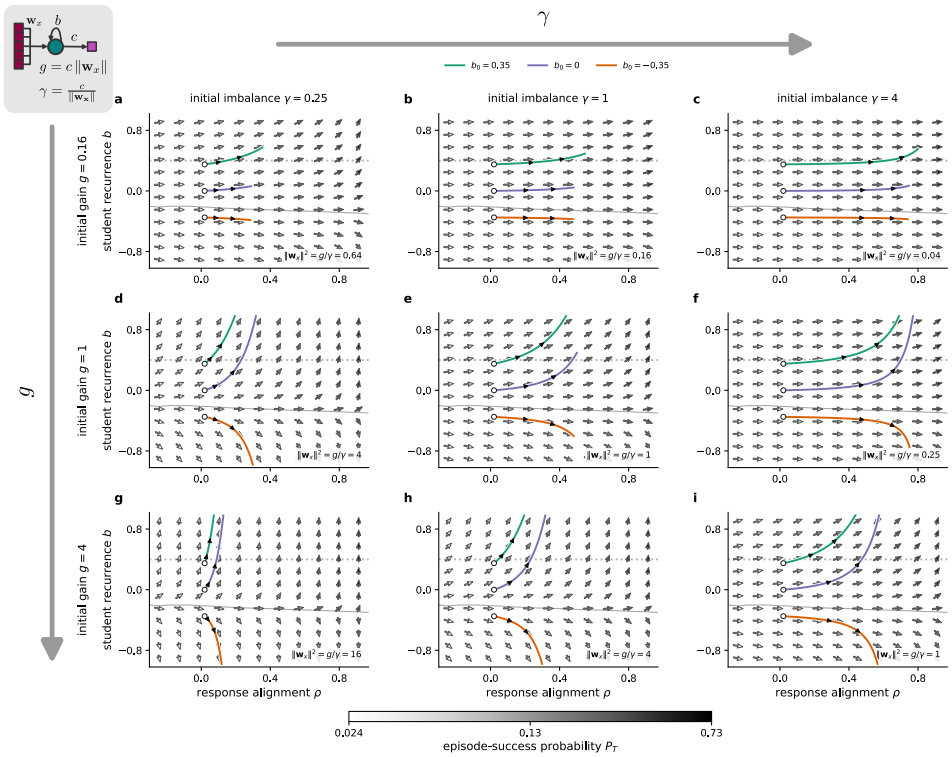



Figure S9: **Jointly varying gain and input-output imbalance reveals that the recurrent weight norm determines the learning trajectory for the single-unit networks.** Learning trajectories for initially matched, feedforward and opposed single-unit teacher-student arangement for different initial input-output gain values (rows) and initial input-output imbalance $\gamma$. In this setting their fraction equals the initial squared recurrent weight norm. These results reveal that in this setting the conditional learning paths (up to time parametrisation) are determined by the recurrent weight norm. Colour conventions and markers same as in Fig. S7.

Instead, the imbalance determines how strongly the parameter updates affect the output. In the single-unit model, a larger readout increases the effect of input-weight changes, while a smaller input norm makes the input more sensitive to rotations induced by transverse updates.

**Why input–readout imbalance affects learning.** For the single-unit student, the output depends on the input weights and scalar readout through their product $c\mathbf{w}_x$. With equal learning rates and pure success reward, we have for the expected update conditioned on success

$$\mathbf{m}_T = \frac{1}{T}\left\langle \sum_{t=1}^{T} y_t \sum_{k=0}^{t-1} b^k \mathbf{x}_{t-k} \,\middle|\, \mathcal{I}_T = 1 \right\rangle.$$

In rescaled learning time, the updates are $\mathrm{d}\mathbf{w}_x/\mathrm{d}\sigma = c\mathbf{m}_T$ and $\mathrm{d}c/\mathrm{d}\sigma = \mathbf{w}_x^\top \mathbf{m}_T$. Their combined effect on the lag-zero response vector is

$$\frac{\mathrm{d}\boldsymbol{\varphi}_0^\top}{\mathrm{d}\sigma} = \frac{\mathrm{d}(c\mathbf{w}_x)}{\mathrm{d}\sigma} = c\frac{\mathrm{d}\mathbf{w}_x}{\mathrm{d}\sigma} + \mathbf{w}_x\frac{\mathrm{d}c}{\mathrm{d}\sigma} = \underbrace{c^2\mathbf{m}_T}_{\text{input update}} + \underbrace{\mathbf{w}_x\mathbf{w}_x^\top\mathbf{m}_T}_{\text{readout update}}. \tag{40}$$

Since the readout $c$ is a scalar and $\boldsymbol{\varphi}_0^\top = c\mathbf{w}_x$, its contribution to the response change is

$$\mathbf{w}_x\frac{\mathrm{d}c}{\mathrm{d}\sigma} = \mathbf{w}_x\mathbf{w}_x^\top\mathbf{m}_T = \frac{\mathbf{w}_x^\top\mathbf{m}_T}{c}\,\boldsymbol{\varphi}_0^\top.$$

This is a scalar multiple of the response vector, and thus all its components change by the same relative amount. For $c > 0$, changing the readout at fixed input weights therefore changes only the

response magnitude, but cannot rotate its direction, and thus makes no direct contribution to the change in normalised teacher-student alignment.

On the other hand, the term that corresponds to the input contribution can also change its direction. Differentiating the normalised teacher-student response overlap therefore results into

$$\frac{\mathrm{d}\rho_0}{\mathrm{d}\sigma} = \gamma \left( \frac{\phi_0^*}{\|\phi_0^*\|} - \rho_0 \frac{\phi_0}{\|\phi_0\|} \right) \mathbf{m}_T, \qquad \gamma = \frac{c}{\|\mathbf{w}_x\|}. \tag{41}$$

At fixed initial gain $g = c\|\mathbf{w}_x\|$, input direction, and recurrence, both the response vectors and the conditional moment $\mathbf{m}_T$ are identical across imbalance conditions. Nevertheless, their alignment velocities differ through $\gamma$, while the initial recurrent velocity, proportional to $g$, remains unchanged. Input-readout imbalance therefore changes the relative evolution of response vector alignment and recurrence despite identical initial behaviour.

Together, these results show that identical initial temporal responses do not imply identical learning dynamics. Both the overall response gain and its distribution between input and readout weights influence how recurrent connectivity evolves. This suggests that their combined effect must be considered to understand the emergence of recurrence during learning.

**Accumulated changes in unrestricted recurrent connectivity.** In the multi-unit unrestricted networks, the recurrent spectra also showed different degrees of eigenvalue displacement, even though reciprocal rescaling of input and readout weights preserves the initial impulse responses.

Fig.S7b uses an unrestricted student with $N_h = 8$, $N_x = 24$, and $T = 4$, trained against a one-unit teacher of recurrence $\beta = 0.5$. The teacher input direction is a normalised Gaussian vector. Student input entries are Gaussian with standard deviation $0.35\sqrt{N_h/N_x}$, readout entries have standard deviation $0.35$, and recurrent entries have standard deviation $0.3/\sqrt{N_h}$.

We advanced the unrestricted weight dynamics in $\sigma$ with step $0.01$. We terminated runs when the recurrent spectral radius exceeds $0.995$ or zero-lag alignment exceeds $0.95$. We compared the initial spectrum with the spectrum at rescaled learning time $\sigma = 0.6$. For the accumulated change $\Delta W_h = W_h(\sigma) - W_h(0)$, we report

$$C_1^{\Delta h} = \frac{\sigma_1(\Delta W_h)^2}{\|\Delta W_h\|_F^2}, \qquad a_c^{\Delta h} = \frac{|u_1(\Delta W_h)^\top c(\sigma)|}{\|c(\sigma)\|}, \tag{42}$$

where $u_1$ is the leading left singular vector. These identify an approximately rank-one, readout-aligned **accumulated change** in the recurrent connectivity. It should be mentioned that these observations are different from the instantaneous joint-field statistic $C_1$ in the main-text shared-subspace comparison.

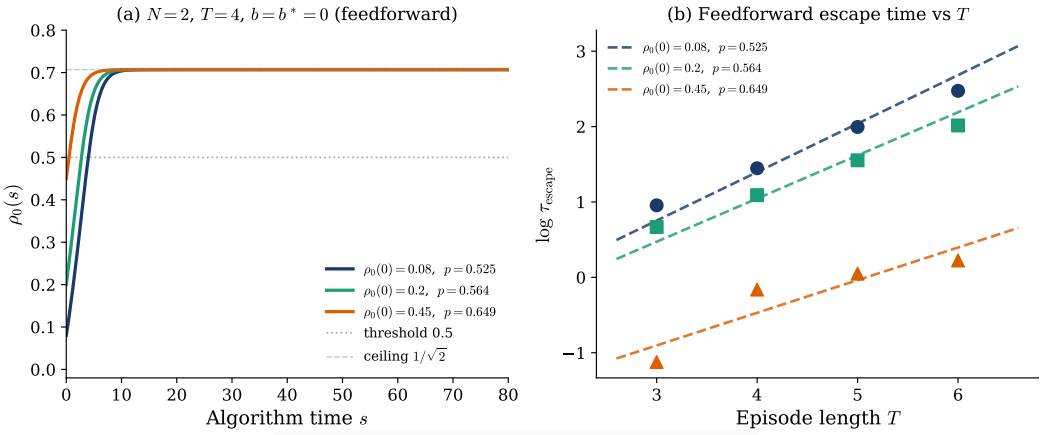



Figure S10: **Feedforward limit.** Setting $b = b^* = 0$ recovers factorized feedforward sparse-reward statistics. (a) Alignment trajectories for three initial $\rho_0(0)$ values in the two-mode anti-aligned construction. (b) Escape-time slopes follow $\tau \propto \mathfrak{p}(\rho_0(0))^{-(T-1)}$.

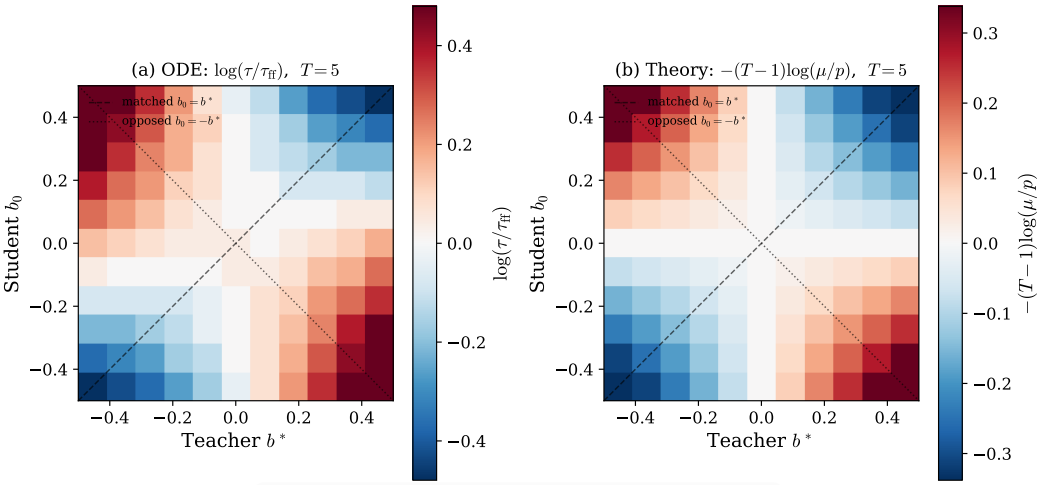



Figure S11: **Plateau-escape phase diagram in the $(b_0, b^*)$ plane.** Blue indicates faster escape than feedforward; red indicates slower escape. The matched diagonal $b_0 = b^*$ accelerates learning, while the opposed diagonal $b_0 = -b^*$ slows it. The approximately hyperbolic contours reflect the perturbative bilinear structure $\mathcal{J} \propto b_0 b^*$ in the symmetric two-mode setting.

-

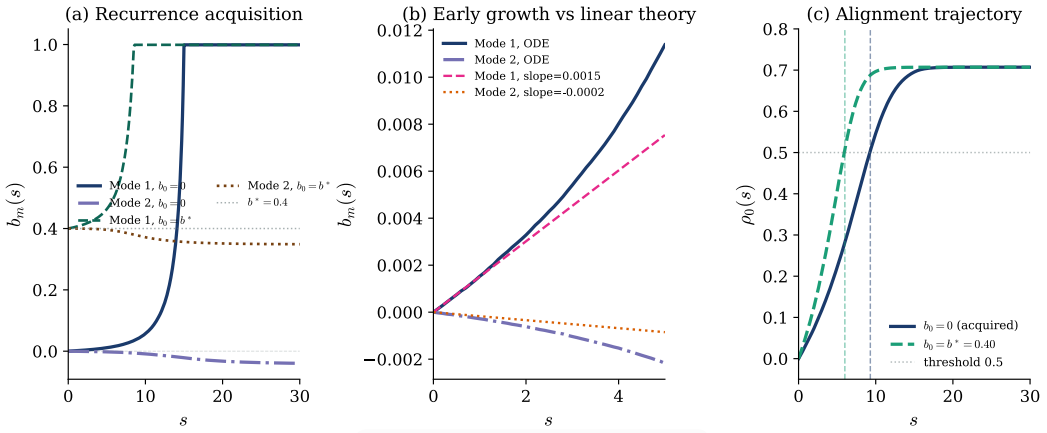



Figure S12: **Sparse reward can induce recurrent-mode growth.** Starting from $b_0 = 0$, the teacher-aligned recurrent mode grows during learning. (a) Recurrent eigenvalues over learning time. (b) Early growth agrees with the source term. (c) A matched recurrent initialization escapes earlier than a feedforward initialisation, consistent with the plateau-rate correction.

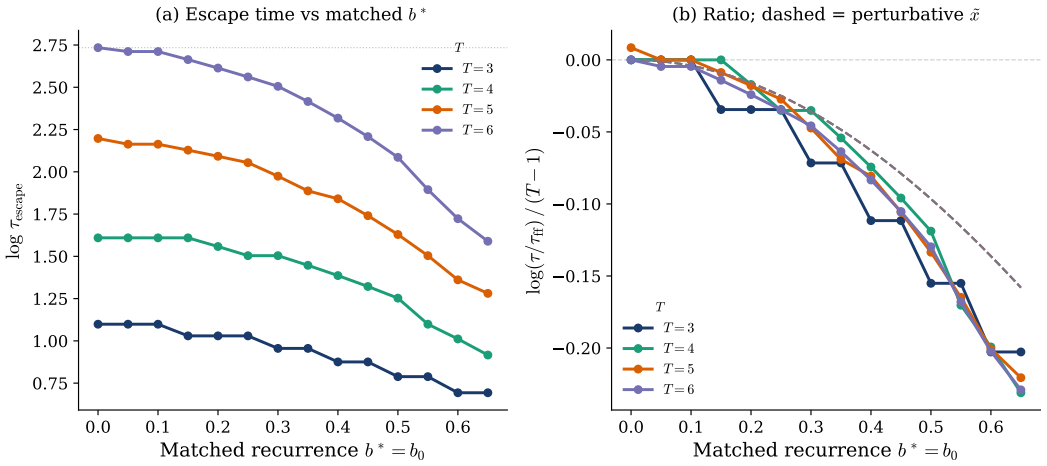



Figure S13: **Matched recurrence accelerates plateau escape in the perturbative regime.** (a) Escape time decreases as the matched recurrent eigenvalue $b_0 = b^*$ increases over the range shown. (b) Ratio-normalized escape times collapse across episode lengths; dashed curves show the perturbative prediction.

## B Order parameters and the reduction to low-dimensional dynamics

In the following section, we present step-by-step the reduction of the learning dynamics for the considered RL-$\ell$RNN teacher-student setting. We start in Section B.1.1 with the presentations of the relevant equations for the general learning problem without considering any reduction. We introduce the two-time correlation kernels that are relevant for this formulation, and discuss the issues they create for closing the equations self-consistently. In Section B.4, we study the structure of the weight derivatives and introduce a reduced representation of the connectivity that separates the diagonal coefficients from factorised off-diagonal interactions. We then derive the learning equations for the reduced model description in Section B.4.3 and show how to assess their evolution using the Gaussian trajectory statistics.

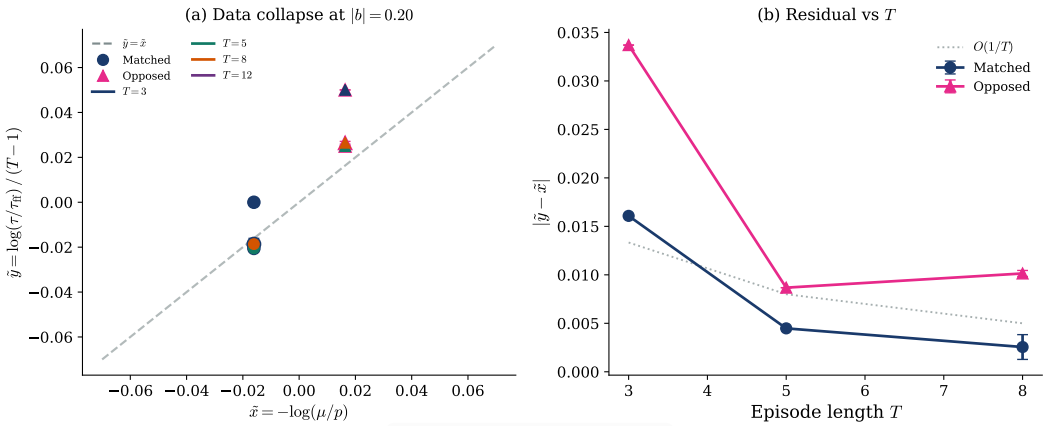



Figure S14: **Data collapse convergence with episode length $T$.** $N = 2$ anti-aligned, $|b| = 0.20$, matched and opposed configurations. (a) $\tilde{y}$ vs $\tilde{x}$; the two configurations lie on opposite sides of the diagonal, consistent with the sign-indeterminate $\mathcal{J}$. Points at different $T$ overlap because $\tilde{x}$ is evaluated at fixed initial order parameters. (b) Residual $|\tilde{y} - \tilde{x}|$ vs $T$; the matched configuration decays approximately as $O(1/T)$ (dotted), consistent with the perturbative analysis.

## B.1 The general setting

### B.1.1 Output pre-activations as Gaussian functionals

We assume a network that receives inputs $\mathbf{x}_t^{(\mu)} \in \mathbb{R}^{N_x}$, has recurrent state $\mathbf{h}_t^{(\mu)} \in \mathbb{R}^{N_h}$, and coupling parameters $W_x \in \mathbb{R}^{N_h \times N_x}$, $W_h \in \mathbb{R}^{N_h \times N_h}$, $W_y \in \mathbb{R}^{1 \times N_h}$. The student output pre-activation at time $t$ is

$$\lambda_t^{(\mu)} = W_y^\top \mathbf{h}_{t+1}^{(\mu)} = \sum_{s=1}^{t} W_y^\top W_h^{t-s} W_x \mathbf{x}_s^{(\mu)}. \tag{43}$$

For convenience, we define the vector

$$\boldsymbol{\varphi}_t \equiv \left( W_x^\top W_h^{(t-1)\top} W_y, W_x^\top W_h^{(t-2)\top} W_y, \ldots, W_x^\top W_y \right) \in \mathbb{R}^{N_x\, t}, \tag{44}$$

so that the output pre-activation can be expressed in terms of $\boldsymbol{\varphi}_t$ as $\lambda_t^{(\mu)} = \boldsymbol{\varphi}_t^\top \mathbf{x}_{1:t}^{(\mu)}$, where $\mathbf{x}_{1:t}^{(\mu)} = (\mathbf{x}_1^{(\mu)}, \ldots, \mathbf{x}_t^{(\mu)}) \in \mathbb{R}^{N_x\, t}$ denotes the trajectory of the received inputs. The output pre-activation is a linear combination of i.i.d. Gaussian random variables, and therefore $\lambda_t^{(\mu)}$ is itself Gaussian for a fixed weight configuration.

In a similar vein, we have for the teacher output pre-activation $\nu_t^{(\mu)} = \boldsymbol{\varphi}_t^{*\top} \mathbf{x}_{1:t}^{(\mu)}$. The joint distribution of the output pre-activations for the teacher and student $Z^{(\mu)} \doteq (\lambda_{1:T}^{(\mu)}, \nu_{1:T}^{(\mu)})$ is a zero-mean $(2\,T)$-dimensional Gaussian distribution.

### B.1.2 Covariance kernels of output pre-activation as order parameters

The joint Gaussian distribution is fully specified by the $(2\,T) \times (2\,T)$ covariance matrix with entries $\begin{pmatrix} \Lambda_{tt'} & M_{tt'} \\ M_{t't} & N_{tt'} \end{pmatrix}$. We define each entry of the covariance matrix as follows

$$\Lambda_{tt'} \equiv \langle \lambda_t^{(\mu)} \lambda_{t'}^{(\mu)} \rangle = \sum_{s=1}^{\min(t,t')} W_y^\top W_h^{t-s} W_x W_x^\top (W_h^{t'-s})^\top W_y, \tag{45}$$

$$N_{tt'} \equiv \langle \nu_t^{(\mu)} \nu_{t'}^{(\mu)} \rangle = \sum_{s=1}^{\min(t,t')} W_y^{*\top} W_h^{*t-s} W_x^* W_x^{*\top} (W_h^{*t'-s})^\top W_y^*, \tag{46}$$

$$M_{tt'} \equiv \langle \lambda_t^{(\mu)} \nu_{t'}^{(\mu)} \rangle = \sum_{s=1}^{\min(t,t')} W_y^\top W_h^{t-s} W_x W_x^{*\top} (W_h^{*t'-s})^\top W_y^*. \tag{47}$$

In the previous equations we have taken into account the input statistics $\langle \mathbf{x}_i^{(\mu)} (\mathbf{x}_j^{(\mu)})^\top \rangle = \delta_{ij} \mathrm{I}_{N_x}$, i.e. independence of the inputs across dimensions. These three $T \times T$ matrices can be viewed as the natural order parameters for studying learning dynamics in this system, since they fully characterize the joint teacher-student output pre-activation distribution at each parameter configuration.

In Eq. 45-47 we have the products of the input matrices

$$\mathcal{W}^{(xx)} = W_x W_x^\top \in \mathbb{R}^{N_h \times N_h}, \qquad \mathcal{W}^{(x^* x^*)} = W_x^* W_x^{*\top}, \qquad \mathcal{W}^{(x x^*)} = W_x W_x^{*\top}, \tag{48}$$

which are $N_h \times N_h$ matrices. As $N_x \to \infty$ with $N_h$ finite, these quantities still depend only on the finite parameters and do not depend on the input realization.

We can rewrite the entries of the covariance matrix in terms of these input matrices

$$\Lambda_{tt'} = \sum_{s=1}^{\min(t,t')} W_y^\top W_h^{t-s} \mathcal{W}^{(xx)} (W_h^{t'-s})^\top W_y, \qquad M_{tt'} = \sum_{s=1}^{\min(t,t')} W_y^\top W_h^{t-s} \mathcal{W}^{(x^* x)} (W_h^{*t'-s})^\top W_y^*, \tag{49}$$

and similarly for $N_{tt'}$.

With this rewritting, it is evident that we can describe the dynamical state of the system at any learning time point $\varsigma$ by
**(i)** the matrices $\mathcal{W}^{(xx)}, \mathcal{W}^{(x^* x)}, \mathcal{W}^{(x^* x^*)}$, which are all $N_h \times N_h$ ($N_h^2$ scalars each, $3N_h^2$ total with $\mathcal{W}^{(x^* x^*)}$ fixed),
**(ii)** the recurrent matrices $W_h, W_h^*$, which are $N_h \times N_h$ ($2N_h^2$ scalars with $W_h^*$ fixed), and
**(iii)** the readouts $W_y, W_y^*$, which are $N_h$-dimensional ($2N_h$ scalars with $W_y^*$ fixed).

Thus the dynamical order parameters of the problem are an $N_h \times N_h$ matrix triplet $(\mathcal{W}^{(xx)}, \mathcal{W}^{(x^* x)}, W_h)$ plus an $N_h$-dimensional vector $(W_y)$. These comprise a finite-dimensional state that doesn't grow with dimension $N_x$ or time $T$, but grows quadratically with the size of the recurrent network $N_h$.

**However, the kernels $\Lambda_{tt'}, N_{tt'}, M_{tt'}$ are *not* themselves independent dynamical variables.** They are quantities resulting from the matrices above. The number of independent entries in the kernels would be $\sim T^2$, but the parametrization shows they are determined by $\sim N_h^2$ scalars.

### B.1.3 Closure of the dynamics in the general formulation

To obtain order-parameter ODEs for the evolution of the afformentioned weight parameters we have to take into account the linearized REINFORCE update and self-averaging. The increments to $\mathcal{W}^{(xx)}, \mathcal{W}^{(x^* x)}, W_h, W_y$ per episode are computable, in the $N_x \to \infty$ limit they become deterministic, and have the general form

$$\frac{\mathrm{d}\mathcal{W}^{(xx)}}{\mathrm{d}\varsigma} = \mathcal{F}_1(\mathcal{W}, W_h, W_y; \text{input statistics}), \tag{50}$$

and similarly for the other matrices, where the right-hand sides involve trajectory expectations under the joint Gaussian determined by the kernels $\Lambda_{tt'}, N_{tt'}, M_{tt'}$. Although this closure is conceptually

straightforward, it is nevertheless computationally heavy with $\sim 3N_h^2 + 2N_h^2 + 2N_h \sim 5N_h^2$ coupled ODEs. This makes it impractical even for a moderately sized $N_h$.

### B.2 Model assumptions and learning rule

Here we outline the essential assumptions for our theory, and introduce the considered learning rule.

**Assumption 1 (whitened input process).** We assume that the input sequence is whitened both across input dimensions and across time $t$ within a learning episode and across episodes $\mu$,

$$\mathbf{x}_t^{(\mu)} \overset{\text{iid}}{\sim} \mathcal{N}(0, I_{N_x}), \qquad \left\langle \mathbf{x}_t^{(\mu)} \mathbf{x}_{t'}^{(\mu')\top} \right\rangle = \delta_{tt'}\delta_{\mu\mu'} I_{N_x}. \tag{51}$$

**Assumption 2 (linear recurrent dynamics and episode resets).** Within each learning episode, the student and teacher state evolves according to linear dynamical equations

$$\begin{aligned} \mathbf{h}_{t+1} &= W_h \mathbf{h}_t + W_x \mathbf{x}_t, & \mathbf{h}_1 &= 0, & \lambda_t &= \mathbf{c}^\top \mathbf{h}_{t+1}, \\ \mathbf{h}^*_{t+1} &= W^*_h \mathbf{h}^*_t + W^*_x \mathbf{x}_t, & \mathbf{h}^*_1 &= 0, & \nu_t &= (\mathbf{c}^*)^\top \mathbf{h}^*_{t+1} \end{aligned} \tag{52}$$

and resets at the begining of it. In the previous equations we write for brevity $\mathbf{c} = W_y^\top$

Their associated actions are thereby

$$y_t = \operatorname{sgn}(\lambda_t), \qquad y_t^* = \operatorname{sgn}(\nu_t). \tag{53}$$

The student weights remain fixed within each episode and update between episodes, while the teacher weights remain fixed throughout learning.

**Reward.** We consider a setting, where the student does not observe the teacher's individual actions. The only feedback it receives regarding the correctness of their actions is the scalar reward $G$.

For a task where the all student actions have to match those of the teacher, we define the **episode-success indicator** $\mathcal{I}_T^{(\mu)}$ and **reward** $G^{(\mu)}$ obtained at episode $\mu$

$$\mathcal{I}_T^{(\mu)} = \prod_{t=1}^{T} \Theta(\lambda_t^{(\mu)} \nu_t^{(\mu)}), \qquad G^{(\mu)} = (r_1 + r_2) I_T^{(\mu)} - r_2, \qquad r_1, r_2 \geq 0. \tag{54}$$

Here $\Theta(z)$ denotes the indicator of $z > 0$, and $r_1 \geq 0$ stands for the reward obtained for reporting matching decisions over all timesteps $[1, T]$, while $r_2 \geq 0$ is the penalty for non-matching ones.

A successful episode thereby has probability

$$\mathrm{P}_T(S^{(\mu)}) = \left\langle \mathcal{I}_T^{(\mu)} \right\rangle_Z = \mathbb{P}(\lambda_t \nu_t > 0 \text{ for all } t), \tag{55}$$

where $S^{(\mu)}$ denotes the system state at episode $\mu$. Its domain is a union of $2^T$ orthants in the joint trajectory space, one orthant for each common action sequence. The agreement of the teacher and the student can persist while both their actions change sign, i.e. from $(+1, +1)$ to $(-1, -1)$.

#### B.2.1 Linearised reward-modulated update rule

The hard-sign deterministic policy we consider for the student is unfortunately not differentiable. Following the linearized REINFORCE/perceptron analogue used in Patel et al. [2025], we replace the derivative of the sign function by the derivative of the preactivation.

Collecting all weight parameters into one variable $\theta = (W_x, W_h, W_y)$, the reward-modulated surrogate amounts to

$$\Delta\theta^{(\mu)} = \frac{\eta G^{(\mu)}}{T} \sum_{t=1}^{T} y_t^{(\mu)} \nabla_\theta \lambda_t^{(\mu)}. \tag{56}$$

We consider the derivatives at the pre-update weights, and apply the same learning rate for the three weight matrices.

This is the learning rule we consider for the rest of the analysis. In the following sections we first study the weight derivatives of the general system, and then introduce a model reduction and study the reduced population learning field.

### B.3 Response representation and Gaussian trajectory statistics

For arbitrary weight matrices $W_x, W_h, w_y$ and $W_x^*, W_h^*, w_y^*$ the output pre-activations $\lambda_t$ and $\nu_t$ are linear functionals of the input history of the network. Their joint law $Z = (\lambda_{1:T}, \nu_{1:T})$ is therefore Gaussian, and it depends on the networks only through their impulse responses.

To determine the episode statistics at a fixed network state $\mathcal{S}$ we introduce the (impulse) response vectors for the teacher and the student that indicate how an observation received at time step $t$ influences the network output at lag $k$, i.e. at timestep $t + k$.

**Lag-dependent response vectors.** By unrolling the state equations (Eq. 52) we express the student and teacher output pre-activations in terms of powers of the recurrent connectivity matrices as follows

$$\lambda_t = \sum_{j=1}^{t} W_y W_h^{t-j} W_x \mathbf{x}_j, \qquad \nu_t = \sum_{j=1}^{t} W_y^* (W_h^*)^{t-j} W_x^* \mathbf{x}_j. \tag{57}$$

We introduce the **lag-$k$ response vectors** for the student and the teacher

$$\boldsymbol{\varphi}_k = W_y W_h^k W_x \ \in \mathbb{R}^{1\times N_x}, \qquad \boldsymbol{\varphi}_k^* = W_y^* (W_h^*)^k W_x^* \ \in \mathbb{R}^{1\times N_x}, \qquad k \geq 0, \tag{58}$$

where the subscript integer $k$ denotes a temporal lag, i.e., $\boldsymbol{\varphi}_k \mathbf{x}_{t-k}$ is the contribution of an input received $k$ steps earlier to the current output action at timestep $t$. Consequently, we express the output pre-activations in terms of the response vetors as

$$\lambda_t = \sum_{k=0}^{t-1} \boldsymbol{\varphi}_k \mathbf{x}_{t-k}, \qquad \nu_t = \sum_{k=0}^{t-1} \boldsymbol{\varphi}_k^* \mathbf{x}_{t-k}. \tag{59}$$

Notice that the row vectors $\boldsymbol{\varphi}_k$ are invariant under invertible changes of recurrent-state coordinates, i.e. a transformation $(W_x, W_h, W_y) \mapsto (SW_x, SW_hS^{-1}, W_yS^{-1})$. Since $(SW_hS^{-1})^k = SW_h^kS^{-1}$, the transformed response vector satisfies [Kalman, 1963]

$$\begin{aligned} \widetilde{\boldsymbol{\varphi}}_k &= W_y S^{-1} (SW_h S^{-1})^k SW_x \\ &= W_y S^{-1} SW_h^k S^{-1} SW_x \\ &= W_y W_h^k W_x = \boldsymbol{\varphi}_k. \end{aligned} \tag{60}$$

This invariance implies that different recurrent-state representations can generate identical input-output responses, and thereby the response vectors $\boldsymbol{\varphi}$ characterise the network input-to-output mapping independent of the specific recurrent representations.

**Joint Gaussian trajectory.** We define the vector comprising the concatenation of the output pre-activations of the student and the teacher over all timesteps

$$Z = (\lambda_1, \ldots, \lambda_T, \nu_1, \ldots, \nu_T)^\top. \tag{61}$$

For fixed student weights, this vector is jointly Gaussian,

$$Z \sim \mathcal{N}(0, \Sigma), \qquad \Sigma = \begin{pmatrix} \Lambda_{tt'} & M_{tt'} \\ M_{tt'}^\top & N_{tt'} \end{pmatrix}. \tag{62}$$

The entries of the covariance matrix of this jointly Gaussian distribution result from impulse-response overlaps as follows

$$\begin{aligned} \Lambda_{tt'} &\equiv \langle \lambda_t \lambda_{t'} \rangle = \sum_{j=1}^{\min(t,t')} \boldsymbol{\varphi}_{t-j} \boldsymbol{\varphi}_{t'-j}^\top, \\ M_{tt'} &\equiv \langle \nu_t \lambda_{t'} \rangle = \sum_{j=1}^{\min(t,t')} \boldsymbol{\varphi}_{t-j} \boldsymbol{\varphi}_{t'-j}^{*\top}, \\ N_{tt'} &\equiv \langle \lambda_t \nu_{t'} \rangle = \sum_{j=1}^{\min(t,t')} \boldsymbol{\varphi}_{t-j}^* \boldsymbol{\varphi}_{t'-j}^{*\top}. \end{aligned} \tag{63}$$

These sums contain full impulse-response overlaps, and the inner products are taken over input coordinates. A singular covariance $\Sigma$ defines a Gaussian measure on its support and is in general allowed.

**Lagged response overlaps.** For later use, we introduce here the **lagged response overlaps** as follows

$$\mathcal{Q}_{jk} = \boldsymbol{\varphi}_j \boldsymbol{\varphi}_k^\top, \quad \mathcal{R}_{jk} = \boldsymbol{\varphi}_j (\boldsymbol{\varphi}_k^*)^\top, \quad \mathcal{Q}^*_{jk} = \boldsymbol{\varphi}^*_j (\boldsymbol{\varphi}^*_k)^\top, \qquad 0 \le j, k < T. \tag{64}$$

Thereby we can rewrite the entries of the covariance matrix in terms of these lagged overlaps
$\Lambda_{tt'} = \sum_{s=1}^{\min(t,t')} \mathcal{Q}_{t-s,t'-s}$,
$M_{tt'} = \sum_{s=1}^{\min(t,t')} \mathcal{R}_{t-s,t'-s}$, and
$N_{tt'} = \sum_{s=1}^{\min(t,t')} \mathcal{Q}^*_{t-s,t'-s}$.
Although the Gaussian inputs are independent across input dimensions and time, this independence does not generally carry over to the recurrent states or output pre-activations. The input and recurrent weight matrices combine contributions from different input dimensions, while the recurrent dynamics introduce temporal dependencies because outputs at successive time steps have shared input history. For instance, $\mathrm{Cov}(\lambda_1, \lambda_2) = \boldsymbol{\varphi}_0 \boldsymbol{\varphi}_1^\top$, which is generally nonzero. The dynamics are still linear in the Gaussian inputs, so the student and teacher output trajectories remain jointly Gaussian. Their covariance kernels thus fully characterise the trajectory statistics needed for the subsequent learning calculations.

### B.4 Learning dynamics

In the following we first formulate the microscopic learing dynamics and identify the shared directions of its matrix components. Inspired by this observation, we then introduce a structured connectivity representation and consequently formulate the reduced learning dynamics through a tangent-space least-squares principle.

#### B.4.1 Weight derivatives and the population learning dynamics

**Output weights derivatives.** We define the **backward readout vectors**

$$\mathbf{d}_k = (W_h^\top)^k W_y = (W_h^\top)^k \mathbf{c}, \qquad k \ge 0, \tag{65}$$

that determine how the network output depends on perturbations of the recurrent state that occured $k$ steps earlier. Here we have for brevity set $W_y = \mathbf{c}$.

For an infinintesimal perturbation of the input matrix, by unrolling the state results in $\delta\lambda_t = \sum_{j=1}^{t} \mathbf{d}_{t-j}^\top \delta W_x \mathbf{x}_j$, while an infinintesimal perturbation of the recurrent state follows $\delta\mathbf{h}_{j+1} = W_h \delta\mathbf{h}_j + \delta W_h \mathbf{h}_j$. By propagating this infinitesimal perturbation to the readout results in

$$\nabla_{W_x} \lambda_t = \sum_{j=1}^{t} \mathbf{d}_{t-j} \mathbf{x}_j^\top, \qquad \nabla_{W_h} \lambda_t = \sum_{j=1}^{t} \mathbf{d}_{t-j} \mathbf{h}_j^\top, \qquad \nabla_{\mathbf{c}} \lambda_t = \mathbf{h}_{t+1}. \tag{66}$$

The $j = 1$ term in the recurrent gradient is zero. At fixed weights, all gradients are linear functions of the within-episode observations.

**Population field and learning time.** We denote $\mathcal{F}(\theta; \mathbf{x}_{1:T})$ as the episode update in Eq. 56 divided by $\eta$, and introduce the learning gradient fields for each weight matrix

$$\begin{gathered} F_x(\theta) = \frac{1}{T} \left\langle G \sum_{t=1}^{T} y_t \nabla_{W_x} \lambda_t \mid \theta \right\rangle, \qquad F_h(\theta) = \frac{1}{T} \left\langle G \sum_{t=1}^{T} y_t \nabla_{W_h} \lambda_t \mid \theta \right\rangle, \\ \mathbf{f}_c(\theta) = \frac{1}{T} \left\langle G \sum_{t=1}^{T} y_t \mathbf{h}_{t+1} \mid \theta \right\rangle. \end{gathered} \tag{67}$$

We introduce the learning time by $\varsigma = \mu\eta$, and by taking the continuous-time limit we obtain the population flow $\dot{\theta} = F(\theta) = (F_x, F_h, \mathbf{f}_c)$ for $\eta \to 0$ at fixed dimensions and horizon [Benaim, 1996].

**Shared left update directions.** The two matrix fields $F_x$ and $F_h$ in Eq. 67 are sums of outer products with left factors within the subspace [Antoulas, 2005]

$$\mathcal{K}_T(W_h, \mathbf{c}) = \operatorname{span}\{\mathbf{c}, W_h^\top \mathbf{c}, \ldots, (W_h^\top)^{T-1}\mathbf{c}\}. \tag{68}$$

Since these vectors are fixed during an episode, the column space of $[F_x\ F_h]$ belongs into the subspace $\mathcal{K}_T$

$$\operatorname{col}[F_x\ F_h] \subseteq \mathcal{K}_T(W_h, \mathbf{c}), \qquad \dim \mathcal{K}_T \leq \min(T, N_h). \tag{69}$$

The input and recurrent derivatives therefore suggest a shared left factor, while the right factors can be distinct.

### B.4.2 Structured connectivity and reduced variables

The shared left directions of the weight derivatives motivate a shared factorization of the input and recurrent matrices. However, this structure does not account for their diagonal entries. We therefore choose to represent the recurrent self-connections and a set of direct input connections separately, allowing them to evolve independently of the shared interactions. We consider this decomposition as defined in a fixed coordinate system.

To properly formalize this separation, we denote the minimum dimension $r = \min(N_h, N_x)$ and define the rectangular diagonal embedding $\mathcal{D}_x : \mathbb{R}^r \to \mathbb{R}^{N_h \times N_x}$ by

$$[\mathcal{D}_x(\mathbf{a})]_{i\alpha} = \begin{cases} a_i, & i = \alpha \leq r, \\ 0, & \text{otherwise.} \end{cases} \tag{70}$$

We denote by $\operatorname{diag}_x(Z) = (Z_{11}, \ldots, Z_{rr})^\top$ the vector of selected diagonal entries of a rectangular matrix $Z$, and by $\operatorname{diag}(Y)$ the diagonal of a square matrix $Y$. Conversely, $\operatorname{Diag}(\mathbf{v})$ denotes the square diagonal matrix with diagonal entries $\mathbf{v}$.

We then introduce the **othogonal projection operators** under the Frobenius inner product

$$\mathcal{P}_x(Z) = Z - \mathcal{D}_x(\operatorname{diag}_x(Z)), \qquad \mathcal{P}_h(Y) = Y - \operatorname{Diag}(\operatorname{diag}(Y)). \tag{71}$$

These operators set the selected diagonal entries to zero, while leaving all other entries unchanged.

For $N_x \geq N_h$, the rectangular diagonal embedding reduces to

$$\mathcal{D}_x(\mathbf{a}) = [\operatorname{Diag}(\mathbf{a})\ \ 0]$$

. Unlike recurrent self-connections, the selected entries of the input matrix have no intrinsic significance, they correspond to a particular pairing of input and hidden coordinates. Initial weights must be representable in the chosen family or initialized with an explicitly quantified approximation.

We now represent the student weight matrices at state $S_L = (\mathbf{a}, \mathbf{b}, \mathbf{c}, U, A, B)$ as

$$\begin{aligned} W_x(S_L) &= \mathcal{D}_x(\mathbf{a}) + \mathcal{P}_x(UA^\top), \\ W_h(S_L) &= \operatorname{Diag}(\mathbf{b}) + \mathcal{P}_h(UB^\top), \\ W_y(S_L) &= \mathbf{c}^\top, \end{aligned} \qquad U^\top U = I_L, \tag{72}$$

where

$$\mathbf{a} \in \mathbb{R}^r, \quad \mathbf{b}, \mathbf{c} \in \mathbb{R}^{N_h}, \quad U, B \in \mathbb{R}^{N_h \times L}, \quad A \in \mathbb{R}^{N_x \times L}. \tag{73}$$

The products $UA^\top$ and $UB^\top$ before the projection have rank at most $L$, while their projected counterparts and the full weight matrices do not need to necessarily be rank deficient

**The state of the reduced model.** The evolving state of the reduced model after emplying Eq. 72 is

$$S_L = (\mathbf{a}, \mathbf{b}, \mathbf{c}, U, A, B), \qquad \Psi(S_L) = (W_x(S_L), W_h(S_L), \mathbf{c}). \tag{74}$$

The number of required parameters to be trackedover learning is

$$r + 2N_h + (2N_h + N_x)L$$

, and for $N_x \geq N_h$, this is $3N_h + (2N_h + N_x)L$. When considering this reduced description we evaluate all response vectors and Gaussian moments at $\theta = \Psi(S_L)$ without the need to formualte the full system trajectory.

### B.4.3 Variational projection and master equations

**Formulation of admissible velocities.** We differentiat Eq. 72 and obtain the update equations

$$\begin{aligned} \dot{W}_x &= \mathcal{D}_x(\dot{\mathbf{a}}) + \mathcal{P}_x(\dot{U}A^\top + U\dot{A}^\top), \\ \dot{W}_h &= \mathrm{Diag}(\dot{\mathbf{b}}) + \mathcal{P}_h(\dot{U}B^\top + U\dot{B}^\top). \end{aligned} \tag{75}$$

We impose the constraints $U^\top \dot{U} = 0$, in order to remove redundant rotations within the current basis. At each system state over learning, we choose the admissible velocity that minimizes

$$\|F_x - \dot{W}_x\|_F^2 + \|F_h - \dot{W}_h\|_F^2 + \|\mathbf{f}_c - \dot{\mathbf{c}}\|^2. \tag{76}$$

This applies the tangent-projection principle of a dynamical model reduction [Feppon and Lermusiaux, 2018] to our structured connectivity representation. In practice, we approximate the population learning field by its orthogonal projection onto the admissible instantaneous weight changes of the considered reduced representation. The resulting dynamics therefore follow, at each learning point, the direction that most closely approximates the learning update of the full system, while still remaining compatible with the chosen connectivity representation.

**Projected learning dynamics.** The diagonal and off-diagonal components are orthogonal under the Frobenius inner product, and consequently, by minimizing the difference between the full and reduced learning velocities results in the evolution of the diagonal coefficients and readout

$$\dot{\mathbf{a}} = \mathrm{diag}_x F_x, \qquad \dot{\mathbf{b}} = \mathrm{diag}\, F_h, \qquad \dot{\mathbf{c}} = \mathbf{f}_c. \tag{77}$$

However, the evolution equations of the interaction factors $UA^\top$ and $UB^\top$ are more involved, because the input and recurrent matrices share the same left factor $U$. Changes in $U$ therefore affect both matrices simultaneously, and therefore their velocities must be determined jointly.

To express this projection, we define the linear map

$$\mathcal{J}_{S_L}(\Omega, C_A, C_B) = \left(\mathcal{P}_x(\Omega A^\top + UC_A^\top), \mathcal{P}_h(\Omega B^\top + UC_B^\top)\right), \qquad U^\top \Omega = 0, \tag{78}$$

which maps the changes in the factors to the corresponding changes in the input and recurrent matrices. Here $\Omega$, $C_A$, and $C_B$ have the same dimensions as $U$, $A$, and $B$, respectively.

The resulting reduced learning dynamics are

$$\begin{aligned} \dot{\mathbf{a}} &= \mathrm{diag}_x F_x(\Psi(S_L)), \\ \dot{\mathbf{b}} &= \mathrm{diag}\, F_h(\Psi(S_L)), \\ \dot{\mathbf{c}} &= \mathbf{f}_c(\Psi(S_L)), \\ (\dot{U}, \dot{A}, \dot{B}) &= \mathcal{J}_{S_L}^\dagger \left(\mathcal{P}_x F_x(\Psi(S_L)), \mathcal{P}_h F_h(\Psi(S_L))\right). \end{aligned} \tag{79}$$

The pseudoinverse determines the factor velocities that best approximate the off-diagonal learning field in the least-squares sense. In underconstrained settings, when several factor velocities would produce the same matrix changes, we select the solution with minimum norm $\|\Omega\|_F^2 + \|C_A\|_F^2 + \|C_B\|_F^2$, subject to $U^\top \Omega = 0$.

**Coupled normal equations.** To determine the evolution of the interaction factors, we minimize the difference between the learning field and the weight changes generated by the reduced representation. We define the residual matrices

$$E_x = \mathcal{P}_x(F_x - \dot{U}A^\top - U\dot{A}^\top), \qquad E_h = \mathcal{P}_h(F_h - \dot{U}B^\top - U\dot{B}^\top). \tag{80}$$

These residuals describe the components of the input and recurrent learning fields that are not reproduced by the factor velocities.

At the least-squares minimum, the residuals must be orthogonal to all admissible changes of the factors. Since the projection operators are self-adjoint under the Frobenius inner product, this results in

$$E_x^\top U = 0, \qquad E_h^\top U = 0, \qquad (I - UU^\top)(E_x A + E_h B) = 0, \qquad U^\top \dot{U} = 0. \tag{81}$$

The first two conditions determine the updates of the right factors $A$ and $B$, while the third determines the evolution of their shared left factor $U$. Because changes in $U$ affect both weight matrices, these equations must be solved jointly.

The projected equations form an autonomous dynamical system in which the current reduced state determines the network weights, the population learning field, and its projection onto the admissible weight changes. Thus, the reduced dynamics are evolved without reference to the full system learning trajectory.

The projection uniquely specifies the instantaneous changes in the represented weight matrices, but the corresponding factor velocities are not necessarily unique. We resolve this ambiguity by using the pseudoinverse to select the minimum-norm solution. For the factors to evolve smoothly, the tangent map must maintain constant rank. In particular, when $A = B = 0$, variations in $U$ have no first-order effect on the weight matrices. Therefore, a positive-rank model with initially vanishing interaction factors requires either a prescribed initial basis $U(0)$ or an explicit rank-adaptive initialisation procedure. Importantly, the products $UA^\top$ and $UB^\top$ have rank at most $L$ before projection, but their off-diagonal parts can have larger rank after the projection.

### B.4.4 Gaussian evaluation of the population learning field

**Weight derivatives and their correlations with the output trajectories.** As stated above, the covariance matrix $\Sigma$ characterises the joint Gaussian distribution of the student and teacher outputs, and therefore determines the probability of receiving a reward. To evaluate the learning dynamics, however, we also need the correlations between these outputs and their derivatives with respect to the student weights. These correlations determine how the different weight parameters contribute to the reward-modulated update.

We consider a scalar entry $w$ of the full-system student weight matrices or readout vector. By differentiating the response representation of $\lambda_t$ results in

$$e_{t,w} = \frac{\partial \lambda_t}{\partial w} = \sum_{j=1}^{t} \frac{\partial \boldsymbol{\varphi}_{t-j}}{\partial w}\, \mathbf{x}_j. \tag{82}$$

Here we take the derivative with respect to the original weight parameter, that is, before imposing the reduced connectivity representation.

At fixed network weights, both $e_{t,w}$ and the output trajectory vector $Z$ are linear functions of the same Gaussian input sequence. They are therefore jointly Gaussian. To characterize their correlations, we introduce the cross-covariance row vector

$$C_{t,w} = \mathrm{Cov}(e_{t,w}, Z \mid \theta) \in \mathbb{R}^{1\times 2T}, \tag{83}$$

and by exploiting the independence of the inputs at different time steps, we express its entries in terms of the impulse responses

$$(C_{t,w})_v = \sum_{j=1}^{\min(t,v)} \frac{\partial \boldsymbol{\varphi}_{t-j}}{\partial w} \boldsymbol{\varphi}_{v-j}^\top, \tag{84}$$

$$(C_{t,w})_{T+v} = \sum_{j=1}^{\min(t,v)} \frac{\partial \boldsymbol{\varphi}_{t-j}}{\partial w} (\boldsymbol{\varphi}^*_{v-j})^\top. \tag{85}$$

The first expression describes correlations with the student outputs, while the second describes correlations with the teacher outputs. Of course, we differentiate only the student responses, since the teacher remains fixed over learning.

These quantities are different from the derivatives of the output covariance matrix. For example, differentiating the student output variance results in

$$\frac{\partial \Lambda_{tt}}{\partial w} = 2(C_{t,w})_t, \tag{86}$$

because the derivative acts on both factors in $\Lambda_{tt} = \mathbb{E}[\lambda_t^2 \mid \theta]$. Thus, although the output covariance determines the reward statistics, the cross-covariances $C_{t,w}$ provide the additional information required to evaluate the learning field. They are computed entirely from the current network weights and their response derivatives.

**Gaussian moments associated with successful episodes.** We first consider the contribution to learning from episodes in which the student agrees with the teacher at every time step. This event corresponds to the agreement region

$$\mathcal{A}_T = \left\{ z \in \mathbb{R}^{2T} : z_t z_{T+t} > 0, \quad t = 1, \dots, T \right\}, \tag{87}$$

with the associated success indicator $I_T$.

The contribution to the update of a parameter $w$ involves the expectation $\langle I_T y_t e_{t,w} \mid \theta \rangle$. To directly evaluate this expectation would require integrating over both the output trajectories and their weight derivatives. However, the joint Gaussian property allows us to eliminate the latter.

We first introduce the vector that can be understood as the the action-weighted first moment over successful trajectories

$$\mathbf{v}_t(\Sigma) = \left\langle Z I_T y_t \mid \theta \right\rangle = \int_{\mathcal{A}_T} z \, \operatorname{sgn}(z_t) \, \mathrm{d}\gamma_\Sigma(z), \tag{88}$$

where $\gamma_\Sigma$ denotes the centered Gaussian measure with covariance $\Sigma$.

Since $(e_{t,w}, Z)$ are jointly Gaussian, their conditional mean is linear in $Z$

$$\langle e_{t,w} \mid Z, \theta \rangle = C_{t,w} \Sigma^\dagger Z. \tag{89}$$

Here $\Sigma^\dagger$ denotes the Moore–Penrose pseudoinverse, which also covers singular covariance matrices. For nonsingular $\Sigma$, it reduces to the ordinary inverse.

Taking the expectation over the successful trajectories [Marsaglia, 1964] therefore gives

$$\begin{aligned} \langle \mathcal{I}_T y_t e_{t,w} \mid \theta \rangle &= \langle \mathcal{I}_T y_t \, \langle e_{t,w} \mid Z, \theta \rangle \mid \theta \rangle \\ &= C_{t,w} \Sigma^\dagger \langle Z \mathcal{I}_T y_t \mid \theta \rangle \\ &= C_{t,w} \Sigma^\dagger \mathbf{v}_t(\Sigma). \end{aligned} \tag{90}$$

This expression separates the two ingredients of the reward-modulated update. The vector $\mathbf{v}_t(\Sigma)$ captures the statistics of the trajectories satisfying the agreement condition, while $C_{t,w}$ determines how a particular weight parameter contributes to the output trajectory. Importantly, the calculation uses the original Gaussian distribution restricted to the agreement region; the distribution conditioned on successful episodes is generally not itself Gaussian.

**Complete population learning field.** The reward contains both a contribution from successful episodes and a constant penalty:

$$G = (r_1 + r_2) I_T - r_2.$$

Consequently, the population learning field contains two expectations. The first has been obtained above. For the second, Gaussian conditioning on the scalar output $\lambda_t$ yields

$$\mathbb{E}[y_t e_{t,w} \mid \theta] = \sqrt{\frac{2}{\pi}} \, \frac{(C_{t,w})_t}{\sqrt{\Lambda_{tt}}}, \tag{91}$$

provided $\Lambda_{tt} > 0$.

Combining these results, the population learning field for each scalar weight parameter $w$ becomes

$$F_w = \frac{1}{T} \sum_{t=1}^{T} \left[ (r_1 + r_2) C_{t,w} \Sigma^\dagger \mathbf{v}_t(\Sigma) - r_2 \sqrt{\frac{2}{\pi}} \, \frac{(C_{t,w})_t}{\sqrt{\Lambda_{tt}}} \right]. \tag{92}$$

The first term describes the update associated with successful episodes, while the second arises from the constant reward penalty. For the learning rule we consider here, the latter is generally nonzero because the output sign and its weight derivative are correlated.

Eq. 92 determines every component of the full population learningfields $F_x$, $F_h$, and $\mathbf{f}_c$, and by evaluating these fields at the weights specified by the reduced representation and applying the tangent-space projection results in the reduced learning equations derived above.

The remaining quantities are Gaussian expectations over the output trajectories. Their numerical evaluation, by quadrature or sampling, is separate from the approximation we introduced by the reduced connectivity representation.

### B.4.5 Learning observables and consistency checks

**Performance from the output correlations.** The reduced learning equations describe the evolution of the network weights. To relate these dynamics to task performance, we reconstruct the student responses at each reduced state and employ them to compute the joint student-teacher covariance matrix.

The correlation between the student and teacher outputs at time $t$ is

$$\rho_t = \frac{M_{tt}}{\sqrt{\Lambda_{tt} N_{tt}}}. \tag{93}$$

For jointly Gaussian outputs with nonzero variances, the probability that their signs agree is

$$p_t = \mathbb{P}(y_t = y_t^* \mid \theta) = \frac{1}{2} + \frac{1}{\pi} \arcsin \rho_t. \tag{94}$$

We use the mean agreement probability

$$\bar{p} = \frac{1}{T} \sum_{t=1}^{T} p_t \tag{95}$$

as a measure of average prediction accuracy within an episode.

Notice that these marginal probabilities do not, in general, determine the probability of success over the entire episode. Since recurrence introduces correlations between outputs at different time points, the probability of agreement at every step must be evaluated from their joint distribution.

**Response alignment and episode success.** We also monitor the alignment between the student and teacher responses at zero time-lag

$$\rho_0 = \frac{\boldsymbol{\varphi}_0 (\boldsymbol{\varphi}_0^*)^\top}{\|\boldsymbol{\varphi}_0\| \|\boldsymbol{\varphi}_0^*\|}. \tag{96}$$

With zero initial recurrent states, the output at the first time step depends only on the current input, so that $\rho_0 = \rho_1$. The subscript in $\rho_0$ denotes the response lag, rather than learning time.

The episode success probability is

$$P_T = \mathbb{P}(\mathcal{I}_T = 1 \mid \theta). \tag{97}$$

For $P_{T-1} > 0$, the ratio

$$\frac{P_T}{P_{T-1}}$$

gives the conditional probability of agreement at the final step, given agreement at all preceding steps. We evaluate both probabilities on the same network weights. For a finite episode length, this ratio is not necessarily equal to an asymptotic persistence eigenvalue.

To quantify the escape from an initially low-performance regime, we specify an observable $\mathcal{O}$ and a threshold $\mathcal{O}_{\mathrm{th}}$, and define the **escape time** as

$$\tau = \inf \{ \varsigma : \mathcal{O}(S_L(\varsigma)) \geq \mathcal{O}_{\mathrm{th}} \}. \tag{98}$$

If the threshold is never reached, we set $\tau = \infty$. This definition allows us to compare escape times consistently across the full and the reduced models.

**Diagonal observables and their relation to the scalar theory.** The diagonal components of the reduced representation also allow us to keep track of the quantities used in the $L = 0$ family of models, that admits a mode-wise description. For each directly driven coordinate $m = 1, \ldots, r$, we define

$$g_m = a_m c_m, \qquad Q_m = g_m^2, \qquad R_m = g_m g_m^*, \qquad \Delta_m = c_m^2 - a_m^2, \tag{99}$$

where

$$g_m^* = (W_x^*)_{mm} c_m^*$$

is a fixed teacher quantity defined in the chosen coordinates.

The product $g_m$ quantifies the contribution of the diagonal input coefficient and readout weight to the output. The quantities $Q_m$ and $R_m$ respectively describe its squared magnitude and its overlap with the corresponding teacher reference coefficient, while $\Delta_m$ measures the difference between the squared readout and input weights.

These quantities provide a direct connection with the diagonal theories encountered in supervised settings Proca et al. [2025], Williams et al. [2026], but their interpretation changes when the factorised interactions are present. The full response vectors then depend on both the diagonal and factorised components, and not on the diagonal elements of the matrix that would constitute its eigenvalue. Moreover, when the teacher is not diagonal, $g_m^*$ is only a coordinate-specific reference quantity, rather than the complete teacher response to input coordinate $m$.

The diagonal quantities therefore are useful observables, but they do not form an autonomous set of order parameters under the considered reward-driven learning update.

Their evolution follows directly from the projected learning equations

$$\begin{aligned}
\dot{g}_m &= c_m (F_x)_{mm} + a_m (\mathbf{f}_c)_m, \\
\dot{Q}_m &= 2 g_m \dot{g}_m, \\
\dot{R}_m &= g_m^* \dot{g}_m, \\
\dot{\Delta}_m &= 2 c_m (\mathbf{f}_c)_m - 2 a_m (F_x)_{mm}.
\end{aligned} \tag{100}$$

These expressions allow the diagonal observables to be monitored alongside the full response statistics. At positive factor rank, $\Delta_m$ does not remain conserved.

**Recovery of the diagonal model.** We can recover a diagonal description akin to Saxe et al. [2013], Williams et al. [2026] as the $L = 0$ member of the reduced representation. To make the connection explicit, we consider first $N_x \geq N_h$, so that

$$W_x = \mathcal{D}_x(\mathbf{a}), \qquad W_h = \mathrm{Diag}(\mathbf{b}).$$

The output pre-activation then separates into independent input-history contributions,

$$\lambda_t = \sum_{m=1}^{N_h} g_m \zeta_t^m, \qquad \zeta_t^m = \sum_{j=1}^{t} b_m^{t-j} x_{j,m}. \tag{101}$$

The derivative of each input-history term with respect to the recurrent coefficient is

$$\dot{\zeta}_t^m = \frac{\partial \zeta_t^m}{\partial b_m} = \sum_{j=1}^{t-1} (t-j) b_m^{t-j-1} x_{j,m}. \tag{102}$$

Consequently, the derivatives of the output with respect to the diagonal parameters are

$$\frac{\partial \lambda_t}{\partial a_m} = c_m \zeta_t^m, \qquad \frac{\partial \lambda_t}{\partial c_m} = a_m \zeta_t^m, \qquad \frac{\partial \lambda_t}{\partial b_m} = a_m c_m \dot{\zeta}_t^m. \tag{103}$$

By introducing the reward-weighted moments

$$E_m = \left\langle G \sum_{t=1}^{T} y_t \zeta_t^m \right\rangle, \qquad J_m = \left\langle G \sum_{t=1}^{T} y_t \dot{\zeta}_t^m \right\rangle, \tag{104}$$

we obtain

$$\dot{a}_m = \frac{c_m}{T} E_m, \qquad \dot{c}_m = \frac{a_m}{T} E_m, \qquad \dot{b}_m = \frac{a_m c_m}{T} J_m. \tag{105}$$

An immediate consequence is that

$$\frac{d}{d\varsigma} (c_m^2 - a_m^2) = 2 c_m \dot{c}_m - 2 a_m \dot{a}_m = 0. \tag{106}$$

Thus, the quantity $\Delta_m$ is conserved in the diagonal model.

For the particular initialization $a_m = c_m$, this equality is preserved by the dynamics. In this case $g_m = a_m^2 = \sqrt{Q_m}$, and the equations reduce further to

$$\dot{Q}_m = \frac{4Q_m}{T} E_m, \qquad \dot{R}_m = \frac{2R_m}{T} E_m, \qquad \dot{b}_m = \frac{\sqrt{Q_m}}{T} J_m. \tag{107}$$

The condition $\Delta_m = 0$ alone is not sufficient to select this branch, since it also allows $a_m = -c_m$. More generally, the signed gain $g_m = a_m c_m$ must be retained.

This establishes the diagonal equations as a limiting case of our considered projected model. However, under the considered linearised REINFORCE update the diagonal description is not preserved over learning.

**Feedforward limit.** We can evaluate our theory when both the student and teacher recurrent matrices vanish, $W_h = W_h^* = 0$, which is similar in vein to Patel et al. [2025]. Their outputs then depend only on the current inputs

$$\lambda_t = \mathbf{c}^\top W_x \mathbf{x}_t, \qquad \nu_t = (\mathbf{c}^*)^\top W_x^* \mathbf{x}_t.$$

Since the inputs are independent across time, the teacher-student output pairs are also independent across time. The agreement probability is constant, and the episode success probability factorises similar to Patel et al. [2025]

$$P_T = \mathfrak{p}^T, \qquad \mathfrak{p} = \frac{1}{2} + \frac{1}{\pi} \arcsin \rho_0. \tag{108}$$

This simplification of the output statistics does not imply that the derivatives with respect to recurrent weights vanish. Indeed, for $t \geq 2$,

$$\nabla_{W_h} \lambda_t = \mathbf{c}(W_x \mathbf{x}_{t-1})^\top. \tag{109}$$

The derivative describes the effect of introducing a recurrent connection, even when the current recurrent matrix is zero.

Nevertheless, when both networks are feedforward, the corresponding population update vanishes. The agreement event is invariant under a simultaneous sign reversal of the student and teacher outputs within any one time step, whereas the preceding input changes sign. Averaging over the independent Gaussian inputs therefore cancels the recurrent update.

Thus, the feedforward limit provides a check on the reward statistics and the population learning field.

**Assessing the accuracy of the reduced dynamics.** Finally, we evaluate the projected model against the full network learning dynamics.

At any reduced state, the residual matrices $E_x$ and $E_h$ defined in Eq. (80) quantify the components of the population learning field that cannot be reproduced by the admissible factor velocities. We define a normalised projection residual as

$$\epsilon_{\text{proj}} = \frac{\left(\|E_x\|_F^2 + \|E_h\|_F^2\right)^{1/2}}{\left(\|F_x\|_F^2 + \|F_h\|_F^2 + \|\mathbf{f}_c\|_2^2\right)^{1/2}}, \tag{110}$$

whenever the denominator is nonzero. This quantifies the fraction of the instantaneous population learning field discarded (not accounted for in) by the projection.

We therefore compare independently integrated reduced and unrestricted trajectories using the full response vectors, the agreement probabilities $\bar{p}$ and $P_T$, and escape times defined by the same threshold criterion.

Approximating the full-model weight matrices at selected learning times quantifies how well the reduced representation can describe an existing trajectory. On the other hand, integrating the projected equations and comparing their predictions with unrestricted learning tests whether the reduced dynamics provide an accurate autonomous description.

### B.5 Invariant balances and the diagonal sign constraint

**A conserved matrix of the unrestricted flow.** For an invertible hidden-coordinate change,

$$(W_x, W_h, \mathbf{c}) \longmapsto (SW_x, SW_hS^{-1}, S^{-\top}\mathbf{c}), \tag{111}$$

every episode's preactivations are unchanged. We differentiate this invariance at $S = I + \epsilon K$ and weight the resulting identity by the episode coefficients of the linearised update, treating these coefficients as constant during differentiation. Because $K$ is arbitrary, the episode field and its population average satisfy

$$W_xF_x^\top + W_hF_h^\top - F_h^\top W_h - \mathbf{f}_c\mathbf{c}^\top = 0. \tag{112}$$

Adding its transpose gives

$$\mathcal{M} = W_xW_x^\top + W_hW_h^\top - W_h^\top W_h - \mathbf{c}\mathbf{c}^\top, \qquad \dot{\mathcal{M}} = 0. \tag{113}$$

This conservation law follows from output invariance and the use of a common Euclidean learning metric, without requiring $F$ to be the gradient of expected reward. It also holds for finite-batch updates computed from exact output derivatives. However, a finite Euler step introduces an $O(\eta^2)$ change in the conserved balance. Using different learning rates across weight blocks, or introducing regularisation, clipping, or inexact projection, can also break these identities.

The projected shared-subspace dynamics do not necessarily conserve the full matrix $\mathcal{M}$. However, the represented family remains invariant under diagonal rescalings of the hidden coordinates, whose generators lie in the admissible tangent space. Because the exact projection residual is orthogonal to these directions, the projected dynamics conserve $\operatorname{diag}\mathcal{M}$ and therefore $\operatorname{tr}\mathcal{M} = |W_x|F^2 - |\mathbf{c}|^2$. At positive factor rank, this does not imply conservation of each $c_m^2 - a_m^2$, since the diagonal entries $\mathcal{M}mm$ also depend on off-diagonal row and column norms.

**The balanced diagonal sign constraint.** For $L = 0$, the recurrent connectivity is diagonal and $\mathcal{M}_{mm} = a_m^2 - c_m^2 = -\Delta_m$ on directly driven coordinates. The scalar equations already given for this family conserve $\Delta_m$. On the positive balanced branch $a_m = c_m$, they further preserve

$$g_m = a_mc_m = a_m^2 \geq 0. \tag{114}$$

The boundary $a_m = c_m = 0$ is invariant, and independent diagonal retuning cannot produce a negative gain from this branch. If the teacher gain $g^*$ has at least one positive entry, the zero-lag alignment consequently satisfies

$$\rho_0 \leq \max_{g\geq 0,\, g\neq 0} \frac{g^\top g^*}{\|g\|\|g^*\|} = \frac{\|(g^*)_+\|}{\|g^*\|}. \tag{115}$$

The bound follows from $g^\top g^* \leq g^\top (g^*)_+$ and Cauchy–Schwarz, with equality attained for $g$ proportional to $(g^*)_+$. For $g^* = (1, -1)^\top$ this gives $1/\sqrt{2}$. It is a constraint on this balanced diagonal branch, not on every diagonal initialisation. The benchmark threshold $\rho_0 = 0.5$ lies below this ceiling. Allowing unrestricted input directions removes this particular nonnegative-gain constraint, but does not by itself prove convergence to perfect imitation.

### B.6 Response-equivalent networks and a matrix-state transformation

The invariance in Eq. 111 preserves the response vectors, reward statistics, and action sequences for any given observation history. However, networks with identical responses do not necessarily follow the same learning trajectories. Indeed, the full network population field at the transformed state satisfies

$$F_x' = S^{-\top}F_x, \qquad F_h' = S^{-\top}F_hS^\top, \qquad f_c' = Sf_c. \tag{116}$$

These inverse-transpose transformations generally differ from the pushforward of weight velocities under Eq. 111, with equivariance guaranteed only for orthogonal transformations $S$. Thus, networks with identical input–output mappings can have different output derivatives and follow different learning trajectories. This distinction explains why the population learning field must account for cross-covariances between outputs and their weight derivatives.

### B.7 The quadratic correction in response coordinates

We consider the common scaling $W_h = \epsilon\overline{W}_h$ and $W_h^* = \epsilon\overline{W}_h^*$, with fixed input maps and readouts. Then $\boldsymbol{\varphi}_k = O(\epsilon^k)$ and likewise for the teacher. Let $q = \|\boldsymbol{\varphi}_0\|^2 > 0$, $q_* = \|\boldsymbol{\varphi}_0^*\|^2 > 0$, and $\rho_0 = \boldsymbol{\varphi}_0(\boldsymbol{\varphi}_0^*)^\top/\sqrt{qq_*}$, with $|\rho_0| < 1$. We define the dominant adjacent-time correlations

$$\begin{aligned} \kappa_1 &= \frac{\boldsymbol{\varphi}_0\boldsymbol{\varphi}_1^\top}{q}, \quad \kappa_2 = \frac{\boldsymbol{\varphi}_0(\boldsymbol{\varphi}_1^*)^\top}{\sqrt{qq_*}}, \\ \kappa_3 &= \frac{\boldsymbol{\varphi}_0^*\boldsymbol{\varphi}_1^\top}{\sqrt{qq_*}}, \quad \kappa_4 = \frac{\boldsymbol{\varphi}_0^*(\boldsymbol{\varphi}_1^*)^\top}{q_*}. \end{aligned} \tag{117}$$

These are the $O(\epsilon)$ parts of the correlations of $(\lambda_t, \nu_t)$ with $(\lambda_{t+1}, \nu_{t+1})$, in that order. The $O(\epsilon^2)$ within-time correlation shift is

$$\delta\rho = \frac{\boldsymbol{\varphi}_1(\boldsymbol{\varphi}_1^*)^\top}{\sqrt{qq_*}} - \frac{\rho_0}{2}\left(\frac{\|\boldsymbol{\varphi}_1\|^2}{q} + \frac{\|\boldsymbol{\varphi}_1^*\|^2}{q_*}\right). \tag{118}$$

For zero resets this shift occurs at each $t \geq 2$ to the order considered, while there is no shift at $t = 1$. Thhis shift does not necessarily vanish when $\rho_0 = 0$.

The four-variable Gaussian expansion used in the original calculation now results in

$$\mathcal{J} = \frac{\mathfrak{p}\,\delta\rho}{\pi\sqrt{1-\rho_0^2}} + \frac{\mathcal{F}(\boldsymbol{\kappa}, \rho_0)}{\pi^2(1-\rho_0^2)}, \qquad \mathfrak{p} = \frac{1}{2} + \frac{\arcsin\rho_0}{\pi}, \tag{119}$$

where

$$\mathcal{F}(\boldsymbol{\kappa}, \rho) = \frac{\rho^2}{2}\sum_{i=1}^{4}\kappa_i^2 + \kappa_1\kappa_4 + \kappa_2\kappa_3 - \rho(\kappa_1\kappa_2 + \kappa_1\kappa_3 + \kappa_2\kappa_4 + \kappa_3\kappa_4). \tag{120}$$

**Extension to a finite horizon.** At zero recurrence the time blocks are independent. Their agreement indicators are even under simultaneous reversal of the two variables within each block. The first covariance correction links neighboring blocks, so its first-order contribution vanishes. At second order, two distinct links leave at least one endpoint block differentiated an odd number of times, and their integrals therefore vanish, including the three-time chain terms. A direct lag-two covariance is also $O(\epsilon^2)$, but contributes only a vanishing first-order cross-block integral. The surviving terms are the square of each adjacent-time perturbation and the within-time shift at $t = 2, \ldots, T$. There are $T - 1$ of each, with the same coefficients to this order, giving

$$P_T = \mathfrak{p}^T + (T-1)\mathfrak{p}^{T-2}\mathcal{J} + O_T(\epsilon^4). \tag{121}$$

The absence of odd orders follows exactly from $\mathbf{x}_t \mapsto (-1)^t\mathbf{x}_t$ and simultaneous reversal of both recurrent matrices. The estimate is a fixed-$T$ expansion. It is not a uniform large-horizon remainder bound, and the marked learning moments have their own parameter derivatives and parity properties.

**Special cases and scope.** When $W_h = bI$ and $W_h^* = b^*I$, the coefficients are $(\kappa_1, \kappa_2, \kappa_3, \kappa_4) = (b, \rho_0 b^*, \rho_0 b, b^*)$ and $\delta\rho = -\rho_0(b - b^*)^2/2$. For $b = b^*$, $\mathcal{J} = (1 - \rho_0^2)b^2/\pi^2$. For $\rho_0 = 0$, $\mathcal{J} = bb^*/\pi^2$. These identities do not imply a general sign rule for arbitrary recurrent connectivity. In general, both the changes in same-time correlations and the correlations between successive time steps contribute to the correction. Moreover, for nonnormal recurrent matrices, small eigenvalues alone are not sufficient to justify the perturbative expansion, which instead relies on the assumed common scaling of the recurrent matrices.

## C The transfer operator construction for the persistence rate

The finite-horizon analysis above describes the episode-success probability $P_T$ through a Gaussian trajectory integral and its expansion in the weak-recurrence regime. We now take a different approach, expressing the same probability as the survival probability of a killed Gaussian Markov process [Nardo, 2005, Maoutsa and Opper, 2022, 2021, Azzimonti and Ginsbourger, 2018]. This formulation applies to arbitrary student and teacher connectivity matrices.

Throughout this section, we keep the network parameters fixed, so that the transfer operator describes the trajectory statistics of a given teacher-student pair, independently of how their connectivity evolves during learning.

### C.1 Joint recurrent-state process and survival representation

**Joint Markov state.** We collect the student and teacher recurrent states associated with the output at time $t$ into

$$\boldsymbol{\xi}_t = \begin{pmatrix} \mathbf{h}_{t+1} \\ \mathbf{h}^*_{t+1} \end{pmatrix} \in \mathbb{R}^{2N_h}. \tag{122}$$

Using the recurrent dynamics, we have

$$\boldsymbol{\xi}_{t+1} = \mathsf{A}\boldsymbol{\xi}_t + \Gamma \mathbf{x}_{t+1}, \qquad \mathsf{A} = \begin{pmatrix} W_h & 0 \\ 0 & W_h^* \end{pmatrix}, \qquad \Gamma = \begin{pmatrix} W_x \\ W_x^* \end{pmatrix}, \tag{123}$$

where $\mathbf{x}_{t+1} \sim \mathcal{N}(0, I_{N_x})$ independently across time. Since both networks are reset at the beginning of the episode,

$$\boldsymbol{\xi}_1 = \Gamma \mathbf{x}_1, \qquad \pi_1 = \mathcal{N}(0, \Gamma\Gamma^\top) \tag{124}$$

on its support.

The student and teacher output pre-activations are linear functions of this joint state

$$\lambda(\boldsymbol{\xi}) = \begin{pmatrix} \mathbf{c} \\ 0 \end{pmatrix}^\top \boldsymbol{\xi}, \qquad \nu(\boldsymbol{\xi}) = \begin{pmatrix} 0 \\ \mathbf{c}^* \end{pmatrix}^\top \boldsymbol{\xi}. \tag{125}$$

We define the active-set indicator (not killed)

$$\chi(\boldsymbol{\xi}) = \mathbf{1}\{\lambda(\boldsymbol{\xi})\nu(\boldsymbol{\xi}) > 0\}, \tag{126}$$

which characterises whether student and teacher agree in sign at the current time.

We write then the full-episode success indicator as a product of active set indicators over the $T$ timesteps

$$\mathcal{I}_T = \prod_{t=1}^{T} \chi(\boldsymbol{\xi}_t). \tag{127}$$

Consequently, the probability of success of the episode with length $T$ is expressed as

$$P_T = \mathbb{P}[\boldsymbol{\xi}_t \in \mathcal{A} \text{ for every } t = 1, \dots, T], \qquad \mathcal{A} = \{\boldsymbol{\xi} : \lambda(\boldsymbol{\xi})\nu(\boldsymbol{\xi}) > 0\}. \tag{128}$$

**Thus episode success is the survival probability of the joint Gaussian Markov process in the agreement region** $\mathcal{A}$. The process is killed when it enters $\mathcal{A}^{\mathrm{c}}$.

### C.2 Killed transfer operator

We denote with $\mathcal{L}$ the free (unconstrained) Markov propagator,

$$(\mathcal{L}f)(\xi) = \mathbb{E}_{\mathbf{x}}\left[f(\mathsf{A}\xi + \Gamma\mathbf{x})\right], \qquad \mathbf{x} \sim \mathcal{N}(0, I_{N_x}), \tag{129}$$

and let $P_\chi$ denote the multiplication by the agreement indicator $\chi$

$$(P_\chi f)(\xi) = \chi(\boldsymbol{\xi}) f(\boldsymbol{\xi}). \tag{130}$$

The backward killed transfer operator is

$$\mathcal{T} = P_\chi \mathcal{L}, \qquad (\mathcal{T}f)(\boldsymbol{\xi}) = \chi(\boldsymbol{\xi}) \Big\langle f(\mathsf{A}\boldsymbol{\xi} + \Gamma\mathbf{x}) \Big\rangle_{\mathbf{x}}. \tag{131}$$

To obtain the finite-horizon survival probability for the episode of length $T$, we define the backward survival function

$$v_k(\boldsymbol{\xi}) = \Big\langle \prod_{t=k}^{T} \chi(\boldsymbol{\xi}_t) \,\Big|\, \boldsymbol{\xi}_k = \boldsymbol{\xi} \Big\rangle_{\mathbf{x}}. \tag{132}$$

At the terminal time we have $v_T = \chi$, while from the the Markov property we have

$$v_k = \mathcal{T} v_{k+1}. \tag{133}$$

Therefore, we can re-express the backward survival funciton in terms of repeated applications of the killed operator

$$v_1 = \mathcal{T}^{T-1}\chi, \tag{134}$$

and then averaging over the reset distribution results in the identity

$$P_T = \langle \pi_1, \mathcal{T}^{T-1}\chi \rangle . \tag{135}$$

Eq. (135) holds at every finite $T$ without requiring any explicit any stability assumption. It is simply a reformulation of the same trajectory orthant probability appearing in the Gaussian description of the outputs.

### C.3 Reachable state space and stationary dynamics

Because student and teacher are driven by the same observation, the one-step transition is generally singular in the ambient $2N_h$-dimensional state space. The appropriate state space is therefore the reachable subspace

$$\mathcal{H} = \operatorname{span}\left\{ \mathsf{A}^k \operatorname{range}(\Gamma) : k \geq 0 \right\}. \tag{136}$$

All subsequent spectral statements are understood as defined on $\mathcal{H}$.

The covariance accumulated over $\ell$ transition steps is the controllability Gramian

$$W_\ell = \sum_{k=0}^{\ell-1} \mathsf{A}^k \Gamma\Gamma^\top (\mathsf{A}^\top)^k. \tag{137}$$

We denote by $\ell_c$ a finite controllability index for which $W_{\ell_c}$ is positive definite on $\mathcal{H}$.

We assume now that the free joint dynamics are stable on the reachable space,

$$\rho(\mathsf{A}|_{\mathcal{H}}) < 1, \tag{138}$$

equivalently $\rho(W_h) < 1$ and $\rho(W_h^*) < 1$ after restricting to dynamically relevant directions. The free process then has a unique stationary Gaussian law on $\mathcal{H}$,

$$\pi = \mathcal{N}(0, \Sigma_\infty), \tag{139}$$

whose covariance satisfies

$$\Sigma_\infty = \mathsf{A}\Sigma_\infty \mathsf{A}^\top + \Gamma\Gamma^\top. \tag{140}$$

For any $\ell$,

$$\Sigma_\infty = \mathsf{A}^\ell \Sigma_\infty (\mathsf{A}^\top)^\ell + W_\ell. \tag{141}$$

Whitening on $\mathcal{H}$ gives

$$C_\ell = \Sigma_\infty^{-1/2} \mathsf{A}^\ell \Sigma_\infty^{1/2}, \tag{142}$$

and therefore

$$I - C_\ell C_\ell^\top = \Sigma_\infty^{-1/2} W_\ell \Sigma_\infty^{-1/2}. \tag{143}$$

For $\ell = \ell_c$, the right-hand side is positive definite, so all singular values of $C_{\ell_c}$ are strictly below one.

This finite-step smoothing is the operator-theoretic counterpart of controllability, implying that although the one-step transition may be degenerate, after $\ell_c$ steps the free Gaussian transition has a nondegenerate density on the reachable space.

### C.4 Persistence rate

We now state the additional spectral conditions needed to obtain a single asymptotic persistence rate. Power compactness of the transition operator alone does not guarantee a unique dominant surviving mode. We therefore assume that, on the alive part of $\mathcal{H}$, the killed process satisfies the usual positivity and mixing conditions required for a simple isolated dominant eigenvalue [Collet et al., 2012, Champagnat and Villemonais, 2016].

Under these conditions, $\mathcal{T}$ has a real positive eigenvalue $\mu$ equal to its spectral radius, separated in modulus from the remaining spectrum

$$\mathcal{T}\psi = \mu\psi, \qquad \mathcal{T}^\dagger \widetilde{\psi} = \mu\widetilde{\psi}. \tag{144}$$

The corresponding right and left eigenobjects can be chosen positive on the surviving state space.

The reset distribution $\pi_1$ can be singular with respect to the stationary law $\pi$. To avoid applying an $L^2(\pi)$ spectral representation directly to this singular measure, we choose $m \geq \ell_c$ and absorb the first $m-1$ survival steps into the finite measure

$$\nu_m(B) = \left\langle \mathbf{1}\{\boldsymbol{\xi}_m \in B\} \prod_{t=1}^{m-1} \chi(\boldsymbol{\xi}_t) \right\rangle. \tag{145}$$

Its free Gaussian envelope has covariance $W_m$, which is nondegenerate on $\mathcal{H}$. For $T \geq m$,

$$P_T = \left\langle \nu_m, \mathcal{T}^{T-m} \chi \right\rangle. \tag{146}$$

The dominant spectral mode then yields

$$P_T = C\mu^{T-1}\left[1 + o(1)\right], \qquad T \to \infty, \tag{147}$$

where the finite initial segment is absorbed into the positive prefactor $C$. Equivalently,

$$\mu = \lim_{T\to\infty} \frac{P_{T+1}}{P_T}, \tag{148}$$

whenever the stated dominant-eigenvalue conditions hold.

Thus $\mu$ is the exponential persistence rate of agreement. It describes the large-horizon statistics of a fixed pair of networks and should not be identified with a finite-horizon conditional probability unless the latter has converged to this limit.

**Feedforward limit.** In the absence of recurrent connections, i.e. when

$$W_h = W_h^* = 0, \tag{149}$$

the recurrent state at each time depends only on the current observation. Success events at different times are therefore independent and the episode success probability

$$P_T = \mathfrak{p}^T, \qquad \mathfrak{p} = \frac{1}{2} + \frac{\arcsin \rho_0}{\pi}, \tag{150}$$

where

$$\rho_0 = \frac{\boldsymbol{\varphi}_0 (\boldsymbol{\varphi}_0^*)^\top}{\|\boldsymbol{\varphi}_0\| \|\boldsymbol{\varphi}_0^*\|}. \tag{151}$$

The killed operator becomes rank one and its unique nonzero eigenvalue is

$$\mu_0 = \mathfrak{p}. \tag{152}$$

Hence the recurrent persistence rate continuously generalizes the feedforward one-step agreement probability.

### C.5 Weak recurrence and connection to the quadratic correction

We consider the common small-recurrence scaling

$$W_h = \epsilon \overline{W}_h, \qquad W_h^* = \epsilon \overline{W}_h^*, \qquad |\epsilon| \ll 1, \tag{153}$$

with $W_x, W_x^*, \mathbf{c}, \mathbf{c}^*$ fixed and $|\rho_0| < 1$.

The Gaussian trajectory probability is invariant under simultaneous sign reversal of both recurrent matrices. Indeed, the measure-preserving transformation

$$\mathbf{x}_t \longmapsto (-1)^t \mathbf{x}_t \tag{154}$$

implies

$$\lambda_t(W_h) \longmapsto (-1)^t \lambda_t(-W_h), \qquad \nu_t(W_h^*) \longmapsto (-1)^t \nu_t(-W_h^*), \tag{155}$$

so the agreement products are unchanged. Therefore,

$$P_T(W_h, W_h^*) = P_T(-W_h, -W_h^*) \tag{156}$$

for every finite $T$. Whenever the dominant eigenvalue varies regularly,

$$\mu(W_h, W_h^*) = \mu(-W_h, -W_h^*). \tag{157}$$

All odd powers of the common perturbation parameter therefore vanish.

The finite-horizon calculation of Appendix B.7 results in

$$P_T = \mathfrak{p}^T + (T-1)\mathfrak{p}^{T-2}\mathcal{J} + O_T(\epsilon^4), \tag{158}$$

where $\mathcal{J} = O(\epsilon^2)$ is expressed entirely through the full zero-lag and lag-one response vectors. Matching the term proportional to $T-1$ with the spectral form $P_T = C(\epsilon)\mu(\epsilon)^{T-1}$ gives

$$\mu = \mathfrak{p} + \frac{\mathcal{J}}{\mathfrak{p}} + O(\epsilon^4). \tag{159}$$

Terms at the same perturbative order that do not grow with $T$ are absorbed into the boundary prefactor $C$. This establishes the connection between the finite-horizon Gaussian calculation and the asymptotic persistence rate without requiring a mode-wise decomposition.

### C.6 Numerical evaluation

The most direct numerical estimate of the persistence rate is obtained from the physical survival probabilities,

$$\widehat{\mu}_T = \frac{\widehat{P}_{T+1}}{\widehat{P}_T}. \tag{160}$$

These probabilities can be estimated directly by simulating the full teacher-student recurrence with fixed weights. This procedure applies without spectral alignment and provides the natural reference for validating any transfer-operator discretization.

For explicit operator calculations, the Gaussian innovation can be reduced to its effective dimension. Let

$$\Gamma\Gamma^\top = LL^\top, \tag{161}$$

where the columns of $L$ span $\operatorname{range}\Gamma$ and $r_\Gamma = \operatorname{rank}\Gamma$. Then the transition can be written

$$\boldsymbol{\xi}_{t+1} = \mathsf{A}\boldsymbol{\xi}_t + L\boldsymbol{\eta}_{t+1}, \qquad \boldsymbol{\eta}_t \sim \mathcal{N}(0, I_{r_\Gamma}), \tag{162}$$

and therefore

$$(\mathcal{T}f)(\boldsymbol{\xi}) = \chi(\boldsymbol{\xi})\mathbb{E}_{\boldsymbol{\eta}}\left[f(\mathsf{A}\boldsymbol{\xi} + L\boldsymbol{\eta})\right]. \tag{163}$$

For sufficiently small $r_\Gamma$ this expectation may be evaluated by Gauss–Hermite or sparse-grid quadrature; otherwise Monte Carlo evaluation is more practical.

In the low-dimensional aligned examples used for the dedicated operator validation, this general construction reduces to the mode-wise representation used in the simulations. We therefore retain those experiments as a special-case numerical check of the operator theory, rather than as an assumption required by the transfer-operator formulation.

The numerical validation uses three complementary checks:

1. in the feedforward limit, $\mu = \mathfrak{p}$;
2. the finite-horizon ratios $P_{T+1}/P_T$ converge with increasing horizon whenever a dominant rate is present;
3. in the symmetric two-mode reference configuration, the measured weak-recurrence coefficient agrees with the analytic specialization derived below.

### C.7 Non-perturbative use and relation to learning

The transfer-operator construction itself does not require weak recurrence. For every fixed stable configuration satisfying the preceding spectral assumptions,

$$P_T = C\mu^{T-1} + O(\mu_{\mathrm{sub}}^{T-1}) \tag{164}$$

at large $T$, even when the perturbative approximation $\mu = \mathfrak{p} + \mathcal{J}/\mathfrak{p} + \cdots$ is no longer quantitatively accurate.

The persistence rate nevertheless characterises reward availability at fixed weights. For a pure success reward setting, $r_2 = 0$, the population field can be written as

$$F(\theta) = r_1 P_T(\theta) V_T(\theta), \tag{165}$$

where

$$V_T(\theta) = \frac{1}{T}\Big\langle \sum_{t=1}^{T} y_t \nabla_\theta \lambda_t \,\Big|\, \mathcal{I}_T = 1, \theta \Big\rangle \tag{166}$$

is the conditional mean update given a successful episode.

If the response statistics and conditional learning progress change slowly over the main plateau interval, without introducing any additional exponential dependence on $T$, the persistence rate suggests the following scaling for reward-limited learning

$$\tau_T \simeq A_T \mu_{\mathrm{pl}}^{-(T-1)}. \tag{167}$$

Here, $A_T$ accounts for the boundary prefactor, the progress made following a successful episode, and the criterion used to define plateau escape. Taking the logarithm gives

$$\frac{\log \tau_T}{T-1} = -\log \mu_{\mathrm{pl}} + \frac{\log A_T}{T-1}. \tag{168}$$

Thus, the persistence rate determines the leading dependence of escape time on sequence length only when the contribution from $A_T$ is subleading.

The transfer operator alone is therefore not sufficient to determine the optimal recurrent matrix or the optimal strength of recurrence. At finite horizons, both the prefactor and subleading spectral modes contribute, while during learning the network parameters and derivative-weighted moments continue to evolve. Determining which recurrent configuration minimizes the overall learning time consequently requires solving the projected master equations or examining the unrestricted learning dynamics through simulations.

Finally, the stability boundary

$$\rho(\mathsf{A}) = 1 \tag{169}$$

lies outside the stationary construction used in this section. At this boundary the stationary Gaussian measure need not exist, and the survival probability need not exhibit the same exponential asymptotics.

# D Dynamical structure of reward-modulated learning

The population learning field allows us to distinguish how often the network receives a reward from how its connectivity changes following successful episodes. We use this distinction to investigate the origin of learning plateaus and how recurrent interactions emerge during learning. We then examine a single-unit benchmark and derive balance relations associated with changes of hidden-state coordinates.

## D.1 Conditional learning paths and sucess-dependent time rescaling

**Time resscaling.** As outlined in Section C.7 for pure success reward ($r_2 = 0$) the population learning equation is the product of two factors with different meanings, $\frac{d\theta}{d\varsigma} = r_1\, P_T(\theta)\, V_T(\theta)$. $P_T$ is the probability that an episode is rewarded, that is, how often the student receives a learning signal at all, while $V_T$ is the average update produced by one rewarded episode, that is, what a student learns from a single success. On a reward-limited plateau, the two factors behave very differently, $P_T$ is exponentially small in $T$, while $V_T$ permits escape without introducing an additional exponential slowing. The scarcity of successful episodes therefore sets the leading exponential growth of the escape time. To separate the effect of the reward probability from the actual direction of the learning field dictated by $V_T$, we rescale time by the episode success probability at each learning time step. To that end we introduce along any learning trajectory with non-zero episode success probability $P_T(\theta) > 0$ the **(reward-)rescaled learning time** $\sigma$,

$$\sigma(\varsigma) = r_1 \int_0^{\varsigma} P_T(\theta(s))\, \mathrm{d}s. \tag{170}$$

This time rescaling results in the dynamical equations

$$\frac{\mathrm{d}\theta}{\mathrm{d}\sigma} = V_T(\theta), \qquad \frac{\mathrm{d}\varsigma}{\mathrm{d}\sigma} = \frac{1}{r_1 P_T(\theta)}. \tag{171}$$

These two equations transverse the same parameter-space path at different speeds. For a fixed horizon and initial state, changing the positive reward amplitude $r_1$ changes only the rescaled learning time.

The same construction applies to the projected shared-subspace model. At a state $\widehat{\theta} = \Psi(S_L)$, the projection is linear in the field, so that

$$\dot{\widehat{\theta}} = r_1 P_T(\widehat{\theta}) \Pi_{S_L} V_T(\widehat{\theta}), \qquad \frac{\mathrm{d}\widehat{\theta}}{\mathrm{d}\sigma} = \Pi_{S_L} V_T(\widehat{\theta}). \tag{172}$$

These identities hold for fixed $T$. Changing the horizon changes the success probability and can also change $V_T$. Consequently, a collapse across horizons in the rescaled learning time should be directly confirmed and is not automatically implied from Eq. 171. Moreover, coincidence of one observable in this clock does not imply coincidence of the complete weight paths. At $P_T = 0$, conditioning and the time change are undefined; in the pure-success population model, the field itself is then zero.

**Feedforward configuration as reference.** When both recurrent matrices vanish, input–readout updates depend only on the current observation, and agreement events at different times are independent. Recall that $\mathfrak{p}$ denotes the single-step agreement probability. We define the single-step input-readout moment as $\mathfrak{m}(\theta) = \langle \chi\, y \nabla_\theta \lambda | \theta \rangle$. The recurrent population field vanishes by the independent sign symmetry of different time blocks. The remaining blocks satisfy

$$F_T = r_1 \mathfrak{p}^{T-1} \mathfrak{m}, \qquad P_T = \mathfrak{p}^T, \qquad V_T = \frac{\mathfrak{m}}{\mathfrak{p}}. \tag{173}$$

For this feedforward reference the conditional field is independent of the horizon. This gives a useful limiting check for the time rescaling.

**Two sources of slow observable dynamics.** For a differentiable observable $O$, we have

$$\frac{\mathrm{d}O}{\mathrm{d}\varsigma} = r_1 P_T \nabla O^\top V_T, \tag{174}$$

with $V_T$ replaced by $\Pi_{S_L} V_T$ in the projected model. A plateau can therefore reflect rare rewards, small progress per rewarded update, or both.

### D.2 Escape as an integral of the rescaled time

For a specified observable $O$ and an associated threshold value $O_{\mathrm{th}}$, we denote with $\sigma_e$ the first crossing time of the observable through $O_{\mathrm{th}}$ along the reward-conditional (time rescaled) trajectory. If the threshold is reached and $P_T$ remains positive up to it, the associated learning time is

$$\tau_{\mathrm{esc}} = \frac{1}{r_1} \int_0^{\sigma_e} \frac{\mathrm{d}\sigma}{P_T(\theta(\sigma))}. \tag{175}$$

**Local approximation.** We introduce the notation for the log episode-success probaility at specific rescaled time and its derivative at the begining of training $f_T(\sigma) = \log P_T(\theta(\sigma))$ and $k_T = f_T'(0)$. We approximate this function by $f_T(0) + k_T \sigma$ over the part of the path that dominates the integral obtaining

$$\tau_{\mathrm{loc}} = \frac{1 - \exp(-k_T \sigma_e)}{r_1 P_T(\theta_0) k_T}, \qquad \tau_{\mathrm{loc}}\big|_{k_T=0} = \frac{\sigma_e}{r_1 P_T(\theta_0)}. \tag{176}$$

This is an approximation to the escape-time integral. Its accuracy is determined by the following inequality: if $|f_T(\sigma) - f_T(0) - k_T \sigma| \le \delta$ throughout the interval, then

$$e^{-\delta} \tau_{\mathrm{loc}} \le \tau_{\mathrm{esc}} \le e^{\delta} \tau_{\mathrm{loc}}. \tag{177}$$

For a positive derivative of the episode success-probability at the start of learning $k_T > 0$, a large $k_T \sigma_e$ suppresses dependence on the distant endpoint **within** the local approximation. A local endpoint approximation additionally requires small curvature on the scale $1/k_T$ and no later region of low reward probability that dominates the integral.

**Relation to the persistence rate.** For fixed weights, the spectral analysis in Appendix C gives $P_T(\theta) = C(\theta)\mu(\theta)^{T-1}[1+o(1)]$. This describes reward availability, but does not by itself determine the escape time, since the escape-time integral also depends on how the network changes along the conditional learning path.

Under the reward-limited plateau approximation of Appendix C.7, a representative persistence factor $\mu_{\mathrm{pl}}$ describes the part of the trajectory that dominates the escape-time integral, giving

$$\tau_T \simeq A_T\, \mu_{\mathrm{pl}}^{-(T-1)}. \tag{178}$$

Here $A_T$ collects the survival prefactor, the reward amplitude, and the contributions of the conditional learning path and escape criterion. The persistence factor determines the leading exponential dependence on episode length only when this plateau approximation remains valid and $\log A_T = o(T)$. In particular, conditional learning must reach the threshold without introducing a competing exponential dependence, and a later region of lower reward probability must not replace the assumed plateau as the dominant source of delay.

At finite horizons, these additional contributions can differ between recurrent configurations and need not cancel in escape-time ratios. We therefore use the finite-horizon escape-time integral and its local approximation to assess learning times, rather than infer them from $\mu_{\mathrm{pl}}$ alone.

### D.3 The linearised REINFORCE update increases the rewarded pre-activation magnitude

For an episode with nonzero pre-activations,

$$\mathcal{E}_T = \nabla_\theta m_T(\theta; \mathbf{x}_{1:T}), \qquad m_T = \frac{1}{T}\sum_{t=1}^{T} |\lambda_t|. \tag{179}$$

For a successful episode, the infinitesimal update therefore increases that same episode's preactivation magnitude.

**Gain growth.** At fixed recurrent connectivity, each output is homogeneous of degree one separately in $W_x$ and $\mathbf{c}$. Consequently, for pure success reward,

$$\langle W_x, F_x\rangle_F = \mathbf{c}^\top \mathbf{f}_c = \frac{r_1}{T}\Big\langle \mathcal{I}_T \sum_{t=1}^{T} |\lambda_t| \Big\rangle. \tag{180}$$

The squared input and readout norms therefore increase at the same rate. The right-hand side is strictly positive whenever success has positive probability and the outputs are nondegenerate. The unrestricted pure-success population flow thus has no finite stationary point in this region. The same readout conclusion holds in the projected model because $\dot{\mathbf{c}} = \mathbf{f}_c$.

**Although positive rescaling of the output leaves a hard-sign policy unchanged, but can change subsequent parameter velocities. Gain is therefore behaviorally redundant without being dynamically irrelevant.**

### D.4 Local recurrent learning at the feedforward state

We now examine which recurrent directions are generated near $W_h = 0$. To that end, we keep here $W_x$ and $\mathbf{c}$ fixed when differentiating the recurrent field. Let

$$\begin{aligned} q &= \|\boldsymbol{\varphi}_0\|^2 > 0, \quad q^* = \|\boldsymbol{\varphi}_0^*\|^2 > 0, \\ \mathbf{u} &= \boldsymbol{\varphi}_0^\top/\sqrt{q}, \quad \mathbf{u}^* = \boldsymbol{\varphi}_0^{*\top}/\sqrt{q^*}, \\ \rho_0 &= \mathbf{u}^\top \mathbf{u}^*, \quad \mathfrak{p} = \frac{1}{2} + \frac{\arcsin \rho_0}{\pi}, \end{aligned} \tag{181}$$

with $|\rho_0| < 1$. For one observation, write $\chi_0 = \mathbf{1}\{(u^\top x)(u^{*\top} x) > 0\}$.

**The agreement-weighted second moment.** The moment $\Omega = \langle \chi_0 \mathbf{x}\mathbf{x}^\top \rangle$ is unnormalised . We define the orthonormal directions

$$\mathbf{e}_+ = \frac{\mathbf{u} + \mathbf{u}^*}{\sqrt{2(1+\rho_0)}}, \qquad \mathbf{e}_- = \frac{\mathbf{u} - \mathbf{u}^*}{\sqrt{2(1-\rho_0)}}. \tag{182}$$

In their plane, agreement consists of two opposite wedges of total angle $2\pi\mathfrak{p}$. Integrating the radial Gaussian moment and the angular second moments gives

$$\Omega = \mathfrak{p} I + \frac{\sqrt{1-\rho_0^2}}{\pi}(\mathbf{e}_+\mathbf{e}_+^\top - \mathbf{e}_-\mathbf{e}_-^\top). \tag{183}$$

Directions orthogonal to this plane contribute $\mathfrak{p} I$. The matrix is positive semidefinite, as also follows directly from its definition. For $|\rho_0| < 1$ its eigenvalues in the plane are $\mathfrak{p} \pm \sqrt{1-\rho_0^2}/\pi > 0$.

**A feedforward teacher – transverse recurrent amplification.** When the teacher is feedforward $W_h^* = 0$, the recurrent field vanishes at $W_h = 0$. Its linearisation is

$$F_h(W_h) = \kappa_T \mathbf{c}\mathbf{c}^\top W_h \Omega_x + O(\|W_h\|_F^2), \qquad \Omega_x = W_x \Omega W_x^\top, \qquad \kappa_T = \frac{r_1 (T-1)\mathfrak{p}^{T-2}}{T\sqrt{2\pi q}}. \tag{184}$$

To obtain this expression, expand $\lambda_t = \boldsymbol{\varphi}_0 \mathbf{x}_t + \mathbf{c}^\top W_h W_x \mathbf{x}_{t-1} + O(\|W_h\|_F^2)$ for $t \geq 2$. The zeroth-order recurrent derivative is $\mathbf{c}(W_x \mathbf{x}_{t-1})^\top$. At zero recurrence the time blocks are independent, and all first-order terms except the adjacent link vanish by an odd sign symmetry in at least one block. For that link, use

$$\mathbf{1}\{\lambda\nu > 0\} \operatorname{sgn}\lambda = \tfrac{1}{2}(\operatorname{sgn}\lambda + \operatorname{sgn}\nu).$$

The derivative with respect to a small additive shift in $\lambda$ is $f_\lambda(0) = 1/\sqrt{2\pi q}$, after averaging the current observation. The previous observation contributes $\Omega_x$, the other $T-2$ time blocks contribute $\mathfrak{p}^{T-2}$, and the sum contains $T-1$ adjacent links.

If $\Omega_x v_j = s_j v_j$, the frozen-coefficient linear map in Eq. (184) has eigen-directions

$$E_j = \widehat{\mathbf{c}}\, v_j^\top, \qquad \lambda_j = \kappa_T \|c\|^2 s_j \geq 0, \qquad \widehat{\mathbf{c}} = \mathbf{c}/\|\mathbf{c}\|. \tag{185}$$

It vanishes on perturbations satisfying $\mathbf{c}^\top E = 0$. Positive $s_j$ identify expanding, rank-one perturbations with a common receiving direction. During feedforward learning, however, $W_x$ and $\mathbf{c}$ generally evolve. This is therefore a transverse linearisation about a feedforward learning trajectory. An exactly zero recurrent matrix remains zero for a feedforward teacher, and an initial perturbation or finite-sample fluctuation is needed for its amplification.

**A recurrent teacher – a source at zero student recurrence.** For an arbitrary recurrent teacher, evaluating the student derivative at $W_h = 0$ gives the identity

$$F_h\big|_{W_h=0} = \frac{r_1}{T}\mathbf{c} \sum_{t=2}^{T} \Big\langle \mathcal{I}_T y_t (W_x x_{t-1})^\top \Big\rangle. \tag{186}$$

Its rank is at most one. At zero student recurrence the input field also has its column space in $\operatorname{span}\{\mathbf{c}\}$, so the instantaneous joint matrix field has rank at most one. The source need not be nonzero for every teacher, but recurrent dependence in the reward event removes the independent-block cancellation that holds for a feedforward teacher.

For a weak teacher recurrence $W_h^* = \epsilon \overline{W}_h^*$, with input and readout weights fixed, the leading coefficient is

$$F_h\big|_{W_h=0} = \kappa_T^* \mathbf{c}\left[W_x \Omega (\boldsymbol{\varphi}_1^*)^\top\right]^\top + O(\epsilon^2), \qquad \kappa_T^* = \frac{r_1(T-1)\mathfrak{p}^{T-2}}{T\sqrt{2\pi q^*}}. \tag{187}$$

The proof follows the adjacent-link calculation above, now differentiating the teacher preactivation. Its Gaussian density at zero supplies $1/\sqrt{2\pi q^*}$. The trace of the source is $\kappa_T^* \boldsymbol{\varphi}_0 \Omega (\boldsymbol{\varphi}_1^*)^\top$ to this order; it is not generally determined by $\boldsymbol{\varphi}_0 (\boldsymbol{\varphi}_1^*)^\top$ alone.

The source establishes a rank-one initial change, $W_h(\varsigma) = \varsigma F_h(\theta_0) + o(\varsigma)$ when $W_h(0) = 0$. It does not establish an invariant rank-one connectivity family. The readout can rotate, and higher backward-readout directions emerge as recurrence develops. Similarly, a rank-one matrix $\mathbf{c}\mathbf{v}^\top$ has $\mathbf{c}$ as a right eigenvector, whereas a closed scalar output recurrence requires $W_h^\top c = bc$. A readout-aligned receiving direction alone does not establish the latter condition.

### D.5 A single-unit example: alignment, memory, and gain

Consider one recurrent unit with input vector $\mathbf{w}_x = r\mathbf{u}$, $r = \|\mathbf{w}_x\| > 0$, readout $\mathbf{c} > 0$, and recurrence $b$. Let the teacher have unit input direction $u^*$ and recurrence $\beta$. Define

$$\rho = \mathbf{u}^\top \mathbf{u}^*, \qquad g = \mathbf{c}r. \tag{188}$$

Here $\rho$ equals the zero-lag response alignment. We use $g$ for the scalar output gain.

For independent observation times, we let $\xi_t = u^\top x_t$ and $\xi_t^* = u^{*\top} x_t$ be standard Gaussian innovations with contemporaneous correlation $\rho$. We define the zero-reset processes

$$z_t = bz_{t-1} + \xi_t, \qquad v_t = \beta v_{t-1} + \xi_t^*, \tag{189}$$

$$\widetilde{z}_t = b\widetilde{z}_{t-1} + \xi_t^*, \qquad d_t = bd_{t-1} + z_{t-1}. \tag{190}$$

The student preactivation is $\lambda_t = \gamma z_t$, and $d_t = \partial_b z_t$. The auxiliary process $\widetilde{z}_t$ filters the teacher's input feature using the student's recurrence and enters the derivative of response alignment. The teacher's positive scale does not change the success event. Introduce

$$\mathfrak{a}_T(\rho, b) = \frac{1}{T}\Big\langle \sum_t |z_t| \mid \mathcal{I}_T = 1 \Big\rangle, \tag{191}$$

$$\mathfrak{b}_T(\rho, b) = \frac{1}{T}\Big\langle \sum_t \operatorname{sgn}(z_t)(\widetilde{z}_t - \rho z_t) \mid \mathcal{I}_T = 1 \Big\rangle, \tag{192}$$

$$\mathfrak{d}_T(\rho, b) = \frac{1}{T}\Big\langle \sum_t \operatorname{sgn}(z_t) d_t \mid \mathcal{I}_T = 1 \Big\rangle. \tag{193}$$

The dependence on $T$ and fixed $\beta$ is understood. Direct differentiation in the successful-episode clock gives

$$\frac{\mathrm{d}r}{\mathrm{d}\sigma} = c\mathfrak{a}_T, \quad \frac{\mathrm{d}c}{\mathrm{d}\sigma} = r\mathfrak{a}_T, \quad \frac{\mathrm{d}\rho}{\mathrm{d}\sigma} = \frac{c}{r}\mathfrak{b}_T, \quad \frac{\mathrm{d}b}{\mathrm{d}\sigma} = cr\mathfrak{d}_T. \tag{194}$$

Thus $c^2 - r^2$ is conserved. On the positive balanced branch $c = r$, the autonomous system closes on three variables:

$$\frac{\mathrm{d}\rho}{d\sigma} = \mathfrak{b}_T(\rho, b), \qquad \frac{\mathrm{d}b}{\mathrm{d}\sigma} = \gamma\mathfrak{d}_T(\rho, b), \qquad \frac{d\gamma}{d\sigma} = 2\gamma\mathfrak{a}_T(\rho, b). \tag{195}$$

**Although the gain does not change the hard-sign policy, it changes the relative speed of recurrent learning and response alignment.**

**Nullclines and trajectories.** Where $g > 0$, the memory nullcline is defined by $\mathfrak{d}_T(\rho, b) = 0$. If we write this curve $b = b_{\mathrm{n}}(\rho)$, then on it

$$\frac{\mathrm{d}}{\mathrm{d}\sigma}\{b - b_{\mathrm{n}}(\rho)\}\Big|_{b=b_{\mathrm{n}}(\rho)} = -b_{\mathrm{n}}'(\rho)\mathfrak{b}_T(\rho, b_{\mathrm{n}}(\rho)). \tag{196}$$

This expression generally does not vanish. The nullcline separates signs of the instantaneous memory drift, not necessarily basins of the full three-variable system.

A two-step calculation illustrates the distinction between the linearised REINFORCE stationarity and successful imitation of the teacher. At $\rho = 1$, the two networks use the same spatial responses, and

$$P_2(b; \beta) = 1 - \frac{|\arctan b - \arctan \beta|}{\pi}. \tag{197}$$

The success probability is maximal at $b = \beta$. Nevertheless, the recurrent linearised REINFORCE component in learning time is

$$F_b = \frac{r_1\gamma}{2\sqrt{2\pi}}\left(\frac{b}{\sqrt{1+b^2}} + \frac{\beta}{\sqrt{1+\beta^2}}\right). \tag{198}$$

This follows from the sign identity we used above and $\langle x_1 \operatorname{sgn}(ax_1 + x_2)\rangle = \sqrt{2/\pi}\, a/\sqrt{1+a^2}$. At matching recurrence,

$$F_b\big|_{b=\beta} = \frac{r_1\gamma}{\sqrt{2\pi}}\frac{\beta}{\sqrt{1+\beta^2}}, \tag{199}$$

which is nonzero for $\beta \neq 0$. Continued learning beyond matching the teacher weights can therefore reduce success if learning continues. Conversely, $b = -\beta$ is a zero of this recurrent component, not a stationary point of the full state because gain continues to grow.

For fixed finite $T$, preactivations and their derivatives are finite polynomials in $b$. The boundaries $b = \pm 1$ concern stationary infinite-horizon dynamics, they are not automatically attractors or singularities of the finite-episode learning field.

## Software and reproducibility

All code is written in Python 3 using NumPy. All simulations are implemented with vectorised NumPy operations. The ODE integrator and stochastic simulation are contained in a single self-contained script included in the supplementary materials. All figures can be reproduced with the provided code on a standard laptop CPU.

## Usage of Large Language Models

During the preparation of the manuscript we employed large language models for brainstorming ideas on how to tackle sub-problems encountered during the formulation of the theory, for polishing of the writing, for plotting simulation results, and for proof-checking milestones of the derivation.

## NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

**The checklist answers are an integral part of your paper submission.** They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We study learning dynamics of recurrent linear neural networks in a reinforcement learning setting.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: We have included a dedicated subsection named limitations in the discussion.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: We provide a subsection with the main assumptions, and in every step of our derivations we detail our assumptions. We provide detailed exposition of the theory in the Supplement.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Yes, we took our best to provide an extensive supplementary information section.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We will submit the codebase used to validate the theory during the review process.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.

- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: [N/A]

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification:We provide error bars in the figures that present resutls from Monte Carlo simulations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In the supplement, although it is a theory paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: [Yes]

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We describe the numerical experiments performed to validate the theory in the supplement.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: [N/A]

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We provide in the supplement a description on how LLMs where used during the preparation of this manuscript.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.