LEARNING TO LEARN DYNAMICAL ASSOCIATIONS WITH REWARD-GATED LOCAL PLASTICITY

Dimitra Maoutsa
dimitra.maoutsa@gmail.com

ABSTRACT

A central question pertinent to neuroscience is how learning creates and reshapes memories under constraints faced by biological circuits: recurrent dynamics, local synaptic plasticity, and delayed feedback signals. Classical theories of associative memory formalize memory formation as content-addressable storage of patterns, often in the form of attractors in Hopfield-like networks. Yet, animals are routinely confronted with tasks that require **dynamical associations**, memories embedded in the evolving neural trajectories rather than in static neural states. How such dynamical memories are formed through biologically plausible learning, and how rule variants shape the resulting circuit solutions, remains unclear. Here, we introduce a meta-learning framework that discovers families of reward-gated local plasticity rules that enable recurrent circuits to acquire dynamical associations from delayed reinforcement signals. We equip synapses with eligibility traces that accumulate pre–post synaptic co-activity and allow for punctuated changes of neuronal interactions upon reward delivery. Rule parameters shape the eligibility dynamics, thereby controlling how co-activation patterns drive plasticity and the formation of dynamical memories. To avoid differentiating through the full learning trajectory, we optimize rule parameters using a policy-gradient estimator of the expected cumulative reward, and use forward-mode differentiation to compute sensitivities of the eligibility dynamics with respect to rule parameters. This framework enables systematic search and analysis of biologically plausible rules for acquiring dynamical associative memories that support learning of common neuroscience tasks from delayed rewards.

1 INTRODUCTION

Learning in biological organisms involves changes in synaptic connections (synaptic plasticity) between neurons (Bailey & Kandel, 1993; Mayford et al., 2012). Synaptic changes are believed to underlie memory formation and are essential for adaptive behaviour (Hopfield, 1982). Experimental evidence suggests that synaptic changes depend on the co-activation of pre- and postsynaptic activity (Bi & Poo, 1998; Sjöström et al., 2001), and possibly other local variables available at the synaptic site (Graupner & Brunel, 2012; Pedrosa & Clopath, 2020). These unsupervised synaptic modifications have explained activity-dependent circuit refinement during development such as the emergence of functional properties like receptive field formation based on naturalistic input statistics (Martin et al., 2000; Blais et al., 1997; Brito & Gerstner, 2016; Gütig et al., 2003).

A complementary perspective is that recurrent circuits can implement content-addressable retrieval, where distributed interactions among neurons store information that can be recovered from partial cues, often formalized through attractor dynamics in Hopfield networks (Hopfield, 1982). However, natural behavior often goes beyond mere retrieval of single static patterns, but rather requires the storage and retrieval of dynamical associations, where what must be learned is embedded in the collective neural trajectory. In this setting, learning is not only about storing patterns, but also about shaping recurrent dynamics so that the circuit implements the appropriate computation at the level of state evolution. However, how such dynamical memories are formed under biological constraints remains an open problem.

For instance, most organisms routinely solve complex tasks that require feedback through explicit supervisory or reinforcement signals. These signals are believed to gate or modulate plasticity, acting in the form of a third factor that scales and also possibly imposes the direction of the required synaptic modifications (Kuśmierz et al., 2017; Sosis & Rubin, 2024) to facilitate long-lasting alignment of representations to behaviourally relevant dimensions (Benezra et al., 2024). How error- or

reward-related information is propagated through the recurrent interactions is not yet clear. While prior work has largely focused on hand-crafted synaptic updates for unsupervised neural circuit self-organization, or biologically plausible approximations of backpropagation (Miconi et al., 2018), the space of plasticity rules capable of supporting structured credit assignment from delayed feedback remains vastly underexplored.

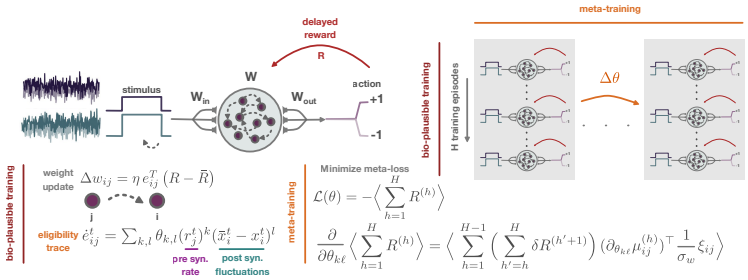


Figure 1: **Outline of the proposed meta-learning framework.**

Backpropagation through time (BPTT), the standard approach for training recurrent neural networks (RNNs), is biologically implausible since it requires symmetric forward and backward connections and non-local information (Lillicrap et al., 2016; Guerguiev et al., 2017). Although recent work has reformulated BPTT into more biologically plausible variants using random feedback (Lillicrap et al., 2016; Murray, 2019), truncated approximations, or by learning feedback pathways (Lindsey & Litwin-Kumar, 2020; Shervani-Tabar & Rosenbaum, 2023), these methods require continuous error signals to refine recurrent connections.

Here, we adopt a bottom-up approach: instead of imposing hand-designed synaptic rules, we discover biologically plausible plasticity rules that support learning through delayed reinforcement signals via meta-optimisation (Schmidhuber et al., 1996). Building on recent work (Confavreux et al., 2023), we parameterise plasticity rules as functions of local signals (presynaptic activity, post-synaptic activity, and synapse size) and meta-learn their parameters within a second reinforcement learning loop.

2 METHOD

Network dynamics. We consider recurrent neural networks of firing rate neurons coupled through a synaptic matrix $\mathbf{W} \in \mathcal{R}^{N\times N}$ (Sompolinsky et al., 1988; Barak, 2017), with additional input and output matrices $\mathbf{W}_{\text{in}} \in \mathcal{R}^{N\times N_{\text{in}}}$ and $\mathbf{W}_{\text{out}} \in \mathcal{R}^{N_{\text{out}}\times N}$ that route task-relevant input into the recurrent circuit and read out network activation to generate task-specific outputs (actions). The equations governing the network dynamics are

$$\frac{\mathrm{d}\mathbf{x}^t}{\mathrm{d}t} = -\mathbf{x}^t + \mathbf{W}\boldsymbol{\phi}(\mathbf{x}^t) + \mathbf{W}_{\text{in}}\mathbf{u}^t, \qquad \mathbf{r}^t = \boldsymbol{\phi}(\mathbf{x}^t) \doteq \tanh(\mathbf{x}^t), \tag{1}$$

where $\mathbf{x}^t \in \mathcal{R}^N$ is the vector of pre-activations (or input currents) to each neuron in the network, $\boldsymbol{\phi}(\cdot) : \mathcal{R}^N \to \mathcal{R}^N$ denotes the single-neuron transfer functions, $\mathbf{r}^t \in \mathcal{R}^N$ is the vector of instantaneous firing rates, $\mathbf{u}^t$ stands for the activity of the N_{in} input neurons. In the terms above, the $\cdot^t$ superscript indicates time dependence. Network outputs $\mathbf{z}^t$ are obtained from linear read-out neurons as $\mathbf{z}^t = \mathbf{W}_{\text{out}}\mathbf{r}^t$.

Sparse feedback and parametrized learning rules. We consider networks that learn cognitive tasks using biologically plausible local learning rules, guided by sparse reinforcement signals R provided only at the end of each training episode. Each synapse between a pre-synaptic unit j and a post-synaptic unit i maintains an eligibility trace e_{ij} (Izhikevich, 2007), which integrates the history of (co-)activation during the episode. We define the evolution of eligibility traces with differential equations of the form

$$\frac{\mathrm{d}e_{ij}^t}{\mathrm{d}t} = \mathcal{H}_\theta(r_j^t, x_i^t) - \frac{e_{ij}^t}{\tau_e} = \sum_{0\le k;\ell\le d;} \theta_{k,\ell} \left(r_j^t\right)^k \left(\bar{x}_i - x_i^t\right)^\ell - \frac{e_{ij}^t}{\tau_e}, \tag{2}$$

where τ_e is a decay time-scale, $\bar{x}_i$ is a running average of the pre-activation of neuron i, and $\theta_{k,\ell} \in \mathcal{R}$ are learnable coefficients. In contrast to eligibility traces based solely on first-order correlations (Gerstner et al., 2018), we use here a polynomial expression that captures richer interactions

between pre- and post-synaptic activity. Each coefficient $\theta_{k,\ell}$ can be construed as a term-specific learning rate, which may be positive (Hebbian), negative (anti-Hebbian), or zero. In our experiments, we set $d = 5$.

The recurrent weight matrix $\mathbf{W}$ gets updated at the end of each training episode according to a reward-modulated learning rule

$$\pi_\Theta\left(\mathbf{\Delta W}^{(h)} \mid \mathbf{\Theta}\right) = \mathcal{MN}\left(\boldsymbol{\mu}_{\mathbf{\Theta}}^{(h)},\, \sigma_w^2\, \mathbf{I}_N, \mathbf{I}_N\right) \quad \text{with} \quad [\boldsymbol{\mu}_{\mathbf{\Theta}}^{(h)}]_{ij} = \eta\, e_{ij}^{T_h}\, \left(R^{(h)} - \bar{R}^{(h)}\right), \quad (3)$$

where with $\mathcal{MN}(\boldsymbol{\mu}, \mathbf{\Sigma}, \mathbf{V})$ we denote the matrix normal distribution with mean $\boldsymbol{\mu} \in \mathcal{R}^{N\times N}$, and $\mathbf{\Sigma}$ and $\mathbf{V}$ the positive semi-definite matrices are the row- and column-variance, while superscript h indicates episode index, η denotes the learning rate, $e_{ij}^{T_s}$ stands for the eligibility trace accumulated till the end of the h-th episode T_h, while R, $\bar{R}$ stand for the obtained and the expected reward. Here, we model reward expectations for each type of trial independently as a running average of past rewards for this trial type (Miconi, 2017).

Meta-learning plasticity rules. While prior work used hand-crafted eligibility traces and synaptic update rules to train RNNs from sparse feedback (Miconi, 2017), we instead meta-learn the plasticity-rule parameters. Our framework consists of two nested training loops: **(i)** an inner loop in which the recurrent network is trained over several episodes using local learning rules and sparse reinforcement signals provided at the end of each episode **(bio-plausible training)**, as described above; and **(ii)** an outer loop that optimizes the plasticity meta-parameters $\mathbf{\Theta} = \{\{\theta_{k,\ell}\}_{k,\ell=0}^{5}\}$ via gradient descent using **tangent-propagation through learning** (forward-mode differentiation through learning) on a meta-loss computed over H training episodes (trials) **(meta-training)**. This approach allows the learning rules themselves to be adapted to the task, rather than be fixed a priori.

Tangent-propagation through learning. Our goal is to optimise the learning rule parameters $\mathbf{\Theta}$ to maximise task performance, quantified as the expected cumulative reward $\langle \sum_h R \rangle$ obtained after a fixed number of learning episodes. However, the reward R depends on the network's output, which is determined by synaptic weights $\mathcal{W} = \{\mathbf{W}_{in}, \mathbf{W}, \mathbf{W}_{out}\}$, with $\mathbf{W}$ evolving under the update rule (Eq.3). Since weights depend on eligibility traces e_{ij}, themselves parameterised by $\mathbf{\Theta}$, the reward depends on the plasticity parameters through $\mathcal{W}$ and $\mathbf{\Theta}$. Directly computing $\nabla_{\mathbf{\Theta}} \langle \sum_h R \rangle$ by backpropagating through the learning dynamics is computationally prohibitive since learning requires several hundreds of trials (Lindsey & Litwin-Kumar, 2020). We therefore, here, adopt a REINFORCE-inspired estimator (Williams, 1992), which involves computing the gradient of an expected value by observing outcomes (*the rewards*) and scaling a measure of what elicited that outcome (*the weight updates*) with the associated reward. Thus, we approximate the gradient of the expected reward by

$$\nabla_\Theta \langle \sum_h R^{(h)} \rangle \approx \langle \sum_h \sum_{h'=h+1}^{H} R^{(h')} \nabla_\Theta \log \pi(\mathbf{\Delta W}^{(h)} \mid \mathbf{\Theta}) \rangle \approx \langle \sum_h \sum_{h'=h+1}^{H} (R^{(h')} - \bar{R}^{(h')})\ \nabla_\Theta \log \pi(\mathbf{\Delta W}^{(h)} \mid \mathbf{\Theta}) \rangle, \quad (4)$$

where $\bar{R}$ stands for the baseline reward, and thus $(R - \bar{R})$ denotes the reward prediction error. Introducing the expression of the plasticity rule, we have in a component-wise formulation for each dimensional component of the plasticity parameters $\mathbf{\Theta}$

$$\frac{\partial}{\partial \theta_{k,\ell}} \left\langle \sum_{h=1}^{H-1} R^{(h)} \right\rangle = \left\langle \sum_{h=1}^{H-1} \left(\sum_{h'=h}^{H-1} \delta R^{(h'+1)} \right) \frac{1}{\sigma_w^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \left(\Delta w_{ij}^{(h)} - \mu_{ij}^{(h)} \right) \frac{\partial \mu_{ij}^{(h)}}{\partial \theta_{k,\ell}} \right\rangle_S, \quad (5)$$

where the expectation $\langle \cdot \rangle_S$ is considered over independent sessions S. This requires the computation of the sensitivity of the mean weight update wrt to the plasticity parameters $\frac{\partial \mu_{ij}^{(h)}}{\partial \theta_{kl}}$ over training. To that end, we propagate the gradients of the within-trial pre-activations $\mathbf{x}^t$, $\boldsymbol{\chi}_{k,\ell}^t \in \mathcal{R}^N$ (**state tangent**), the pre-activation trace $\bar{\mathbf{x}}^t$, $\boldsymbol{\psi}_{k,\ell}^t \in \mathcal{R}^N$ (**trace tangent**), and of the eligibility traces of each synaptic pair ij, e_{ij}^t, $[\mathbf{z}_{k,\ell}^t]_{ij}$ (**eligibility tangent**), as well as inter-trial sensitivities of weight matrices (**weight matrix tangent**), $\mathbf{U}_{k,\ell}^{(h)}$ (c.f. Appendix Sec. B).

To make the tangent propagation computationally tractable independent of the size of the parameter space, we optimize a low-dimensional affine reparametrisation of the tangent system, amounting to projected the gradient vector on a different random subspace in each iteration.

REFERENCES

Craig H Bailey and Eric R Kandel. Structural changes accompanying memory storage. *Annual Review of Physiology*, 1993.

Omri Barak. Recurrent neural networks as versatile tools of neuroscience research. *Current Opinion in Neurobiology*, 46:1–6, 2017.

Sam E Benezra, Kripa B Patel, Citlali Pérez Campos, Elizabeth MC Hillman, and Randy M Bruno. Learning enhances behaviorally relevant representations in apical dendrites. *Elife*, 13:RP98349, 2024.

Guo-qiang Bi and Mu-ming Poo. Synaptic modifications in cultured hippocampal neurons: dependence on spike timing, synaptic strength, and postsynaptic cell type. *Journal of Neuroscience*, 18 (24):10464–10472, 1998.

Brian Blais, Nathan Intrator, Harel Shouval, and Leon Cooper. Receptive field formation in natural scene environments: comparison of single cell learning rules. *Advances in Neural Information Processing Systems*, 10, 1997.

Carlos SN Brito and Wulfram Gerstner. Nonlinear hebbian learning as a unifying principle in receptive field formation. *PLoS computational biology*, 12(9):e1005070, 2016.

Basile Confavreux, Poornima Ramesh, Pedro J Goncalves, Jakob H Macke, and Tim Vogels. Meta-learning families of plasticity rules in recurrent spiking networks using simulation-based inference. *Advances in Neural Information Processing Systems*, 36:13545–13558, 2023.

Wulfram Gerstner, Marco Lehmann, Vasiliki Liakoni, Dane Corneil, and Johanni Brea. Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules. *Frontiers in Neural Circuits*, 12:53, 2018.

Michael Graupner and Nicolas Brunel. Calcium-based plasticity model explains sensitivity of synaptic changes to spike pattern, rate, and dendritic location. *Proceedings of the National Academy of Sciences*, 109(10):3991–3996, 2012.

Jordan Guerguiev, Timothy P Lillicrap, and Blake A Richards. Towards deep learning with segregated dendrites. *Elife*, 6:e22901, 2017.

Robert Gütig, Ranit Aharonov, Stefan Rotter, and Haim Sompolinsky. Learning input correlations through nonlinear temporally asymmetric hebbian plasticity. *Journal of Neuroscience*, 23(9): 3697–3714, 2003.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982.

Eugene M Izhikevich. Solving the distal reward problem through linkage of stdp and dopamine signaling. *Cerebral Cortex*, 17(10):2443–2452, 2007.

Łukasz Kuśmierz, Takuya Isomura, and Taro Toyoizumi. Learning with three factors: modulating hebbian plasticity with errors. *Current Opinion in Neurobiology*, 46:170–177, 2017.

Christiane Lemieux. Control variates. *Wiley StatsRef: Statistics Reference Online*, pp. 1–8, 2014.

Timothy P Lillicrap, Daniel Cownden, Douglas B Tweed, and Colin J Akerman. Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7 (1):13276, 2016.

Jack Lindsey and Ashok Litwin-Kumar. Learning to learn with feedback and local plasticity. *Advances in Neural Information Processing Systems*, 33:21213–21223, 2020.

Stephen J Martin, Paul D Grimwood, and Richard GM Morris. Synaptic plasticity and memory: an evaluation of the hypothesis. *Annual Review of Neuroscience*, 23(1):649–711, 2000.

Mark Mayford, Steven A Siegelbaum, and Eric R Kandel. Synapses and memory storage. *Cold Spring Harbor perspectives in biology*, 4(6):a005751, 2012.

Thomas Miconi. Biologically plausible learning in recurrent neural networks reproduces neural dynamics observed during cognitive tasks. *Elife*, 6:e20899, 2017.

Thomas Miconi, Kenneth Stanley, and Jeff Clune. Differentiable plasticity: training plastic neural networks with backpropagation. In *International Conference on Machine Learning*, pp. 3559–3568. PMLR, 2018.

James M Murray. Local online learning in recurrent networks with random feedback. *Elife*, 8: e43299, 2019.

Victor Pedrosa and Claudia Clopath. Voltage-based inhibitory synaptic plasticity: network regulation, diversity, and flexibility. *bioRxiv*, pp. 2020–12, 2020.

Juergen Schmidhuber, Jieyu Zhao, and MA Wiering. Simple principles of metalearning. *Technical report IDSIA*, 69:1–23, 1996.

Navid Shervani-Tabar and Robert Rosenbaum. Meta-learning biologically plausible plasticity rules with random feedback pathways. *Nature Communications*, 14(1):1805, 2023.

Per Jesper Sjöström, Gina G Turrigiano, and Sacha B Nelson. Rate, timing, and cooperativity jointly determine cortical synaptic plasticity. *Neuron*, 32(6):1149–1164, 2001.

Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural networks. *Physical Review Letters*, 61(3):259, 1988.

Baram Sosis and Jonathan E Rubin. Distinct dopaminergic spike-timing-dependent plasticity rules are suited to different functional roles. *bioRxiv*, 2024.

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

A PLASTICITY GRADIENT WITH REINFORCE APPROXIMATION

Following the REINFORCE estimator (Williams, 1992), we approximate the gradient of the expected reward by

$$\nabla_\theta \langle R \rangle = \left\langle (R - \bar{R})\, \nabla_\theta \log \pi(\mathbf{\Delta W} \mid \theta) \right\rangle . \tag{6}$$

This results from applying the log-derivative trick on the expectation of Eq. 6

$$\begin{aligned}
\nabla_\theta \langle R \rangle &= \nabla_\theta \int \pi(\mathbf{\Delta W} \mid \theta)\, R \, \mathrm{d}\mathbf{\Delta W} \\
&= \int \nabla_\theta \pi(\mathbf{\Delta W} \mid \theta)\, R \, \mathrm{d}\mathbf{\Delta W} && \text{(Leibniz integral rule)} \\
&= \int \pi(\mathbf{\Delta W} \mid \theta)\, \frac{\nabla_\theta \pi(\mathbf{\Delta W} \mid \theta)}{\pi(\mathbf{\Delta W} \mid \theta)}\, R \, \mathrm{d}\mathbf{\Delta W} \\
&= \int \pi(\mathbf{\Delta W} \mid \theta)\, \nabla_\theta \log \pi(\mathbf{\Delta W} \mid \theta)\, R \, \mathrm{d}\mathbf{\Delta W} && \text{(log-derivative trick)} \\
&= \langle R\, \nabla_\theta \log \pi(\mathbf{\Delta W} \mid \theta) \rangle_\pi \\
&\approx \left\langle \underbrace{(R - \bar{R})}_{\substack{\text{reward prediction} \\ \text{error}}}\ \nabla_\theta \log \pi(\mathbf{\Delta W} \mid \theta) \right\rangle .
\end{aligned}$$

In the last expression we have introduced the baseline reward $\bar{R}$ as a control variate (Lemieux, 2014) commonly used for variance reduction of the expectation. This heuristic uses the **reward prediction error** $\delta R = R - \bar{R}$ as a scaling factor for the direction of the update. This approximation assumes that R is a smooth functional of $\mathbf{W}$ and that changes in θ affect R primarily through their effect on the connectivity $\mathbf{W}$.

B TANGENT-PROPAGATION THROUGH LEARNING

Tangent-propagation through single trial time. To be able to compute the gradient of the weight updates with respect to the plasticity parameters, we propagate the gradients of the within-trial pre-activations $\mathbf{x}^t$, $\boldsymbol{\chi}^t_{k,\ell} \in \mathcal{R}^N$ (**state tangent**), the pre-activation trace $\bar{\mathbf{x}}^t$, $\boldsymbol{\psi}^t_{k,\ell} \in \mathcal{R}^N$ (**trace tangent**), and of the eligibility traces of each synaptic pair ij, e^t_{ij}, $[\mathbf{z}^t_{k,\ell}]_{ij}$ (**eligibility tangent**). Thus we define the following within-trial tangents (sensitivities) with respect to the plasticity parameter $\theta_{k,\ell}$

$$\boldsymbol{\chi}^t_{k,\ell} \doteq \frac{\partial \mathbf{x}^t}{\partial \theta_{k,\ell}}, \qquad \boldsymbol{\psi}^t_{k,\ell} \doteq \frac{\partial \bar{\mathbf{x}}^t}{\partial \theta_{k,\ell}}, \qquad \mathbf{z}^t_{k,\ell} \doteq \frac{\partial \mathbf{e}^t}{\partial \theta_{k,\ell}}. \tag{7}$$

We assume that the reward and baseline reward R and $\bar{R}$ are not directly related to the plasticity parameters $\theta_{k,\ell}$, and thus we treat these variables and the reward prediction error as θ-independent.

For convenience we define $\alpha = \mathrm{d}t/\tau$, and denote the derivative of the single neuron activation function with $\mathrm{d}\phi(x) = \phi'(x)\,\mathrm{d}x$. The forward equations for these sensitivity parameters are

$$\begin{aligned}
\boldsymbol{\chi}^{t+1}_{k,\ell} &= \boldsymbol{\chi}^t_{k,\ell} + \alpha\Big(-\boldsymbol{\chi}^t_{k,\ell} + \mathbf{W}^{(h)} \left(\mathrm{diag}(\phi'(\mathbf{x}^t)) \cdot \boldsymbol{\chi}^t_{k,\ell}\right) + \mathbf{U}^{(h)}_{k,\ell}\, \mathbf{r}^t \Big) \\
\boldsymbol{\psi}^{t+1}_{k,\ell} &= \alpha_x\, \boldsymbol{\psi}^t_{k,\ell} + (1-\alpha_x)\, \boldsymbol{\chi}^{t+1}_{k,\ell}, \\
\mathbf{z}^{t+1}_{k,\ell} &= \mathbf{z}^t_{k,\ell} + \mathrm{d}t\, (\boldsymbol{\Delta}\mathbf{x}^t)^\ell \otimes \left(\mathbf{r}^t\right)^k \\
&\quad + \mathrm{d}t \sum_{\kappa,\lambda} \Big[\theta_{\kappa,\lambda}\, \lambda\, (\boldsymbol{\Delta}\mathbf{x}^t)^{\lambda-1} (\boldsymbol{\psi}^t_{k,\ell} - \boldsymbol{\chi}^{t+1}_{k,\ell}) \otimes \left(\mathbf{r}^t\right)^\kappa + \theta_{\kappa,\lambda}\, (\boldsymbol{\Delta}\mathbf{x}^t)^\lambda \otimes \kappa\, \left(\mathbf{r}^t\right)^{\kappa-1} \left(\mathrm{diag}(\phi'(\mathbf{x}^t)) \cdot \boldsymbol{\chi}^t_{k,\ell}\right) \Big],
\end{aligned} \tag{8}$$

where $\mathrm{diag}(\mathbf{y})$ denotes the matrix with $\mathbf{y}$ in the main diagonal, $\otimes$ denotes the outer product, while

$$\mathbf{U}^{(h)}_{k,\ell} \doteq \frac{\partial \mathbf{W}^{(h)}}{\partial \theta_{k,\ell}} \tag{9}$$

stands for the inter-trial **weight matrix tangent**. The initial conditions for the three sensitivity parameters are zero at the beginning of each trial $\boldsymbol{\chi}^0_{k,\ell} = \mathbf{0}$, $\boldsymbol{\psi}^0_{k,\ell} = \mathbf{0}$, and $\mathbf{z}^0_{k,\ell} = \mathbf{0}$.

At the end of each trial h we have

$$\frac{\partial \boldsymbol{\mu}^{(h)}}{\partial \theta_{k,\ell}} = \eta\, \delta R^{(h)}\, \mathbf{z}^{T_h}_{k,\ell}\,. \tag{10}$$

Propagating sensitivities through-learning (across trials). The derivative of the weights of trial $h+1$ w.r.t. $\theta_{k,\ell}$ accumulates the across trial sensitivities

$$\mathbf{U}^{(h+1)}_{k,\ell} \;=\; \mathbf{U}^{(h)}_{k,\ell} + \frac{\partial \boldsymbol{\mu}^{(h)}}{\partial \theta_{k,\ell}}, \qquad \text{with } \mathbf{U}^{(0)}_{k,\ell} = 0. \tag{11}$$

This sensitivity $\mathbf{U}^{(h)}_{k,\ell}$ couples back into the state tangent through the $\mathbf{U}^{(h)}_{k,\ell}\, \mathbf{r}^t$ term, which captures how $\theta_{k,\ell}$ affects later trials through the modified weights of earlier trials.

B.1 VALIDATION OF WEIGHT UPDATES GRADIENTS WRT PLASTICITY PARAMETERS

To validate the gradients wrt plasticity parameters obtained through forward mode differentiation, we compare both single-trial and multi-trial gradients (for $H = 500$ trials) obtained with finite differences (FD) to those computed through forward mode differentiation (FM). To avoid observing discrepancies between the two versions of the gradients, we employ the same noise and environment inputs in all simulations. For this experiment we considered only $\theta_{3,3} = 1$ nonzero, while all other entries of $\boldsymbol{\Theta}$ were zero. For the finite difference calculation, we run the training with $\theta + \epsilon$ and $\theta - \epsilon$ for $\epsilon = 10^{-4}$ and approximate the gradient of the weight update with respect to the plasticity parameter as

$$\frac{\mathrm{d}\boldsymbol{\Delta}\mathbf{W}}{\mathrm{d}\theta_{k,\ell}} \approx \frac{\boldsymbol{\Delta}\mathbf{W}^+ (\theta_{k,\ell} + \varepsilon) - \boldsymbol{\Delta}\mathbf{W}^- (\theta_{k,\ell} - \varepsilon)}{2\varepsilon}. \tag{12}$$

The resulting two versions of the gradients are in very close agreement throughout all 500 trials (Fig. 2).

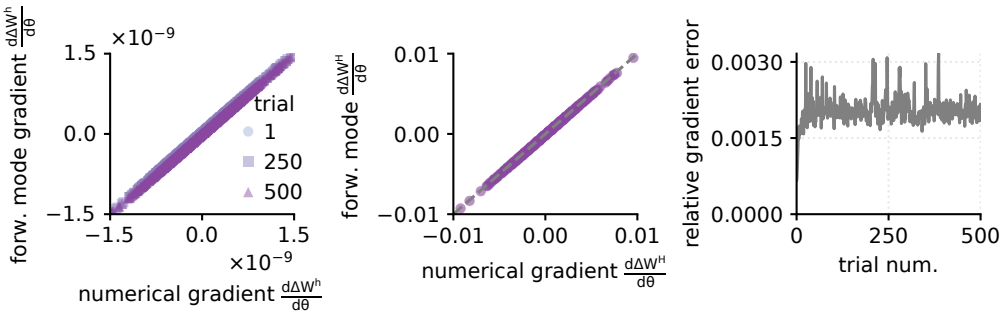


Figure 2: **Validation of forward-mode gradient (FM) computation for the weight update wrt plasticity parameters θ against numerical gradients (FD). a.** Comparison of numerical gradient for per-trial weight update $\mathbf{\Delta W}^{(h)}$ wrt plasticity parameter $\theta_{3,3}$ against the gradient obtained through forward mode differentiation for trials $1, 250, 500$. **b.** Comparison of cumulative gradient computed over 500 trials for the weight update wrt plasticity parameters obtained numerically and through forward mode differentiation. The forward-mode differentiation provides an exact estimation of the plasticity update gradients. **c.** Relative gradient error per trial computed as $\|\frac{\mathrm{d}\mathbf{\Delta W}}{\mathrm{d}\theta}^{FM} - \frac{\mathrm{d}\mathbf{\Delta W}}{\mathrm{d}\theta}^{FD}\|/\|\frac{\mathrm{d}\mathbf{\Delta W}}{\mathrm{d}\theta}^{FD}\|$.

B.2 Local validation of the meta-objective on a two-dimensional subspace

We first considered a simple associative task in which the network must learn an input-output mapping from a two-dimensional input stream. On each trial, one of the two input streams is elevated, fluctuating around 1 depending on the target output. Through learning, the network must associate large activity in the first input channel with the output $+1$, and large activity in the second input channel with the output -1, so as to maximize reward across sessions. The output must be withheld and produced only during the designated decision period.

We empirically verified that the meta-objective has a meaningful local structure for optimization in low-dimensional parameter subspace. In particular, we examined the dependence of the objective $J(\theta)$ on two dimensions of the plasticity-rule parameter space ($\theta_{3,3}$ and $\theta_{1,2}$) while keeping all remaining parameters fixed for different values of the recurrent weight amplitudes that control the regime of the recurrent network dynamics. This directly verifies that changes in the rule parameters induce systematic and measurable changes in the network performance, rather than only producing random fluctuations among individual repetitions.

We varied the selected plasticity parameters within a range of $[-1, +1]$, while holding all other components fixed at their random initialisation values and for the same network structure, while varying only the recurrent weight amplitude. For each parameter pair, we computed the expected meta-objective by averaging over $M = 8$ independent Monte Carlo runs of the learning trajectory,

$$\bar{J}(\theta_{3,3}, \theta_{1,2}) = \frac{1}{M} \sum_{m=1}^{M} J^{(m)}(\theta_{3,3}, \theta_{1,2}), \tag{13}$$

where each realization uses the same task structure, but samples the stochastic plasticity trajectory independently.

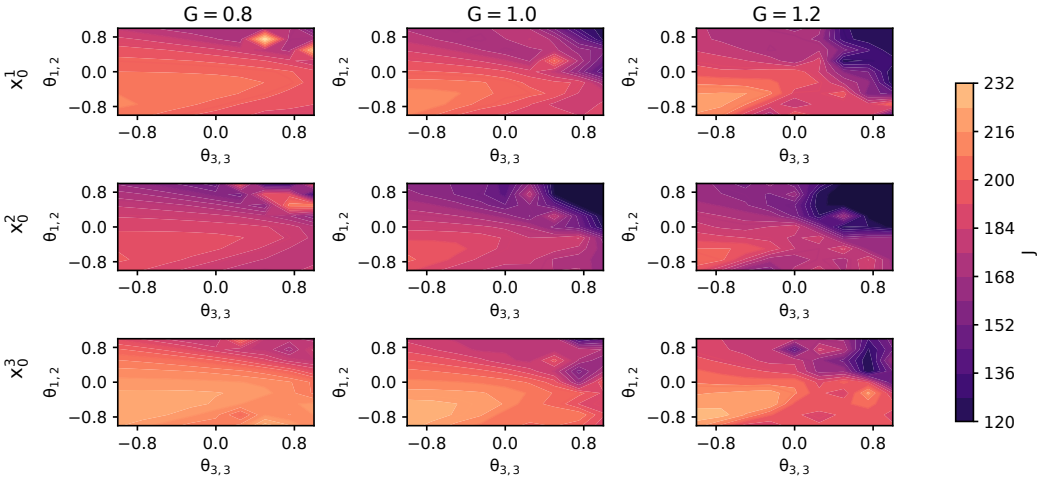
To isolate the dependence on the rule parameters from variability due to connectivity and recurrent state initialization, we fixed the the network weights and repeated the analysis for multiple initial hidden states $\mathbf{x}_0^q$ (indicated by the different rows in Fig. B.2). To quantify the dependence of the meta-objective landscape on the dynamical regime of the network, for fixed network connectivity we repeated the analysis for different values of the recurrent gain parameter G (different columns in Fig. B.2). This allowed us to assess whether the local geometry of the meta-objective is stable across initial conditions and across different dynamical regimes.

Among our different experiments we consistently observed that for the considered task, initial conditions sampled uniformly within the $[-1, +1]$ interval for each neuron independently, did not have considerable influence on the optimization landscape, while, as expected, optimisation was reasonable for networks with recurrent gain $G \geq 1$.

In these numerical experiments we employed plasticity noise amplitude $\sigma_w = 10^{-4}$. We found that optimisation to be well behaved for noise amplitudes in the range $[10^{-4}, 10^{-3}]$. Within this range the divergence among the M individual learning trajectories was moderate, and thus averaging over independent Monte Carlo runs to compute the objective of Eq. 13 (and in general of the full objective) would result in smaller variance.

At the same time, the choice of σ_w involves a trade-off in our framework. Smaller noise amplitudes increase the consistency among the learning trajectories across independent runs, but they also amplify variance of the policy-gradient estimator because of the $1/\sigma_w$[1] prefactor in Eq. 4. Larger noise amplitudes, in contrast, encourage broader exploration in the space of plasticity parameters, but they also produce more divergent learning trajectories, thereby increasing the variance across Monte Carlo repetitions.

Figure B.2 shows representative two-dimensional sections of the meta-objective resulting from these experiments. Across conditions, the main qualitative conclusion is that the plasticity parameters do influence the long-horizon objective J in a structured way, but the local geometry may vary substantially with the recurrent gain G, while the initial recurrent state did not have considerable influence within this setting.



A long-horizon ($H = 250$ trials) experiment on the two-dimensional parameter landscape considered above (Fig. 3) indicates that local optimization results in an improving direction (Fig. 3 **a.**), but the the underlying landscape is highly non-linear as indicated by the non-monotonic evolution of the meta-objective.

[1]One $1/\sigma_w$ drops out when we replace the difference in the bracket with the exact noise variables samples in the Monte Carlo run.

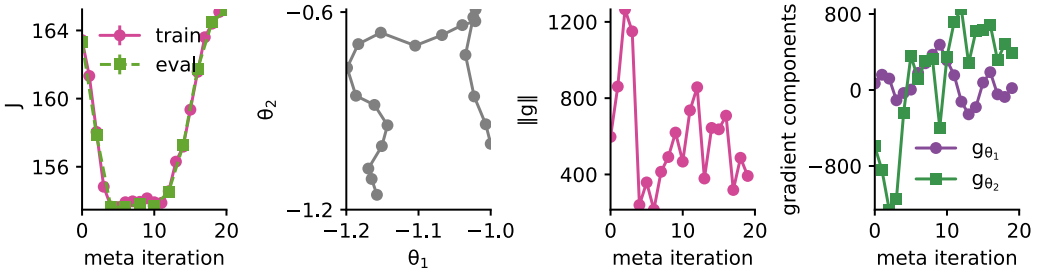


Figure 3: **Short meta-optimization of a $N = 100$ neuron network with recurrent gain $G = 1.2$ results in improvement of the meta-objective and considerable exploration of the parameter subspace. a.** Evolution of the meta-objective J over meta-iterations. Magenta curve indicates average and standard error among 8 realisations of the meta-objective employed for training, while green curve indicates the evaluation of the meta-objective of 5 independent runs not employed in the meta-optimisation. **b.** Evolution of the plasticity parameters in the considered two-dimensional parameter landscape. here $\theta_1 = \theta_{3,3}$ and $\theta_2 = \theta_{1,2}$. **c.** Gradient norm remains stable over the meta-iterations, while **d.** the magnitude of the second dimensional component of the gradient g_2 shows considerably larger magnitude changes compared to the first g_1.