EFFICIENCY AND ROBUSTNESS PARTITION THE SOLUTION SPACE FOR MEMORY IN RECURRENT NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

In computational neuroscience, task-trained recurrent neural networks (RNNs) are commonly used as a testbed for exploring the space of recurrent circuit solutions compatible with a particular function. However, these networks are subject to inductive biases that may be misaligned with those of biological circuits, which operate under various efficiency and robustness constraints. Here, using a minimal stimulus recall task, we systematically characterize how the solution space of recurrent neural networks is shaped by the desiderata of weight efficiency, activity efficiency, and robustness to noise. We show that weight efficient solutions exhibit a low-rank bias that aligns with the inductive biases of task-trained networks and favors dynamical motifs involving slow, persistent dynamics, whereas activity efficiency encourages high-rank solutions that rely on transient amplification. Moreover, we find that activity efficient solutions exhibit a strong dissociation between high-variance modes of activity and those that causally influence network output. Unlike in static feedforward settings, we demonstrate that noise robustness and activity efficiency are generally in opposition, rather than synergistic. Using these findings, we propose a normative account of misaligned coding phenomena recently identified in premotor cortex.

1 INTRODUCTION

A central goal of neuroscience is to understand the representations and dynamical motifs that underlie neural computation, and why they might arise. A large body of work seeks to use task-optimized artificial neural networks (ANNs) to address this question (Mante et al., 2013; Yamins et al., 2014; Rajan et al., 2016; Sorscher et al., 2019; Richards et al., 2019; Schrimpf et al., 2020). In many cases, task-trained ANNs have been argued to recapitulate key features of computations in biological brains, despite obvious mismatches in their computational substrates and learning rules. This perspective, while enjoying remarkable success in recent years, still leaves unanswered the question of why a particular solution has been chosen by biology when many task-compatible solutions exist.

Relatedly, a growing body of work has sought to characterize the solution space of RNNs trained on neuroscience-style tasks, with variation induced by random seeds, architectural variations, learning rules, feature learning strength, or explicit diversity-seeking regularizers (Maheswaranathan et al., 2019; Turner et al., 2021; Kurtkaya et al., 2025; Huang et al., 2025; Qian & Pehlevan, 2025; Pagan et al., 2025; Bordelon et al., 2026; Clark et al., 2026; Ostrow et al., 2026). Among the vast sea of solutions, what organizing principles select the solutions actually implemented in biological circuits? Normative theories such as efficient coding propose that neural circuits select metabolically efficient solutions, an insight that is typically operationalized through constraints on activation sparsity, firing rates, or synaptic weight norms (Attneave, 1954). This perspective has most famously been used to explain sensory representations in visual cortex (Olshausen & Field, 1996; 1997). More recently, similar ideas have been extended beyond sensory coding to neural computations supporting cognition and behavior (Motiwala et al., 2022; Schütt et al., 2024; Stroud et al., 2025; Dorrell et al., 2026). While the effects of efficiency and robustness constraints have been extensively studied in feedforward networks (Atick & Redlich, 1990; Bishop, 1995; Karklin & Simoncelli, 2011; Whittington et al., 2022; Braun et al., 2025), a precise understanding of how these desiderata shape the solution space of recurrent circuits is still lacking.

Standard procedures for training RNNs on neuroscience-style tasks already favor one particular notion of efficiency–small weight norm–whether implicitly through the inductive biases of gradient-

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

based optimization or through explicit regularization. In particular, task-trained RNNs often learn minimal solutions with approximately low-rank structure in their recurrent weights, with rank roughly proportional to task dimensionality (Schuessler et al., 2020; Bordelon et al., 2026; Clark et al., 2026). Often, this low-dimensional description of neural dynamics seems apt, especially when the effective dimensionality of neural activity observed in experimental recordings is small. Yet, this perspective can fail to capture important structure when modes of activity that are highly functionally relevant contribute only a small fraction of the explained variance, and are instead discarded as task-irrelevant noise. Indeed, recent work has shown that dynamics in anterior lateral motor cortex (ALM) during a memory-guided motor task are well-described by task coding directions, which explain a large fraction of the variance in activity, and low-variance residual directions, which ultimately exert outsized behavioral influence (Daie et al., 2023; Pereira-Obilinovic et al., 2026). Why might this "misaligned" coding be favorable?

Motivated by these observations, we study how weight efficiency, activity efficiency, and noise robustness partition the space of recurrent solutions to memory functions, using these desiderata as interpretable axes along which to evaluate solutions. Under an analytically tractable model of stimulus recall, we show that weight efficiency pushes networks towards a minimal, rank-1 solution where stimulus identity, along with input, readout, and recurrent weights align to a common direction, coinciding with the simplicity bias of standard task-trained networks. In contrast, all of these components become misaligned in activity efficient solutions, which instead use high rank recurrent weights to transiently amplify the stimulus in an encoding distributed across many modes of activity. Further, we show that noise robustness constraints set fundamental limits on how activity efficient a solution can be. We demonstrate that these findings extend qualitatively to more complex settings in nonlinear RNNs. Finally, by varying solutions along these axes, we show that pairing noise robustness with an intermediate level of activity efficiency best recapitulates the misaligned coding phenomena described above in a model of anterior lateral motor cortex.

2 The solution space of stimulus recall

We begin by studying the solution space of a simple stimulus recall task (Fig. 1a), which we have formulated to be deliberately minimal so as to aid both analytical tractability and conceptual clarity. This setting isolates the maintenance of task-relevant information across a delay, a basic computational motif underlying memory-guided behavior. Consider a scalar input-output mapping governed by a linear RNN:

$$\dot{\boldsymbol{x}} = -\boldsymbol{x} + \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}u(t), \tag{1}$$

$$y(t) = \boldsymbol{c}^\top \boldsymbol{x}(t), \tag{2}$$

where $\boldsymbol{W} \in \mathbb{R}^{N\times N}$, $\boldsymbol{b} \in \mathbb{R}^N$, and $\boldsymbol{c} \in \mathbb{R}^N$ represent the recurrent, input, and readout weights, respectively. In each trial, the network receives a pulse input $u(t) = z\delta(t)$, where $z \sim P(z)$ represents an observed scalar stimulus, and the network is simply required to reproduce the observed stimulus z after a delay of time τ. Crucially, we do not explicitly constrain the readout $y(t)$ at any time other than the readout time τ. Solving this task then amounts to the condition $y(\tau) = z$ for any stimulus z, or equivalently, requiring that $g(\tau) = \boldsymbol{c}^\top e^{(-\boldsymbol{I}+\boldsymbol{W})\tau}\boldsymbol{b} = 1$, where $g(\tau)$ is the input-to-output gain at time τ.

A natural way to find a solution for this task is to minimize the mean squared error $\mathbb{E}_z[(y(\tau)-z)^2]$ via gradient descent over the parameters $\boldsymbol{\Theta} = \{\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}\}$. More generally however, the solution space for this task is vastly degenerate; minimizing MSE via gradient descent only imposes a particular tiebreaker for this degeneracy. It is well known that gradient descent has an implicit bias towards simple solutions, and in particular solutions with small parameter norm (Neyshabur et al., 2014; Gunasekar et al., 2018). Indeed, although the standard minimum-norm result for linear regression does not apply because the map from $\boldsymbol{\Theta}$ to $y(\tau) = g(\tau)z$ is nonlinear and the resulting MSE objective is nonconvex, we show in App. B.2 that gradient descent from small initialization recovers an aligned, rank-1 solution similar to the minimum norm solution (under similar assumptions as in (Bordelon et al., 2026), detailed in B.2).

However, biological circuits operate under metabolic and energetic constraints that may not necessarily align with the simplicity-seeking bias of gradient descent. In particular, solutions that require large weights or excess network activity are both potentially undesirable. Thus, we consider the

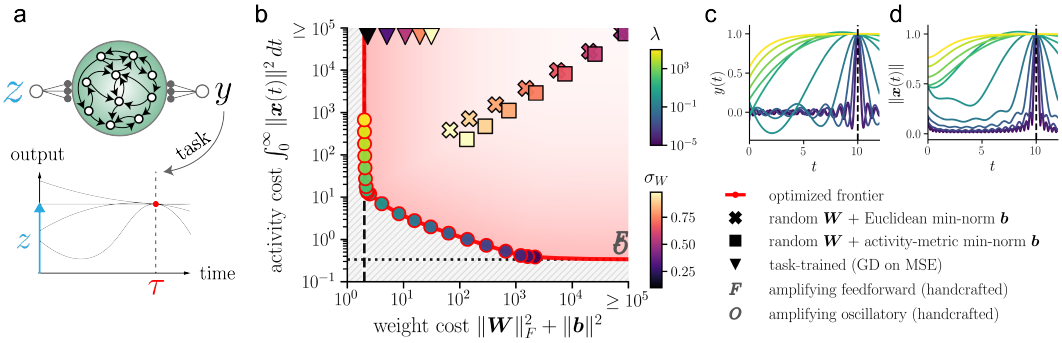


Figure 1: **Characterizing the space of recurrent solutions to stimulus recall by their efficiency. a.** Schematic for the stimulus recall task, showing multiple task-compatible output traces. **b.** Weight and activity costs across solution families, including networks with unstructured recurrent weights, task-trained networks, cost-optimized networks, and handcrafted networks. Where relevant, σ_W indicates the initialization scale of the recurrent weights. The red boundary indicates the optimized frontier obtained over λ. Dashed lines indicate lower bounds on the costs. **c, d.** Readout and activity norm traces across cost-optimized networks at different λ, shown for unit input $z = 1$.

constrained optimization problem

$$\min_{\boldsymbol{W},\boldsymbol{b}} \quad \mathbb{E}_z \left[\int_0^\infty \|\boldsymbol{x}(t)\|^2 dt \right] + \lambda \left(\|\boldsymbol{W}\|_F^2 + \|\boldsymbol{b}\|^2 \right) \qquad \text{s.t. } g(\tau) = 1, \tag{3}$$

where λ is a parameter governing the relative importance of the activity and weight costs. The activity cost is defined with an infinite upper limit to model the total activity over a long horizon, and converges to a finite value so long as $\boldsymbol{W}$ is asymptotically stable (i.e., eigenvalue spectrum has maximum real part less than unity). We fix a choice of readout vector $\boldsymbol{c}$ with $\|\boldsymbol{c}\| = 1$, which can be justified for two reasons: by rotational symmetry of linear RNNs, we can always fix the direction of $\boldsymbol{c}$ without loss of generality. We choose to also fix the readout norm because in the highly activity-constrained regime $\lambda \to 0$, the scaling symmetry $\boldsymbol{c} \to \alpha \boldsymbol{c}$, $\boldsymbol{b} \to \alpha^{-1}\boldsymbol{b}$ would allow for arbitrarily small activity by making the readout vector large and the input vector tiny, thereby making the optimization problem ill-posed. Thus, we require activity efficiency independent of output scale.

This constrained optimization problem can be reduced to the unconstrained optimization $\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W})$, with an objective dependent solely on the recurrent weights $\boldsymbol{W}$ (App. A.2):

$$\mathcal{L}(\boldsymbol{W}) = \lambda \|\boldsymbol{W}\|_F^2 + \frac{1}{\boldsymbol{q}(\tau)^\top (\boldsymbol{R} + \lambda \boldsymbol{I})^{-1} \boldsymbol{q}(\tau)}, \tag{4}$$

where we define the effective dynamics matrix $\boldsymbol{A} = -\boldsymbol{I} + \boldsymbol{W}$, the Gramian matrix $\boldsymbol{R} = \int_0^\infty e^{\boldsymbol{A}^\top t} e^{\boldsymbol{A} t} dt$, the adjoint state $\boldsymbol{q}(\tau) = e^{\boldsymbol{A}^\top \tau} \boldsymbol{c}$, and without loss of generality have taken the stimulus variance as $\sigma_z^2 = 1$. Here, we have integrated out the optimal input weights $\boldsymbol{b}^\star \propto (\boldsymbol{R} + \lambda \boldsymbol{I})^{-1} \boldsymbol{q}(\tau)$, normalized to satisfy $g(\tau) = 1$. For fixed recurrent weights, the second term in Eq. 4 can be interpreted as the minimum combined input weight and activity cost necessary to solve the task.

Although this objective is nonconvex and difficult to optimize analytically in full generality, we can gain intuition for the classes of solutions it favors by considering its two limiting regimes. Where an exact characterization of the global optimum is unavailable, we construct explicit families of solutions that approach optimality in relevant limits and use them to illustrate the dynamical mechanisms associated with near-optimal solutions.

2.1 Weight and activity efficiency favor distinct dynamical regimes

In the limit $\lambda \to \infty$, we recover the minimum norm objective, given by $\mathcal{L}_{\lambda \to \infty}(\boldsymbol{W}) \propto \|\boldsymbol{W}\|_F^2 + \frac{1}{\|\boldsymbol{q}(\tau)\|^2}$. For any delay period τ, the optimal solution is then given by a readout-aligned, symmetric rank-1 spike of the form $W = \alpha \boldsymbol{c}\boldsymbol{c}^\top$, for some eigenvalue $\alpha \in [0, 1)$. In particular, for $\tau \geq 1$, the optimal solution uses marginally stable dynamics, with $\alpha \to 1^-$ (App. B.3). Thus, a single persistent mode governed by line attractor-like dynamics (Seung, 1996) is selected by the minimum norm objective.

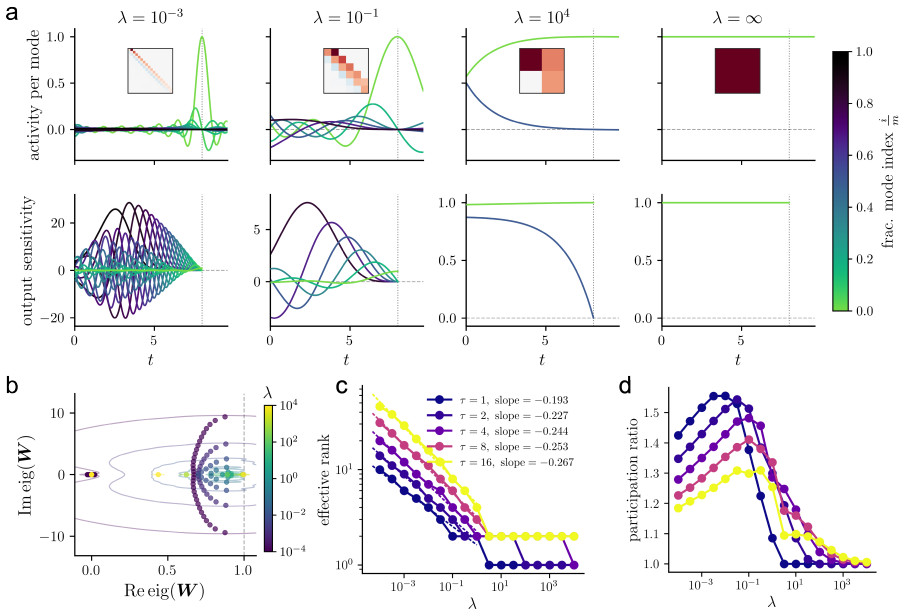


Figure 2: **Activity efficiency promotes high-rank, transiently amplifying solutions characterized by a dissociation between explained variance and causal influence. a.** Top: Per-mode activity of optimized networks at different λ, with modes defined by the Arnoldi basis. Traces are colored by their zero-indexed fractional mode index, so that light green corresponds to the first mode, the readout direction. Insets show the recurrent weights in this basis. Dotted lines indicate the readout time. Bottom: Sensitivity of the output at the readout time with respect to each mode of activity over time. **b.** Eigenvalue spectrum of optimized recurrent weights across λ. The contours indicate pseudospectra (Trefethen, 2005) at $\varepsilon = 0.03$. **c.** Effective rank of the recurrent weights across λ, as estimated by the number of singular values greater than 1e-2. Power law exponents are estimated over $\lambda \leq 1$. **d.** Participation ratio of activity versus λ.

If we instead seek an activity-minimizing solution $\lambda \to 0$, the resulting objective is given by $\mathcal{L}_{\lambda\to 0}(\boldsymbol{W}) = \left[\boldsymbol{q}(\tau)^\top \boldsymbol{R}^{-1}\boldsymbol{q}(\tau)\right]^{-1}$. Intuitively, an activity-efficient solution should seek to transiently amplify activity around the readout time τ, so as to not spend excess activity energy during the delay or after the readout. Moreover, the temporal resolution with which activity can be selectively amplified should be limited by the network size N. Indeed, we show that this activity-minimizing objective is lower bounded as $\mathcal{L}_{\lambda\to 0}(\boldsymbol{W}) \geq \frac{\tau}{N}$ (App. B.4).

A natural candidate for achieving the highly temporally specific transient amplification necessary for saturating this bound is feedforward chain connectivity (Goldman, 2009; Ganguli et al., 2008) of the form $\boldsymbol{A} = -\boldsymbol{I} + \boldsymbol{W} = -a\boldsymbol{I} + k\boldsymbol{\Gamma}$, where $\boldsymbol{\Gamma}_{ij} = \delta_{i+1,j}$, and $a, k > 0$ represent the decay rate and transient gain, respectively (we have chosen a basis where the readout $\boldsymbol{c} = \boldsymbol{e}_1$, the first natural basis vector). With this choice of connectivity, the stimulus encoding is distributed and rotates through N orthogonal directions before aligning with the readout direction. If we choose the decay rate to scale as $a = N/\tau$, and the amplification ratio $\rho = k/a$ to be large, we find that the activity-minimizing objective approaches $\mathcal{L}_{\lambda\to 0}(\boldsymbol{W}) \approx \frac{\pi\tau}{N}$, thus achieving the bound up to a constant factor (App. B.4). Thus, by increasing the length of the chain, this activity cost can be made arbitrarily small. We note that this construction is not unique; see B.5 for a choice of $\boldsymbol{W}$ based on high-dimensional, nonnormal oscillatory dynamics that can also saturate the bound, but up to a different constant factor.

We summarize in Fig. 1b how various families of solutions differ in their respective weight and activity costs. Task-trained networks learn a recurrent mode of long timescale aligned with the input and readout directions, allowing for high weight efficiency at the expense of activity efficiency (App. B.3). On the other hand, networks with high-dimensional nonnormal feedforward structure can be constructed so as to be highly activity efficient but weight inefficient (App. B.4). For solutions with unstructured recurrent weights $\boldsymbol{W}_{ij} \sim \mathcal{N}(0, \sigma_W^2/N)$, both costs are comparably large, as such networks require large input vectors (and therefore large activity) so that the stimulus can still be read

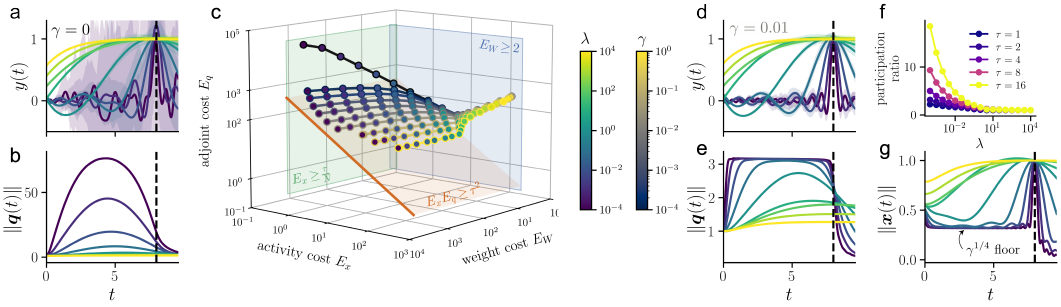


Figure 3: **Robustness to noise requires trading off against activity efficiency. a.** Output traces for optimized networks after dynamics are corrupted with process noise. Shaded regions indicate confidence intervals. **b.** Adjoint norm predicts noise sensitivity. **c.** Cost properties of networks optimized across both λ and γ. The black trace corresponds to $\gamma = 0$ (no adjoint penalty). Planes indicate theoretical bounds that constrain the solution space. **d-g.** Properties of networks optimized at nonzero adjoint cost $\gamma = 0.01$. **d.** Output traces under process noise no longer exhibit extreme noise sensitivity. Adjoint norms are constrained (**e**), and activity norms no longer approach zero during the delay (**g**). **f.** Effective dimensionality of activity now grows as λ approaches zero.

out at the proper scale by the readout time (App. B.1). Finally, solving the constrained optimization in Eq. 4 at different λ yields the Pareto frontier of solutions for this task (Fig. 1c,d).

The optimal solutions selected by both limits of λ are diametrically opposite extremes: the minimum norm solution uses a single persistent attracting mode, yet is highly activity inefficient, while the minimum activity solution is well-approximated by an arbitrarily high-rank chain with large amplification ratio, and thus is highly weight inefficient. These two extremes, while each pathological in their own way, are useful baselines for grounding the distinct inductive biases imposed by these normative objectives. At small but nonzero λ, it is natural to expect that the optimal solution would lie between these extremes, with a dimensionality controlled by λ. If we again consider a chain ansatz for $\boldsymbol{W}$, but with an intermediate length $m \leq N$, the resulting activity term still approximately scales as $\left[\boldsymbol{q}^\top (\boldsymbol{R} + \lambda \boldsymbol{I})^{-1} \boldsymbol{q}\right]^{-1} \sim \frac{\tau}{m}$, whereas the weight cost scales as $\|\boldsymbol{W}\|_F^2 = m[(1-a)^2 + k^2] \sim \frac{m^3}{\tau^2}$. Balancing these costs yields an optimal chain length scaling as $m^\star \sim \lambda^{-1/4} \tau^{3/4}$.

2.2 Activity-efficient solutions exhibit misaligned coding

While these scaling exponents are not expected to be universal, and arise from a simplified ansatz for $\boldsymbol{W}$, the conclusion that an extensive number of recurrent modes can be optimal to solve this simple memory task is robust to these modeling choices. Indeed, numerically optimizing the joint objective $\mathcal{L}(\boldsymbol{W})$ (App. D.1) at different λ yields a spectrum of increasingly high dimensional solutions as λ approaches zero (Fig. 2). For each numerically optimized solution $\boldsymbol{W}$, we obtain an informative basis for understanding the properties of the solution via Arnoldi iteration, i.e., orthonormalizing the Krylov vectors $[\boldsymbol{c}, \boldsymbol{W}^\top \boldsymbol{c}, \left(\boldsymbol{W}^\top\right)^2 \boldsymbol{c}, \ldots \left(\boldsymbol{W}^\top\right)^{m-1} \boldsymbol{c}]$ generated by the readout direction, where m is the effective rank of the recurrent weights. Viewed under this basis, we find that highly activity efficient networks can indeed be described as having nonnormal, amplifying chain-like connectivity, but with feedback connections rather than purely feedforward. Interestingly, despite the high dimensional nature of highly activity-constrained solutions, this structure is not immediately apparent from activity. In particular, the effective dimensionality of activity $\{\boldsymbol{x}(t)\}$ remains $\mathcal{O}(1)$ across the whole range of λ (Fig. 2d). Continuing with the chain analogy, this could be explained by the fact that such solutions produce activity that is amplified exponentially as it propagates along the chain, so that the activity along the readout-aligned mode still remains the dominant contribution of explained variance. To quantify causal influence of each mode of activity $\boldsymbol{v}_i$ on network output, we consider perturbations $\delta \boldsymbol{x}(t) = \epsilon \boldsymbol{v}_i$, and compute $\partial_\varepsilon y(\tau)|_{\epsilon=0} = \boldsymbol{v}_i^\top \boldsymbol{q}(\tau - t)$ during the delay. Consistent with the feedforward chain analogy, we find that a dissociation between modes of activity that explain the most variance and those that maximally causally influence output emerges at low λ (Fig. 2a). Thus, such misaligned coding can arise simply from optimizing for activity efficiency.

2.3 Noise robustness competes with activity efficiency

Another common desideratum for biological circuits is robustness to noise. Unlike in feedforward networks, RNNs subject to process noise must contend with potential noise amplification arising

from the successive application of the same recurrent weight matrix across timesteps. It is thus natural to ask whether optimizing for activity efficiency, which encourages temporally selective transient amplification of inputs, has the adverse consequence of increasing sensitivity to noise. Indeed, we find that networks optimized to solve the stimulus recall task in a highly activity efficient manner ($\lambda \ll 1$) were most sensitive to process noise (Fig. 3a).

If we augment the RNN's dynamics with white process noise as $\frac{d\boldsymbol{x}}{dt} = -\boldsymbol{x} + \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}u(t) + \boldsymbol{\eta}(t)$, where $\boldsymbol{\eta} \in \mathbb{R}^N$ satisfies $\mathbb{E}[\boldsymbol{\eta}(t)] = 0$ and $\mathbb{E}[\boldsymbol{\eta}(t)\boldsymbol{\eta}(t')^\top] = \sigma_\eta^2 \boldsymbol{I}\delta(t-t')$, this induces noise in the output, with variance given by $\mathrm{Var}_\eta[y(\tau)] = \sigma_\eta^2 \boldsymbol{c}^\top \left(\int_0^\tau e^{\boldsymbol{A}t} e^{\boldsymbol{A}^\top t} dt \right) \boldsymbol{c} = \sigma_\eta^2 \int_0^\tau \|\boldsymbol{q}(t)\|^2 dt$. Thus, process noise induces noise in the readout with variance proportional to the integrated squared norm of the adjoint dynamics variable $\boldsymbol{q}(t)$. To quantify sensitivity to noise in a task-agnostic (horizon-free) manner, it is therefore natural to consider an "adjoint activity" given by the quantity

$$E_q = \int_0^\infty \|\boldsymbol{q}(t)\|^2 dt, \tag{5}$$

which measures the stationary variance along the readout due to isotropic process noise, and upper bounds the readout noise for any finite time horizon.

This apparent tradeoff between activity efficiency and noise robustness can be made precise by observing a standard conservation relation (Luchini & Bottaro, 2014) between the time-reversed adjoint $\widetilde{\boldsymbol{q}}(t) = \boldsymbol{q}(\tau - t)$, defined for $t \in [0, \tau]$, and the activity variable $\widetilde{\boldsymbol{x}}(t) = \boldsymbol{x}(t)/z = e^{\boldsymbol{A}t}\boldsymbol{b}$. In particular, we have that $\frac{d}{dt}\widetilde{\boldsymbol{q}}(t)^\top \widetilde{\boldsymbol{x}}(t) = \frac{d}{dt}(\boldsymbol{c}^\top e^{\boldsymbol{A}(\tau-t)} e^{\boldsymbol{A}(t)} \boldsymbol{b}) = 0$, and thus $\widetilde{\boldsymbol{q}}(t)^\top \widetilde{\boldsymbol{x}}(t) = \widetilde{\boldsymbol{q}}(0)^\top \widetilde{\boldsymbol{x}}(0) = 1$ for all times $t \in [0, \tau]$, where the latter equality follows from the task constraint $g(\tau) = \boldsymbol{c}^\top e^{\boldsymbol{A}\tau}\boldsymbol{b} = 1$. Cauchy-Schwarz then yields the uncertainty relation-like bound

$$E_x E_q \geq \tau^2, \tag{6}$$

where $E_x = \int_0^\infty \|\widetilde{\boldsymbol{x}}(t)\|^2 dt$ represents the average activity cost. In other words, minimizing activity necessarily demands that the noise sensitivity of the readout, as measured by the adjoint "activity", grows to compensate, and vice versa. For the amplifying feedforward chain, this divergence of adjoint cost as the network becomes increasingly activity efficient can be shown explicitly (App. C).

We remark that this notion of noise sensitivity based on adjoint norm measures absolute noise scale, as opposed to a relative measure such as signal-to-noise ratio. This is appropriate for our setting, where the task constraint has fixed the scale of the "signal" as $g(\tau) = 1$. Moreover, a mean-squared error style task objective of the form $\mathbb{E}_{z,\eta}[(y(\tau) - z)^2]$ decomposes additively as $\sigma_z^2 \left(g(\tau) - 1\right)^2 + \sigma_\eta^2 \int_0^\tau \|\boldsymbol{q}(t)\|^2 dt$, thus making this a relevant notion of noise sensitivity for task-optimized RNNs as well. However, in settings where signal-to-noise ratio is optimized instead, amplifying nonnormal connectivity similar to that discussed here, paired with a judicious choice of input weights, can be optimal for extensive memory, as both signal and noise are amplified (Ganguli et al., 2008).

If we now introduce the adjoint cost to the constrained optimization as

$$\min_{\boldsymbol{W},\boldsymbol{b}} E_x + \lambda E_W + \gamma E_q \qquad \text{s.t. } g(\tau) = 1, \tag{7}$$

with weighting γ, we can damp the extreme sensitivity of solutions to process noise by tuning γ. In the limit of vanishing weight cost $\lambda \to 0$, optimal solutions hug the lower bound $E_x E_q \geq \tau^2$ (Fig. 3e,g). Together with the approximately constant activity and adjoint norms during the delay observed numerically, this yields an activity scale $\|\boldsymbol{x}(t)\| \approx \gamma^{1/4}$ throughout the delay (Fig. 3g, App. C.2), in contrast to the vanishing delay activity observed before in Fig. 1d. Moreover, modes of activity orthogonal to the readout direction become resolved rather than suppressed to infinitesimal magnitude (Fig. E.1), leading to an effective dimensionality of activity that grows with decreasing λ and tracks the increase in recurrent rank, rather than staying $\mathcal{O}(1)$ (Fig. 3f).

3 Efficiency-robustness tradeoffs persist in nonlinear RNNs

Thus far, we have considered linear RNNs subject to a temporally pointwise constraint on the network output, a tractable albeit restrictive setting. To test whether the qualitative tradeoffs identified above depend on these assumptions, we next consider nonlinear RNNs whose output is constrained over a continuous time window. Specifically, we train RNNs with $\phi = \tanh$ activation to solve a variant of the constrained optimization problem in Eq. 3, but with the activ-

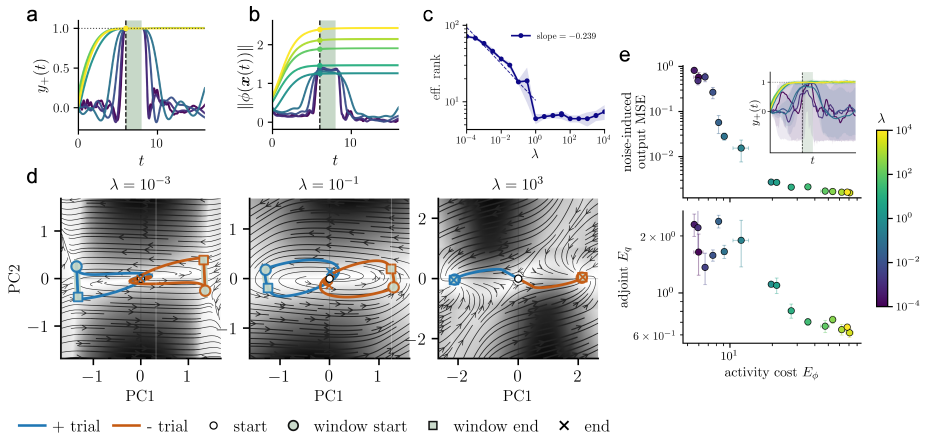


Figure 4: **Nonlinear RNNs exhibit similar changes in recurrent dynamics and noise sensitivity with increasing activity efficiency. a-b.** Output and activity over time across different λ, shown for a $+$ trial. The output is only constrained during the window highlighted in green. **c.** Effective rank of learned solutions across λ, averaged over five seeds. **d.** Flow fields projected onto the top two principal components of activity, with $+/-$ trial trajectories overlaid. Darkness of the background indicates flow velocity. **e.** Top: Output error due to injected process noise versus activity cost. Inset shows noised readouts for a $+$ trial. Each point is an average over seeds, with error bars indicating standard error of the mean. Bottom: Adjoint versus activity cost.

ity cost $E_\phi = \mathbb{E}_z\left[\int_0^T \|\phi(\boldsymbol{x}(t))\|^2 dt\right]$ evaluated over a finite horizon T, the readout constraint $y(t) = \boldsymbol{c}^\top \phi(\boldsymbol{x}(t)) \approx z$ applied over $t \in [\tau, \tau + \xi]$, and the stimulus drawn discretely as $z \sim \text{Unif}(\{-1, +1\})$ (App. D.2). As it is difficult to solve this constrained optimization problem for nonlinear RNNs even numerically, we approximate doing so by adding the weight and activity costs as soft penalties, and assigning a large relative weighting to a soft task cost $\mathbb{E}_z\left[\int_\tau^{\tau+\xi}(y(t) - z)^2 dt\right]$ (App. D.2). Unlike linear RNNs, nonlinear RNNs can store memories of discrete stimuli across multiple stable fixed points. Indeed, we find that weight efficiency constrained networks find a low-rank solution that stores stimulus identity in distinct stable fixed points, as is typically observed in task-trained RNNs performing memory tasks (Sussillo & Barak, 2013; Maheswaranathan et al., 2019; Driscoll et al., 2024). However, below a critical value of λ, solutions that rely on stable fixed points disappear (Fig. 4). Instead, highly activity-constrained solutions learn high rank recurrent weights that produce choreographed transient trajectories, with a period of high activity aligned only to the readout window. Mirroring our findings in the simplified linear setting, we again find that the effective rank of trained solutions grows approximately as a power law with decreasing λ (Fig. 4). Thus, despite the richer dynamical repertoire available to nonlinear RNNs, activity efficient solutions still favor high rank recurrent weights.

Next, we assess how network performance, as measured by mean-squared error of the output, degrades with injected neuronal noise. Consistent with our analysis of linear RNNs, we observe that greater activity efficiency is associated with greater sensitivity to noise (Fig. 4e). While the conservation relation underlying Eq. 6, and hence an exact uncertainty bound, does not hold in general in the nonlinear setting, the cumulative squared norm of the adjoint variable $\tilde{\boldsymbol{q}}(t) = \frac{\partial y(\tau)}{\partial \boldsymbol{x}(t)}$ still scales roughly inversely to activity cost (Fig. 4e).

3.1 Efficiency and robustness recapitulate misaligned coding in ALM

Finally, we ask whether the same efficiency and robustness axes can recapitulate key dynamical properties of anterior lateral motor cortex during a memory-guided motor task. As in Pereira-Obilinovic et al. (2026), we use networks with positive firing rates and learnable maximum firing rates per neuron:

$$\dot{\boldsymbol{x}}_i = -\boldsymbol{x}_i + \sum_j \boldsymbol{W}_{ij}\phi_j(\boldsymbol{x}_j) + \sum_a \boldsymbol{B}_{ia}\boldsymbol{u}_a(t) + \boldsymbol{b}_i^0 + \boldsymbol{\eta}_i \tag{8}$$

$$y(t) = \sum_i \boldsymbol{c}_i \phi_i(\boldsymbol{x}_i(t)), \tag{9}$$

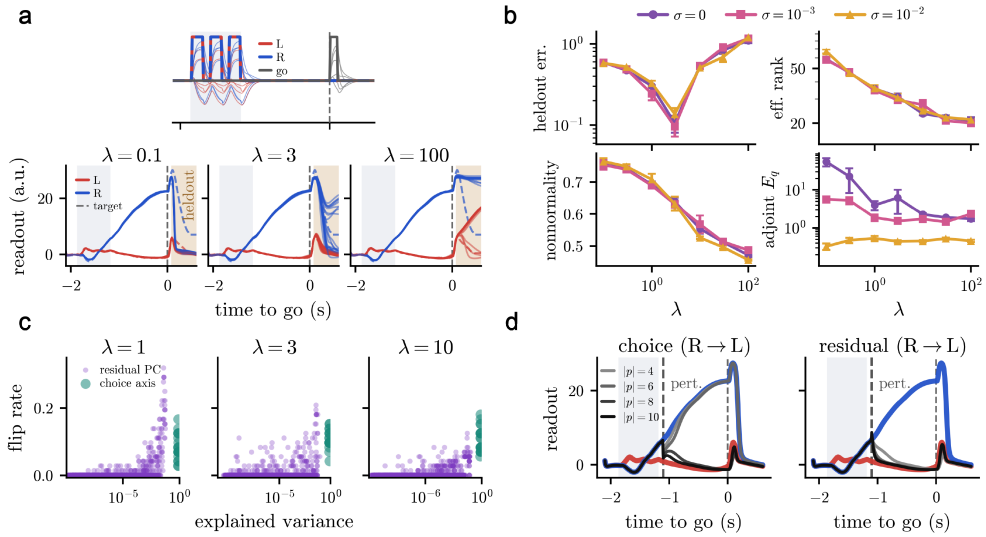


Figure 5: **An intermediate level of activity efficiency, rather than the minimum norm solution, recapitulates signatures of misaligned coding in anterior lateral motor cortex. a.** Top: Schematic of the task input. Low dimensional input pulses corresponding to left trial, right trial, and go cues are filtered by a fixed random reservoir before being passed as input to the network. Bottom: Readout traces across networks trained at different λ, shown for 20 random seeds. **b.** Properties of solutions trained at different process noise levels σ and weight cost levels λ. **c.** Rate of flipping for perturbations aligned with residual principal components (purple) and choice axis (green), pooled over all 20 trained seeds. **d.** Example perturbation-induced responses for a representative network trained at $\lambda = 1$ and $\sigma = 1e-2$, shown for different perturbation magnitudes.

with saturating nonlinearity $\phi_i(x) = \frac{r_i^{\max}}{2}(\tanh(x) + 1)$. Here, $\boldsymbol{B}$ denotes the input weights, $\boldsymbol{b}^0$ denotes trainable biases, and $\boldsymbol{\eta}$ is isotropic process noise. The task consists of "left" or "right" sample cues, a fixed delay, and then a subsequent "go" cue, after which the network should produce a response corresponding to the sample cue that was observed (Fig. 5a). These low-dimensional task inputs are first filtered through a fixed random reservoir to produce the high-dimensional input $\boldsymbol{u}(t)$ that is passed into the RNN (App. D.3). Here, rather than densely supervising the RNNs against recorded neural activity at the single-unit level, we ask if much weaker data supervision, along with some combination of the desiderata of weight efficiency, activity efficiency, and noise robustness could yield networks with qualitatively similar phenomenology as observed in Pereira-Obilinovic et al. (2026). Specifically, we trained networks so that a readout of the firing rates reproduces the corresponding condition-averaged choice projection of neural activity recorded from anterior lateral motor cortex in mice performing the same task. In this manner, networks are constrained only through a one-dimensional, task-relevant projection of recorded activity. Importantly, this target displays a complex time-dependent profile with stimulus-dependent preparatory ramping, rather than simply tracking the stimulus. Among all such task-solving networks, we vary the weight-to-activity cost ratio λ and the level of noise to determine their effect on the dynamical properties of resulting solutions (App. D.3). We penalize recurrent and input weight norms as before, here $\|\boldsymbol{W}\|_F^2 + \|\boldsymbol{B}\|_F^2$, but do not penalize the biases. In addition, we fix the norm of the readout $\boldsymbol{c}$ as $\|\boldsymbol{c}\| = 1$, but allow its direction to be trainable.

To assess heldout predictivity, we supervised networks only up to a time just after the onset of the go cue, leaving the response period unconstrained. We find that both extremes of λ led to networks that were poor predictors of the heldout targets (Fig. 5b); highly activity-constrained networks produced readouts that decayed too quickly during the response period, whereas highly weight-constrained networks produced readouts that saturated and failed to decay at all. We also evaluated network performance on catch trials where the go cue was delayed or withheld, finding that training in the presence of process noise was necessary for networks to maintain stable choice-selective activity throughout the extended delay, consistent with experimental catch trials (Pereira-Obilinovic et al., 2026)(Fig. E.2). We next assessed whether perturbations during the delay could flip the trained networks' readout trajectories from one outcome (i.e. left) to the other. For each network, we estimate sample, choice, and response directions by computing the average firing rate difference

$\langle \boldsymbol{r}^R(t) - \boldsymbol{r}^L(t) \rangle$ between left and right trials over the sample, late-delay, and post-go cue periods, respectively (App. D.3.1). We then defined residual directions as principal components of activity orthogonal to the subspace spanned by these task coding dimensions. Perturbing along both coding and residual dimensions, we find that decreasing λ, i.e., increasing activity efficiency, is associated with the emergence of residual directions with high causal influence, as measured by flip rate (Fig. 5c,d, E.3). Further, consistent with prior analyses, we find that such highly activity efficient solutions require high rank, nonnormal recurrent weights (Fig. 5b), mirroring the nonnormal amplification implicated in the causal influence of residual directions found by Pereira-Obilinovic et al. (2026). Overall, these results suggest that an intermediate level of activity efficiency, paired with robustness to process noise, is sufficient to recapitulate both the temporal structure of choice-related activity and the disproportionate causal influence of low-variance residual dimensions observed in ALM.

4 Discussion

Understanding how different forms of efficiency and robustness shape the space of solutions in ANNs is crucial for identifying and improving upon gaps in their analogy to biological neural circuits. In feedforward linear networks, constraints on the solution space imposed by minimizing total representation norm (activity) across layers are largely nested within those imposed by minimizing total weight norm (Braun et al., 2025). Moreover, an objective that penalizes weight norm and/or representation norm arises naturally from optimizing for robustness (in the mean-squared error sense) of solutions to additive perturbations (Bishop, 1995; Braun et al., 2025; Dorrell et al., 2026). In contrast, we show that in recurrent neural networks, these desiderata of weight efficiency, activity efficiency, and noise robustness are in general in competition with each other, with each imposing distinct pressures on selected solutions. Although our analytical picture of these tradeoffs originates from a simplified setting of a linear RNN solving a simple memory task, we show that qualitatively similar tradeoffs arise even when dynamics are nonlinear, activations are bounded and positive, and targets have complex time-dependent profiles.

While solutions where task-relevant dynamics evolve in low-dimensional subspaces have been the focus of much of the computation-through-dynamics literature (Mante et al., 2013; Mastrogiuseppe & Ostojic, 2018; Valente et al., 2022), we have identified activity efficiency as a biologically-relevant, interpretable scaling axis along which non-minimal, higher-dimensional solutions to even simple tasks may emerge. Further, such non-minimal solutions arising from optimizing for an intermediate level of activity efficiency appear to be more consistent with misaligned coding phenomena observed in anterior lateral motor cortex.

Our work connects to prior studies linking recurrent dynamics supporting memory to underlying optimization principles (Ganguli et al., 2008; Lim & Goldman, 2011; Stroud et al., 2024; Ritter & Chadwick, 2025; Stroud et al., 2025; Dorrell et al., 2026). In the context of working memory, Stroud et al. (2023) argued that inputs should be optimally loaded along directions that best take advantage of nonnormal structure in recurrent weights, and that such a scheme is consistent with neural recordings. Ritter & Chadwick (2025) find that working memory networks optimized for signal-to-noise ratio and activity efficiency learn high-dimensional, nonnormal rotational dynamics, with qualitative similarities to the activity-efficient regime described here. Our results complement these studies by characterizing the interplay between weight, activity, and robustness costs in a setting where key aspects of their effects on recurrent weights and dynamics can be described analytically.

The present study has several limitations. First, we focus on simple memory tasks where only a single variable needs to be remembered to inform a later response. Correlations among stimuli and representational interference across time in sequence memory-style tasks (Sharma et al., 2025; Dorrell et al., 2026) could substantially alter the nature of tradeoffs identified here. In addition, while we focus on particular forms of activity and weight efficiency, other desiderata may be more biologically relevant. For instance, due to homeostatic pressures, it may be more relevant to measure activity cost by the deviation from some baseline level of activity, rather than by overall scale (Chintaluri & Vogels, 2023). Finally, while we focused on how these different costs shape optimal solutions, we did not address the learning process that gives rise to such optimally efficient or robust solutions. Weight efficiency has a natural implementation as a local learning rule in the form of weight decay. It is less clear, however, how local learning rules would construct the high-rank, nonnormal connectivity favored by activity efficiency. Identifying plasticity rules that construct solutions favored by these less standard objectives is an important direction for future work.

AI USE STATEMENT

We used generative AI tools, including ChatGPT/Codex (5.4, 5.5) and Claude Code (Opus 4.8), to 1) assist with numerical simulations and plotting, 2) check for correctness of analytic results, and 3) trim and polish text. We also credit LLMs for useful discussions regarding asymptotics relevant to the activity efficiency analysis of the feedforward chain in App. B.4. All results were carefully reviewed by the authors, who take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Assumptions underlying analytic results are outlined in Appendices A, B, and C. We further clarify both in the main text and in the Appendix which results are heuristic and which are exact. Settings, optimizers, and hyperparameter choices for reproducing numerical experiments and analyses are detailed in Appendix D.

REFERENCES

Joseph J. Atick and A. Norman Redlich. Towards a theory of early visual processing. *Neural Computation*, 2(3):308–320, 1990. ISSN 1530-888X. doi: 10.1162/neco.1990.2.3.308.

Fred Attneave. Some informational aspects of visual perception. *Psychological Review*, 61(3): 183–193, 1954. ISSN 1939-1471. doi: 10.1037/h0054663.

Chris M. Bishop. Training with Noise is Equivalent to Tikhonov Regularization. *Neural Computation*, 7(1):108–116, January 1995. ISSN 0899-7667. doi: 10.1162/neco.1995.7.1.108.

Blake Bordelon, Jordan Cotler, Cengiz Pehlevan, and Jacob A. Zavatone-Veth. Dynamics of learning to integrate in linear recurrent neural networks, June 2026.

Lukas Braun, Erin Grant, and Andrew M. Saxe. Not all solutions are created equal: An analytical dissociation of functional and representational similarity in deep linear neural networks. In *Forty-Second International Conference on Machine Learning*, June 2025.

Chaitanya Chintaluri and Tim P. Vogels. Metabolically regulated spiking could serve neuronal energy homeostasis and protect from reactive oxygen species. *Proceedings of the National Academy of Sciences*, 120(48):e2306525120, November 2023. doi: 10.1073/pnas.2306525120.

David G. Clark, Blake Bordelon, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. Structure, disorder, and dynamics in task-trained recurrent neural circuits. *bioRxiv*, pp. 2026.03.02.708943, May 2026. ISSN 2692-8205. doi: 10.64898/2026.03.02.708943.

Kayvon Daie, Lorenzo Fontolan, Shaul Druckmann, and Karel Svoboda. Feedforward amplification in recurrent networks underlies paradoxical neural coding. *bioRxiv*, pp. 2023.08.04.552026, August 2023. doi: 10.1101/2023.08.04.552026.

William Dorrell, Peter E. Latham, Timothy E. J. Behrens, and James C. R. Whittington. An Efficient Computing Theory of Prefrontal Structured Working Memory Representations. *bioRxiv*, pp. 2026.02.16.706126, March 2026. ISSN 2692-8205. doi: 10.64898/2026.02.16.706126.

Laura N. Driscoll, Krishna Shenoy, and David Sussillo. Flexible multitask computation in recurrent networks utilizes shared dynamical motifs. *Nature Neuroscience*, 27(7):1349–1363, July 2024. ISSN 1546-1726. doi: 10.1038/s41593-024-01668-6.

P. J. Forrester, N. E. Frankel, and T. M. Garoni. Asymptotic form of the density profile for Gaussian and Laguerre random matrix ensembles with orthogonal and symplectic symmetry. https://arxiv.org/abs/math-ph/0508031v1, August 2005.

Surya Ganguli, Dongsung Huh, and Haim Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 105(48):18970–18975, December 2008. ISSN 1091-6490. doi: 10.1073/pnas.0804451105.

Mark S. Goldman. Memory without Feedback in a Neural Network. *Neuron*, 61(4):621–634, February 2009. ISSN 0896-6273. doi: 10.1016/j.neuron.2008.12.012.

Suriya Gunasekar, Jason Lee, Daniel Soudry, and Nathan Srebro. Characterizing Implicit Bias in Terms of Optimization Geometry. In *International Conference on Machine Learning*, pp. 1832–1841. PMLR, July 2018.

Ann Huang, Satpreet Harcharan Singh, Flavio Martinelli, and Kanaka Rajan. Measuring and Controlling Solution Degeneracy across Task-Trained Recurrent Neural Networks. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, October 2025.

Yan Karklin and Eero P. Simoncelli. Efficient coding of natural images with a population of noisy Linear-Nonlinear neurons. *Advances in neural information processing systems*, 24:999–1007, December 2011. ISSN 1049-5258.

Bariscan Kurtkaya, Fatih Dinc, Mert Yuksekgonul, Marta Blanco-Pozo, Ege Cirakman, Mark Schnitzer, Yucel Yemez, Hidenori Tanaka, Peng Yuan, and Nina Miolane. Dynamical phases of short-term memory mechanisms in RNNs. In *Forty-Second International Conference on Machine Learning*, June 2025.

Sukbin Lim and Mark S. Goldman. Noise Tolerance of Attractor and Feedforward Memory Models. *Neural computation*, 24(2):332, November 2011. doi: 10.1162/NECO_a_00234.

Paolo Luchini and Alessandro Bottaro. Adjoint Equations in Stability Analysis. pp. 493–517, January 2014. doi: 10.1146/annurev-fl-46-011514-200001.

Niru Maheswaranathan, Alex Williams, Matthew Golub, Surya Ganguli, and David Sussillo. Universality and individuality in neural dynamics across large populations of recurrent networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Valerio Mante, David Sussillo, Krishna V. Shenoy, and William T. Newsome. Context-dependent computation by recurrent dynamics in prefrontal cortex. *Nature*, 503(7474):78–84, November 2013. ISSN 1476-4687. doi: 10.1038/nature12742.

Francesca Mastrogiuseppe and Srdjan Ostojic. Linking Connectivity, Dynamics, and Computations in Low-Rank Recurrent Neural Networks. *Neuron*, 99(3):609–623.e29, August 2018. ISSN 0896-6273. doi: 10.1016/j.neuron.2018.07.003.

Christian Mehl, Volker Mehrmann, and Punit Sharma. Stability Radii for Linear Hamiltonian Systems with Dissipation Under Structure-Preserving Perturbations. *SIAM Journal on Matrix Analysis and Applications*, 37(4):1625–1654, January 2016. ISSN 0895-4798. doi: 10.1137/16M1067330.

Asma Motiwala, Sofia Soares, Bassam V. Atallah, Joseph J. Paton, and Christian K. Machens. Efficient coding of cognitive variables underlies dopamine response and choice behavior. *Nature Neuroscience*, 25(6):738–748, June 2022. ISSN 1546-1726. doi: 10.1038/s41593-022-01085-7.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. *International Conference on Learning Representations Workshop Track*, December 2014.

B. A. Olshausen and D. J. Field. Natural image statistics and efficient coding. *Network: Computation in Neural Systems*, 7(2):333, May 1996. ISSN 0954-898X. doi: 10.1088/0954-898X/7/2/014.

Bruno A. Olshausen and David J. Field. Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research*, 37(23):3311–3325, December 1997. ISSN 0042-6989. doi: 10.1016/S0042-6989(97)00169-7.

Mitchell Ostrow, Adam J. Eisen, Leo Kozachkov, William T. Redman, and Ila Fiete. A metric for comparing complex systems by their dynamics. *bioRxiv*, pp. 2026.07.16.738953, July 2026. ISSN 2692-8205. doi: 10.64898/2026.07.16.738953.

Marino Pagan, Vincent D. Tang, Mikio C. Aoi, Jonathan W. Pillow, Valerio Mante, David Sussillo, and Carlos D. Brody. Individual variability of neural computations underlying flexible decisions. *Nature*, 639(8054):421–429, March 2025. ISSN 1476-4687. doi: 10.1038/s41586-024-08433-6.

Ulises Pereira-Obilinovic, Kayvon Daie, Susu Chen, Karel Svoboda, and Ran Darshan. Neural dynamics outside task-coding dimensions drive decision trajectories through transient amplification. *bioRxiv*, pp. 2025.11.20.689599, April 2026. ISSN 2692-8205. doi: 10.1101/2025.11.20.689599.

William Qian and Cengiz Pehlevan. Discovering alternative solutions beyond the simplicity bias in recurrent neural networks. In *The Fourteenth International Conference on Learning Representations*, October 2025.

Kanaka Rajan, Christopher D. Harvey, and David W. Tank. Recurrent Network Models of Sequence Generation and Memory. *Neuron*, 90(1):128–142, April 2016. ISSN 0896-6273. doi: 10.1016/j.neuron.2016.02.009.

Blake A. Richards, Timothy P. Lillicrap, Philippe Beaudoin, Yoshua Bengio, Rafal Bogacz, Amelia Christensen, Claudia Clopath, Rui Ponte Costa, Archy de Berker, Surya Ganguli, Colleen J. Gillon, Danijar Hafner, Adam Kepecs, Nikolaus Kriegeskorte, Peter Latham, Grace W. Lindsay, Kenneth D. Miller, Richard Naud, Christopher C. Pack, Panayiota Poirazi, Pieter Roelfsema, João Sacramento, Andrew Saxe, Benjamin Scellier, Anna C. Schapiro, Walter Senn, Greg Wayne, Daniel Yamins, Friedemann Zenke, Joel Zylberberg, Denis Therien, and Konrad P. Kording. A deep learning framework for neuroscience. *Nature Neuroscience*, 22(11):1761–1770, November 2019. ISSN 1546-1726. doi: 10.1038/s41593-019-0520-2.

Laura Ritter and Angus Chadwick. Efficient Working Memory Maintenance via High-Dimensional Rotational Dynamics, September 2025. ISSN 2692-8205.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J. Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *bioRxiv*, pp. 407007, January 2020. doi: 10.1101/407007.

Friedrich Schuessler, Francesca Mastrogiuseppe, Alexis Dubreuil, Srdjan Ostojic, and Omri Barak. The interplay between randomness and structure during learning in RNNs. In *Advances in Neural Information Processing Systems*, volume 33, pp. 13352–13362. Curran Associates, Inc., 2020.

Heiko H. Schütt, Dongjae Kim, and Wei Ji Ma. Reward prediction error neurons implement an efficient code for reward. *Nature Neuroscience*, 27(7):1333–1339, July 2024. ISSN 1546-1726. doi: 10.1038/s41593-024-01671-x.

H. S. Seung. How the brain keeps the eyes still. *Proceedings of the National Academy of Sciences*, 93(23):13339–13344, November 1996. doi: 10.1073/pnas.93.23.13339.

Pratyaksh Sharma, Alexandra Maria Proca, Lucas Prieto, and Pedro A. M. Mediano. Temporal superposition and feature geometry of RNNs under memory demands. In *The Fourteenth International Conference on Learning Representations*, October 2025.

Ben Sorscher, Gabriel Mel, Surya Ganguli, and Samuel Ocko. A unified theory for the origin of grid cells through the lens of pattern formation. *Advances in Neural Information Processing Systems*, 32, 2019.

Jake P. Stroud, Kei Watanabe, Takafumi Suzuki, Mark G. Stokes, and Máté Lengyel. Optimal information loading into working memory explains dynamic coding in the prefrontal cortex. *Proceedings of the National Academy of Sciences*, 120(48):e2307991120, November 2023. doi: 10.1073/pnas.2307991120.

Jake P. Stroud, John Duncan, and Máté Lengyel. The computational foundations of dynamic coding in working memory. *Trends in Cognitive Sciences*, 28(7):614–627, July 2024. ISSN 1364-6613, 1879-307X. doi: 10.1016/j.tics.2024.02.011.

Jake Patrick Stroud, Michal Wojcik, Kristopher Torp Jensen, Makoto Kusunoki, Mikiko Kadohisa, Mark J. Buckley, John Duncan, Mark G. Stokes, and Mate Lengyel. Effects of noise and metabolic cost on cortical task representations. https://elifesciences.org/articles/94961, January 2025.

David Sussillo and Omri Barak. Opening the Black Box: Low-Dimensional Dynamics in High-Dimensional Recurrent Neural Networks. *Neural Computation*, 25(3):626–649, March 2013. ISSN 0899-7667. doi: 10.1162/NECO_a_00409.

Lloyd N. Trefethen. *Spectra and Pseudospectra: The Behavior of Nonnormal Matrices and Operators*. Princeton University Press, Princeton, 2005. ISBN 978-0-691-11946-5 978-0-691-21310-1.

Elia Turner, Kabir V. Dabholkar, and Omri Barak. Charting and Navigating the Space of Solutions for Recurrent Neural Networks. *Advances in Neural Information Processing Systems*, 34:25320–25333, December 2021.

Adrian Valente, Jonathan W. Pillow, and Srdjan Ostojic. Extracting computational mechanisms from neural data using low-rank RNNs. *Advances in Neural Information Processing Systems*, 35: 24072–24086, December 2022.

James C. R. Whittington, Will Dorrell, Surya Ganguli, and Timothy Behrens. Disentanglement with Biological Constraints: A Theory of Functional Cell Types. In *The Eleventh International Conference on Learning Representations*, September 2022.

Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, Darren Seibert, and James J. DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, June 2014. doi: 10.1073/pnas.1403112111.

A Problem formulation

A.1 General formulation and reduction

We first formulate the constrained optimization problem for a general distribution of readout times and a finite task-error tolerance. This makes clear which aspects of the reduction are generic, before specializing to the fixed-readout-time setting analyzed in the main text.

Consider a single-input single-output mapping governed by a linear RNN:

$$\dot{\boldsymbol{x}} = -\boldsymbol{x} + \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}u(t) \tag{10}$$

$$y(t) = \boldsymbol{c}^\top \boldsymbol{x}(t) \tag{11}$$

where $\boldsymbol{W} \in \mathbb{R}^{N\times N}$, $\boldsymbol{b} \in \mathbb{R}^N$, and $\boldsymbol{c} \in \mathbb{R}^N$.

On each trial, the network sees a pulse input $u(t) = z\delta(t)$, where $z \sim \mathcal{N}(0,1)$ represents an observed signal, and the network is simply required to reproduce the observed signal z at a time $T \sim p(T)$. Crucially, on each trial, we do not explicitly constrain the readout $y(t)$ at any time other than the readout time T. We are interested in exploring the solution space of weights $\boldsymbol{\Theta} = [\boldsymbol{W}, \boldsymbol{b}, \boldsymbol{c}]$ that allow the network to solve the task, up to an ε-slack in the expected squared-error task performance:

$$\mathbb{E}_{z,T}\left[(y(T) - z)^2\right] \leq \varepsilon. \tag{12}$$

As in the main text, we will fix a choice of readout vector $\boldsymbol{c}$, with $\|\boldsymbol{c}\| = 1$, and focus on choices of $\boldsymbol{W}$ and $\boldsymbol{b}$. We would like to characterize the constrained optimization problem

$$\min_{\boldsymbol{W},\boldsymbol{b}} \quad \mathbb{E}_z\left[\int_0^\infty \|\boldsymbol{x}(t)\|^2 dt\right] + \lambda\left(\|\boldsymbol{W}\|_F^2 + \|\boldsymbol{b}\|^2\right) \tag{13}$$

$$\text{s.t. } \mathbb{E}_{z,T}\left[(y(T) - z)^2\right] \leq \varepsilon. \tag{14}$$

where λ is a parameter governing the relative importance of the activity and weight costs.

Performing the expectation over z, the activity penalty becomes $\boldsymbol{b}^\top \boldsymbol{R}\boldsymbol{b}$, where we define $\boldsymbol{R}(\boldsymbol{W}) = \int_0^\infty e^{\boldsymbol{A}^\top t} e^{\boldsymbol{A}t} dt$ and $\boldsymbol{A} = -\boldsymbol{I} + \boldsymbol{W}$. Similarly, performing the expectation in the constraint yields

$$\boldsymbol{b}^\top \boldsymbol{Q}\boldsymbol{b} - 2\boldsymbol{q}^\top \boldsymbol{b} + 1 \leq \varepsilon \tag{15}$$

where we have defined $\boldsymbol{Q}(\boldsymbol{W}) = \mathbb{E}_T[e^{\boldsymbol{A}^\top T}\boldsymbol{c}\boldsymbol{c}^\top e^{\boldsymbol{A}T}]$ and $\boldsymbol{q}(\boldsymbol{W}) = \mathbb{E}_T[e^{\boldsymbol{A}^\top T}\boldsymbol{c}]$.

Enforcing the constraint via Lagrange multiplier $\mu \geq 0$, we have:

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{b}, \mu) = \boldsymbol{b}^\top \boldsymbol{R}\boldsymbol{b} + \lambda\left(\|\boldsymbol{W}\|_F^2 + \|\boldsymbol{b}\|^2\right) + \mu(\boldsymbol{b}^\top \boldsymbol{Q}\boldsymbol{b} - 2\boldsymbol{q}^\top \boldsymbol{b} + 1 - \varepsilon). \tag{16}$$

This objective is quadratic in $\boldsymbol{b}$:

$$\mathcal{L}(\boldsymbol{W}, \boldsymbol{b}, \mu) = \boldsymbol{b}^\top(\boldsymbol{R} + \mu\boldsymbol{Q} + \lambda\boldsymbol{I})\boldsymbol{b} - 2\mu\boldsymbol{q}^\top \boldsymbol{b} + \mu(1-\varepsilon) + \lambda\|\boldsymbol{W}\|_F^2 \tag{17}$$

Minimizing in $\boldsymbol{b}$, we find

$$\boldsymbol{b}^* = \mu(\boldsymbol{R} + \mu\boldsymbol{Q} + \lambda\boldsymbol{I})^{-1}\boldsymbol{q}, \tag{18}$$

yielding the reduced objective

$$\mathcal{L}(\boldsymbol{W}) = \lambda\|\boldsymbol{W}\|_F^2 + \sup_{\mu\geq 0}\left[\mu(1-\varepsilon) - \mu^2 \boldsymbol{q}^\top(\boldsymbol{R} + \mu\boldsymbol{Q} + \lambda\boldsymbol{I})^{-1}\boldsymbol{q}\right]. \tag{19}$$

While simple to write down, it is difficult to make further analytical progress because the optimization over μ yields only an implicit fixed point equation for μ^*.

A.2 Fixed readout time

We now specialize to the fixed readout-time setting of the main text, $p(T) = \delta(T - \tau)$, for which $\boldsymbol{Q} = \boldsymbol{q}\boldsymbol{q}^\top$. By Sherman-Morrison, we have

$$\boldsymbol{q}^\top(\boldsymbol{R} + \mu\boldsymbol{Q} + \lambda\boldsymbol{I})^{-1}\boldsymbol{q} = \boldsymbol{q}^\top(\boldsymbol{R} + \lambda\boldsymbol{I})^{-1}\boldsymbol{q} - \mu\frac{(\boldsymbol{q}^\top(\boldsymbol{R} + \lambda\boldsymbol{I})^{-1}\boldsymbol{q})^2}{1 + \mu\boldsymbol{q}^\top(\boldsymbol{R} + \lambda\boldsymbol{I})^{-1}\boldsymbol{q}} \tag{20}$$

$$= \frac{\boldsymbol{q}^\top(\boldsymbol{R} + \lambda\boldsymbol{I})^{-1}\boldsymbol{q}}{1 + \mu\boldsymbol{q}^\top(\boldsymbol{R} + \lambda\boldsymbol{I})^{-1}\boldsymbol{q}}. \tag{21}$$

Thus, defining $\alpha(\boldsymbol{W}) = \boldsymbol{q}^\top(\boldsymbol{R} + \lambda \boldsymbol{I})^{-1}\boldsymbol{q}$, the optimization over μ becomes straightforward:

$$\sup_{\mu \geq 0}\left[\mu(1-\varepsilon) - \frac{\mu^2\alpha}{1+\mu\alpha}\right] = \sup_{\mu \geq 0}\left[\frac{\mu}{1+\mu\alpha} - \varepsilon\mu\right] \tag{22}$$

Differentiating with respect to μ gives $\mu^* = \frac{\varepsilon^{-1/2}-1}{\alpha}$, thus yielding

$$\sup_{\mu \geq 0}\left[\frac{\mu}{1+\mu\alpha} - \varepsilon\mu\right] = \frac{(1-\sqrt{\varepsilon})^2}{\alpha}. \tag{23}$$

Thus, the reduced objective is

$$\mathcal{L}(\boldsymbol{W}) = \lambda\|\boldsymbol{W}\|_F^2 + \frac{(1-\sqrt{\varepsilon})^2}{\boldsymbol{q}^\top(\boldsymbol{R}+\lambda\boldsymbol{I})^{-1}\boldsymbol{q}}. \tag{24}$$

Note that this is similar to the objective in the main text, but with the latter term rescaled by a factor $(1-\sqrt{\varepsilon})^2$. Therefore, at least for the fixed readout-time constraint, permitting some slack $\varepsilon > 0$ does not appreciably change the space of solutions.

Similarly, we can simplify the optimal input vector as

$$\boldsymbol{b}^* = \frac{(1-\sqrt{\varepsilon})(\boldsymbol{R}+\lambda\boldsymbol{I})^{-1}\boldsymbol{q}}{\boldsymbol{q}^\top(\boldsymbol{R}+\lambda\boldsymbol{I})^{-1}\boldsymbol{q}}. \tag{25}$$

To recover the setup of the main text, we set $\varepsilon = 0$:

$$\mathcal{L}(\boldsymbol{W}) = \lambda\|\boldsymbol{W}\|_F^2 + \frac{1}{\boldsymbol{q}^\top(\boldsymbol{R}+\lambda\boldsymbol{I})^{-1}\boldsymbol{q}}. \tag{26}$$

B Weight and activity efficiency

B.1 Unstructured recurrent weights

Suppose we select and fix recurrent weights randomly as $\boldsymbol{W}_{ij} \sim \mathcal{N}\left(0, \frac{\sigma_W^2}{N}\right)$, fix readout weights $\boldsymbol{c}$ such that $\|\boldsymbol{c}\| = 1$, and choose input weights $\boldsymbol{b}$ as

$$\boldsymbol{b}^* = \text{argmin}_{\boldsymbol{b}}\|\boldsymbol{b}\| \tag{27}$$

$$\text{s.t.} \quad g(\tau) = 1. \tag{28}$$

This corresponds to a reservoir-like regime. The input weights are then given as $\boldsymbol{b}^* = \frac{\boldsymbol{q}(\tau)}{\|\boldsymbol{q}(\tau)\|^2}$, where $\boldsymbol{q}(\tau) = e^{\boldsymbol{A}^\top\tau}\boldsymbol{c}$. We would then like to estimate the weight and activity costs E_W and E_x under this choice of parameters. Conveniently, all quantities of interest can be estimated from order parameters of the dynamical mean field theory description of random RNNs, which yields closed-form expressions here as we are considering linear RNNs. We give a brief self-contained version of the calculation here, which follows that of Bordelon et al. (2026).

Observe that $\boldsymbol{q}(t)$ obeys the dynamics

$$\frac{d\boldsymbol{q}}{dt} = -\boldsymbol{q} + \boldsymbol{W}^\top\boldsymbol{q}, \tag{29}$$

subject to initial condition $\boldsymbol{q}(0) = \boldsymbol{c}$. Squaring these dynamics yields

$$(1+\partial_t)(1+\partial_{t'})\boldsymbol{q}(t)\boldsymbol{q}(t')^\top = \boldsymbol{W}^\top\boldsymbol{q}(t)\boldsymbol{q}(t')^\top\boldsymbol{W}. \tag{30}$$

Defining the autocorrelation function $C(t,t') = \frac{1}{N}\langle\boldsymbol{q}(t)^\top\boldsymbol{q}(t')\rangle$, we have that

$$(1+\partial_t)(1+\partial_{t'})C(t,t') = \frac{1}{N}\,\text{Tr}(\langle\boldsymbol{W}^\top\boldsymbol{q}(t)\boldsymbol{q}(t')^\top\boldsymbol{W}\rangle) \tag{31}$$

$$= \frac{1}{N}\langle\boldsymbol{q}(t')^\top\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{q}(t)\rangle \tag{32}$$

$$\approx \frac{1}{N}\langle\boldsymbol{q}(t')^\top\langle\boldsymbol{W}\boldsymbol{W}^\top\rangle\boldsymbol{q}(t)\rangle \tag{33}$$

$$= \sigma_W^2 C(t,t'). \tag{34}$$

This PDE has closed-form solution

$$C(t,t') = C(0,0)e^{-t-t'} I_0\left(2\sqrt{\sigma_W^2 tt'}\right), \tag{35}$$

where I_0 is a modified Bessel function. Thus,

$$\|\boldsymbol{b}^*\|^2 = \frac{1}{NC(\tau,\tau)} \approx 2\sqrt{\pi\sigma_W\tau}e^{2(1-\sigma_W)\tau}, \tag{36}$$

where we have used the leading asymptotics $I_0(x) \sim \frac{e^x}{\sqrt{2\pi x}}$, and the initial condition $C(0,0) = \frac{1}{N}\boldsymbol{c}^\top\boldsymbol{c} = \frac{1}{N}$. The same mean-field description of the statistics of the corresponding activity variable $\widetilde{\boldsymbol{x}}(t) = e^{\boldsymbol{A}t}\boldsymbol{b}^*$ is difficult to apply directly because the initial condition $\boldsymbol{b}^*$ is strongly dependent on the particular choice of $\boldsymbol{W}$, but we nonetheless can obtain a bound corresponding to the activity resulting from a generic initial condition of the same norm (valid under mean-field assumptions):

$$E_x = \int_0^\infty \|\widetilde{\boldsymbol{x}}(t)\|^2 dt \gtrsim N\int_0^\infty C(t,t)|_{C(0,0)=\frac{1}{N}\|\boldsymbol{b}^*\|^2} dt \tag{37}$$

$$= \frac{\|\boldsymbol{b}^*\|^2}{2\sqrt{1-\sigma_W^2}}. \tag{38}$$

Thus, both the total weight cost $E_W = N\sigma_W^2 + \|\boldsymbol{b}^*\|^2$ and the activity cost E_x scale at least as large as $\sim e^{2(1-\sigma_W)\tau}$.

One can also choose input weights $\boldsymbol{b}$ to be minimum norm under the activity metric, i.e.

$$\boldsymbol{b}^* = \text{argmin}_{\boldsymbol{b}}\|\boldsymbol{b}\|_{\text{act}} \tag{39}$$

$$\text{s.t.} \quad g(\tau) = 1, \tag{40}$$

where $\|\boldsymbol{b}\|_{\text{act}}^2 = \boldsymbol{b}^\top\boldsymbol{R}\boldsymbol{b}$, corresponding to a minimizer $\boldsymbol{b}^* = \frac{\boldsymbol{R}^{-1}\boldsymbol{q}(\tau)}{\boldsymbol{q}(\tau)^\top\boldsymbol{R}^{-1}\boldsymbol{q}(\tau)}$. Numerically, this yields solutions with moderately smaller activity cost and moderately larger weight cost (Fig. 1b).

B.2 Gradient descent from small initialization

Here, we show that gradient descent from small initialization recovers the solution of minimum norm, provided that we clamp the recurrent weights $\boldsymbol{W}$ to be stable. Initialize parameters as $\boldsymbol{W}_{ij}^0, \boldsymbol{b}_i^0, \boldsymbol{c}_i^0 \sim \mathcal{N}\left(0, \frac{\sigma_W^2}{N}\right)$. The loss can be expressed as

$$L = \mathbb{E}_z[(y(\tau) - z)^2] = \sigma_z^2(g(\tau) - 1)^2. \tag{41}$$

For convenience, we set $\sigma_z^2 = 1$.

We then have the gradient flow dynamics

$$\frac{d}{ds}\boldsymbol{W}^s = -\eta(g(\tau) - 1)\int_0^\tau \boldsymbol{q}(t)\widetilde{\boldsymbol{x}}(\tau - t)^\top dt \tag{42}$$

$$\frac{d}{ds}\boldsymbol{b}^s = -\eta(g(\tau) - 1)\boldsymbol{q}(\tau) \tag{43}$$

$$\frac{d}{ds}\boldsymbol{c}^s = -\eta(g(\tau) - 1)\widetilde{\boldsymbol{x}}(\tau), \tag{44}$$

where we have absorbed the factor of 2 into the learning rate η. At early training times s, assuming $\sigma_W \ll 1$, we have, up to first-order in σ_W, that

$$\frac{d}{ds}\boldsymbol{W}^s \approx \eta\tau e^{-\tau}\boldsymbol{c}^s(\boldsymbol{b}^s)^\top \tag{45}$$

$$\frac{d}{ds}\boldsymbol{b}^s = \eta e^{-\tau}\boldsymbol{c}^s \tag{46}$$

$$\frac{d}{ds}\boldsymbol{c}^s = \eta e^{-\tau}\boldsymbol{b}^s \tag{47}$$

where we used the early-training-time approximations $\boldsymbol{q}(t) \approx e^{-t}\boldsymbol{c}^s + \mathcal{O}(\sigma_W^2)$ and $\widetilde{\boldsymbol{x}}(t) \approx e^{-t}\boldsymbol{b}^s + \mathcal{O}(\sigma_W^2)$. As in (Bordelon et al., 2026), $\boldsymbol{b}^s$, $\boldsymbol{c}^s$, and $\boldsymbol{W}^s$ quickly align to the common direction $\boldsymbol{r} \propto \frac{1}{2}(\boldsymbol{b}^0 + \boldsymbol{c}^0)$, yielding the reduced-order description $\boldsymbol{W}^s \approx w(s)\boldsymbol{r}\boldsymbol{r}^\top$, $\boldsymbol{b}^s \approx b(s)\boldsymbol{r}$, $\boldsymbol{c}^s \approx c(s)\boldsymbol{r}$

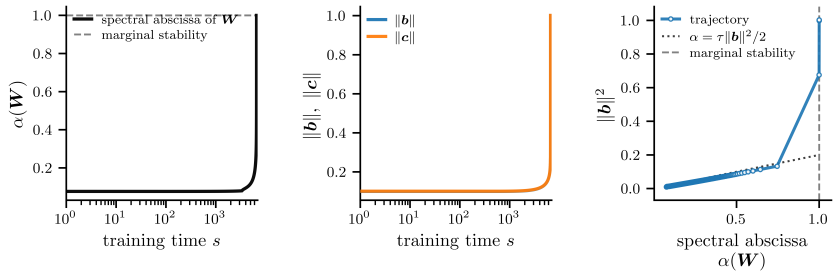


Figure B.1: **Learning the stimulus recall task via gradient descent on the MSE loss from small initialization.** Left: Spectral abscissa of the recurrent weights over training time. Middle: Norms of the input and output weights over training time. Right: Learning trajectory in the spectral abscissa versus input weight norm plane closely tracks the conservation relation in Eq. 55.

for $\|\boldsymbol{r}\| = 1$. Plugging this ansatz into the full gradient flow then yields the reduced flow

$$\frac{d}{ds}w(s) = -\eta\left(b(s)c(s)e^{(w(s)-1)\tau} - 1\right)b(s)c(s)e^{(w(s)-1)\tau}\tau \tag{48}$$

$$\frac{d}{ds}b(s) = -\eta\left(b(s)c(s)e^{(w(s)-1)\tau} - 1\right)e^{(w(s)-1)\tau}c(s) \tag{49}$$

$$\frac{d}{ds}c(s) = -\eta\left(b(s)c(s)e^{(w(s)-1)\tau} - 1\right)e^{(w(s)-1)\tau}b(s) \tag{50}$$

Observe that this flow obeys the conservation law $\frac{d}{ds}(b(s)^2 - c(s)^2) = 0$. Thus, assuming a balanced initialization $b(0) = c(0)$, we can take $c(s) = b(s)$. This yields

$$\frac{d}{ds}w(s) = -\eta\left(b(s)^2e^{(w(s)-1)\tau} - 1\right)b(s)^2e^{(w(s)-1)\tau}\tau \tag{51}$$

$$\frac{d}{ds}b(s) = -\eta\left(b(s)^2e^{(w(s)-1)\tau} - 1\right)e^{(w(s)-1)\tau}b(s). \tag{52}$$

We then have that

$$\frac{d}{ds}b(s)^2 = 2b(s)b'(s) = -2\eta\left(b(s)^2e^{(w(s)-1)\tau} - 1\right)e^{(w(s)-1)\tau}b(s)^2. \tag{53}$$

Thus,

$$\frac{dw(s)}{d(b(s)^2)} = \frac{\tau}{2}, \tag{54}$$

implying that $w(s) - \frac{\tau}{2}b(s)^2$ is conserved. Since we assume small initialization, this yields the relation

$$w(s) = \frac{\tau}{2}b(s)^2, \tag{55}$$

which holds over all training times s. The endpoint of training is given by the condition for solving the task $g(\tau) = 1$, giving

$$b^2e^{(w-1)\tau} = \frac{2}{\tau}we^{(w-1)\tau} = 1. \tag{56}$$

However, for $\tau > 2$, the corresponding w satisfying Eq. 56 requires $w > 1$, yielding unstable recurrent weights. Thus, if we replace w with a clamped variable $w_c(s) = \min(w(s), 1)$, then $w_c(s)$ obeys the relation $w_c(s) = \frac{\tau}{2}b(s)^2$ only up to the first time s_c such that $w_c(s_c) = 1$. Once w_c is clamped, the input/output norm obeys the dynamics

$$\frac{d}{ds}b(s)^2 = -2\eta(b(s)^2 - 1)b(s)^2 \tag{57}$$

from initial condition $b(s_c)^2 = \frac{2}{\tau}$. Since this is now a logistic ODE, we have that $\lim_{s\to\infty} b(s)^2 = 1$.

Thus, with stability clamping of the recurrent weights, GD from small initialization yields the minimum norm solution: the recurrent weights, input, and output weights align along the same single direction, with marginally stable recurrent dynamics. In the absence of stability clamping, the weights still align in this manner, but the recurrent weights become unstable.

B.3 WEIGHT-CONSTRAINED REGIME

Taking λ to be large, we find:

$$\mathcal{L}(\boldsymbol{W}) = \lambda \left(\|\boldsymbol{W}\|_F^2 + \frac{1}{\|\boldsymbol{q}\|^2} + \frac{1}{\lambda} \frac{\boldsymbol{q}^\top \boldsymbol{R} \boldsymbol{q}}{\|\boldsymbol{q}\|^4} + O\left(\frac{1}{\lambda^2}\right) \right). \tag{58}$$

We can bound $\|\boldsymbol{q}\|$ as:

$$\|\boldsymbol{q}\| = e^{-\tau} \|e^{\boldsymbol{W}^\top \tau} \boldsymbol{c}\| \leq e^{-\tau} \|e^{\boldsymbol{W}^\top \tau}\|_2 \|\boldsymbol{c}\| \tag{59}$$

$$\leq e^{\tau(\|\boldsymbol{W}\|_2 - 1)} \tag{60}$$

$$\leq e^{\tau(\|\boldsymbol{W}\|_F - 1)}. \tag{61}$$

Thus, we have a lower bound

$$\mathcal{L}/\lambda \geq \alpha^2 + e^{2\tau(1-\alpha)} + O(1/\lambda), \tag{62}$$

where $\alpha = \|\boldsymbol{W}\|_F$. Taking $W = \alpha \boldsymbol{c}\boldsymbol{c}^\top$ saturates this lower bound. The optimal α is then given by the minimizer of $f(\alpha) = \alpha^2 + e^{2\tau(1-\alpha)}$. Observing that $f'(0) < 0$ and $f'(1) = 2(1-\tau)$, for $0 < \tau < 1$, the minimizer is the solution to $f'(\alpha^*) = 2(\alpha^* - \tau e^{-2(\alpha^*-1)\tau}) = 0$. Thus, $W \propto \boldsymbol{c}\boldsymbol{c}^\top$ is globally optimal for $\tau \leq 1$. However, for $\tau > 1$, the minimizer requires $\alpha^* > 1$, which leads to divergent activity, and thus is disallowed even at infinitesimal effective activity cost strength λ^{-1}.

We will now prove that even for $\tau > 1$, provided we restrict $\boldsymbol{W}$ to be stable, a global optimum (in the infimum sense) is given by $W \propto \boldsymbol{c}\boldsymbol{c}^\top$, and in particular by $\boldsymbol{W}_0 = \alpha \boldsymbol{c}\boldsymbol{c}^\top$ for $\alpha \to 1^-$.

Defining

$$\tilde{\mathcal{L}} = \|\boldsymbol{W}\|_F^2 + \frac{1}{\|\boldsymbol{q}\|^2}, \tag{63}$$

we have that $\tilde{\mathcal{L}}(\boldsymbol{W}_0) = 2$.

It thus suffices to show that for all stable $\boldsymbol{W}$ (i.e., $\max(\Re(\lambda(\boldsymbol{W})) < 1$), $\tilde{\mathcal{L}}(\boldsymbol{W}) \geq 2$.

For $\|\boldsymbol{W}\|_F \geq \sqrt{2}$, the claim immediately follows from the nonnegativity of $1/\|\boldsymbol{q}\|^2$.

For $\|\boldsymbol{W}\|_F < 1$, since the lower bound of $\tilde{\mathcal{L}}$, $\alpha^2 + e^{2\tau(1-\alpha)}$, is monotonically decreasing in α over $0 < \alpha < 1$ for $\tau > 1$, the optimal loss achievable for $0 < \alpha < 1$ is given by $\alpha \to 1^-$, yielding the desired result.

Thus it remains to show that the claim still holds for $1 < \|\boldsymbol{W}\|_F < \sqrt{2}$ and $\tau > 1$. The claim in this case is nontrivial because nonnormal amplification can in general make $\|\boldsymbol{q}\|$ grow beyond 1, and standard exponential growth bounds on $\|\boldsymbol{q}(t)\|$ are too loose. The proof of this subcase is not terribly illuminating, but we give it here for completeness.

From $\|\boldsymbol{q}\| = \|e^{\boldsymbol{A}^\top \tau} \boldsymbol{c}\| \leq \|e^{\boldsymbol{A}\tau}\|_2$, it suffices to prove the stronger statement that

$$\|e^{\boldsymbol{A}\tau}\|_2^2 \leq \frac{1}{2 - \|\boldsymbol{W}\|_F^2}. \tag{64}$$

We first prove this statement in the special case of real upper-triangular 2x2 matrices of the form $\boldsymbol{W} = \begin{pmatrix} a & b \\ 0 & d \end{pmatrix}$, for $a, d < 1$ (stability criterion), and $1 \leq \|\boldsymbol{W}\|_F^2 = a^2 + b^2 + d^2 < 2$. We will later show that all other cases can be reduced to this case by induction.

We have that

$$e^{(-\boldsymbol{I} + \boldsymbol{W})\tau} = \begin{pmatrix} e^{(a-1)\tau} & bC \\ 0 & e^{(d-1)\tau} \end{pmatrix}, \tag{65}$$

where $C = \int_0^\tau e^{(a-1)(\tau-t)} e^{(d-1)t} dt$. Then, to show that $\|e^{(-\boldsymbol{I}+\boldsymbol{W})\tau}\|_2^2 \leq \frac{1}{2-\|\boldsymbol{W}\|_F^2}$, we would like to prove

$$\frac{1}{2 - \|\boldsymbol{W}\|_F^2} \boldsymbol{I} - \left(e^{(-\boldsymbol{I}+\boldsymbol{W})\tau}\right)^\top e^{(-\boldsymbol{I}+\boldsymbol{W})\tau} \succeq \boldsymbol{0}. \tag{66}$$

Since this is 2x2, this reduces to the conditions $e^{2(a-1)\tau} \leq \frac{1}{\Delta}$ (top left) and the determinant condition

$$\left(1 - \Delta e^{2(a-1)\tau}\right)\left(1 - \Delta e^{2(d-1)\tau}\right) \geq \Delta b^2 C^2, \tag{67}$$

where we introduced the shorthand $\Delta = 2 - \|\boldsymbol{W}\|_F^2$. The top left condition follows trivially from observing that $0 < \Delta \leq 1$, and thus $\frac{1}{\Delta} \geq 1$ and $e^{2(a-1)\tau} \leq 1$. For the determinant condition, we can rewrite

$$(1 - \Delta e^{2(a-1)\tau})(1 - \Delta e^{2(d-1)\tau}) = \left(1 - \Delta e^{(a+d-2)\tau}\right)^2 - \Delta\left(e^{(a-1)\tau} - e^{(d-1)\tau}\right)^2. \tag{68}$$

We also have the explicit integral $C = \frac{1}{a-d}\left(e^{(a-1)\tau} - e^{(d-1)\tau}\right)$. Thus, to show 67, we would like to show that

$$\left(1 - \Delta e^{(a+d-2)\tau}\right)^2 \geq \Delta((a-d)^2 + b^2)C^2. \tag{69}$$

To prove this, it will be convenient to first show that

$$\Delta((a-d)^2 + b^2) \leq (a+d-2)^2. \tag{70}$$

From $\Delta = 2 - a^2 - b^2 - d^2$, we have that

$$\Delta((a-d)^2 + b^2) = \Delta(2 - 2ad - \Delta), \tag{71}$$

thus giving

$$\begin{aligned}(a+d-2)^2 - \Delta((a-d)^2 + b^2) &= ((a-1) + (d-1))^2 - \Delta(2 - 2ad - \Delta) && (72)\\ &= ((a-1) + (d-1) + \Delta)^2 + 2\Delta(1-a)(1-d) && (73)\\ &\geq 0. && (74)\end{aligned}$$

From 69, it then suffices to prove

$$1 - \Delta e^{(a+d-2)\tau} \geq (2 - a - d)C. \tag{75}$$

We have the pointwise bound

$$((1-a) + (1-d))e^{(a-1)(t-s)}e^{(d-1)s} \leq (1-a)e^{(a+d-2)(t-s)} + (1-d)e^{(a+d-2)s}, \tag{76}$$

which integrating from 0 to τ, yields

$$(2 - a - d)C \leq 1 - e^{(a+d-2)\tau}. \tag{77}$$

Since $0 \leq \Delta < 1$, we have that $1 - e^{(a+d-2)\tau} \leq 1 - \Delta e^{(a+d-2)\tau}$, thus showing 75, completing the proof for the upper triangular 2x2 case.

For generic $\boldsymbol{W}$, we have the complex Schur decomposition $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{T}\boldsymbol{U}^*$, for $\boldsymbol{U}, \boldsymbol{T} \in \mathbb{C}^{N\times N}$, $\boldsymbol{U}^*\boldsymbol{U} = \boldsymbol{U}\boldsymbol{U}^* = \boldsymbol{I}$, and $\boldsymbol{T}$ upper triangular. Define $\boldsymbol{T}^R$ such that $T^R_{ii} = \Re T_{ii}$, $T^R_{ij} = |T_{ij}|$ for $i < j$. Observe that $\|e^{\boldsymbol{A}\tau}\|_2 = \|e^{(-\boldsymbol{I}+\boldsymbol{T})\tau}\|_2$. We then would like to show that it is sufficient to consider only the corresponding real upper triangular matrix $\boldsymbol{T}^R$ by bounding

$$\|e^{(-\boldsymbol{I}+\boldsymbol{T})\tau}\|_2 \leq \|e^{(-\boldsymbol{I}+\boldsymbol{T}^R)\tau}\|_2. \tag{78}$$

To show this, we will show a stronger pointwise statement that $|e^{(-\boldsymbol{I}+\boldsymbol{T})\tau}\boldsymbol{v}| \leq e^{(-\boldsymbol{I}+\boldsymbol{T}^R)\tau}|\boldsymbol{v}|$, for any $\boldsymbol{v} \in \mathbb{C}$, where $|\cdot|$ operates entrywise. Defining $\boldsymbol{x}(t) = e^{(-\boldsymbol{I}+\boldsymbol{T})t}\boldsymbol{v}$ and $\boldsymbol{y}(t) = e^{(-\boldsymbol{I}+\boldsymbol{T}^R)t}|\boldsymbol{v}|$, we have

$$x_i(t) = e^{(-1+T_{ii})t}v_i + \int_0^t e^{(-1+T_{ii})(t-s)}\sum_{i<j} T_{ij}x_j(s)ds. \tag{79}$$

We will now use backwards induction starting from $i = N$ down to $i = 1$ to show this pointwise bound. For the base case $i = N$, $x_N(t) = e^{(-1+T_{NN})t}v_N$, so

$$|x_N(t)| = e^{(-1+\Re T_{NN})t}|v_N| = y_N(t). \tag{80}$$

For general i, we have

$$\begin{aligned}|x_i(t)| &\leq e^{(-1+\Re T_{ii})t}|v_i| + \int_0^t e^{(-1+\Re T_{ii})(t-s)}\sum_{i<j}|T_{ij}||x_j(s)|ds && (81)\\ &\leq^{(-1+\Re T_{ii})t} |v_i| + \int_0^t e^{(-1+\Re T_{ii})(t-s)}\sum_{i<j}|T_{ij}|y_j(s)ds && (82)\\ &= y_i(t), && (83)\end{aligned}$$

where the latter inequality follows from the induction hypothesis. Thus, we have shown that $|e^{(-\boldsymbol{I}+\boldsymbol{T})\tau}\boldsymbol{v}| \leq e^{(-\boldsymbol{I}+\boldsymbol{T}^R)\tau}|\boldsymbol{v}|$. Therefore,

$$\|e^{(-\boldsymbol{I}+\boldsymbol{T})\tau}\boldsymbol{v}\|_2 \leq \|e^{(-\boldsymbol{I}+\boldsymbol{T}^R)\tau}|\boldsymbol{v}|\|_2 \tag{84}$$

$$\leq \|e^{(-\boldsymbol{I}+\boldsymbol{T}^R)\tau}\|_2\|\boldsymbol{v}\|_2. \tag{85}$$

Taking the supremum over $\boldsymbol{v} \neq 0$ yields the desired bound 78. Combining this with the observation $\|\boldsymbol{T}^R\|_F \leq \|\boldsymbol{T}\|_F = \|\boldsymbol{W}\|_F$, we have that for any $\boldsymbol{W}$ obeying $1 < \|\boldsymbol{W}\|_F < \sqrt{2}$ and $\tau > 1$, it suffices to show the bound for the corresponding real upper triangular matrix $\boldsymbol{T}^R$,

$$\left\|e^{(-I+\boldsymbol{T}^R)\tau}\right\|_2^2 \leq \frac{1}{2 - \|\boldsymbol{T}^R\|_F^2}. \tag{86}$$

We will now prove 86 by induction on dimension N. For the base case $N = 1$, let $r = T_{11}^R$. Recall that $r < 1$ (stability) and $r^2 < 2$ (Frobenius norm bound). Since $\tau > 1$,

$$e^{2(-1+r)\tau} \leq e^{2(-1+r)} \leq \frac{1}{2 - r^2}, \tag{87}$$

where the second inequality follows from $2 - r^2 \leq 1 + 2(1 - r) \leq e^{2(1-r)}$.

For $N = k$ for some $k > 1$, we can write

$$\boldsymbol{T}^R = \begin{pmatrix} \boldsymbol{B} & \boldsymbol{v} \\ \boldsymbol{0}^\top & d \end{pmatrix}, \tag{88}$$

for $\boldsymbol{B} \in \mathbb{R}^{(N-1)\times(N-1)}$ an upper triangular matrix, $\boldsymbol{v} \in \mathbb{R}^{N-1}$, and $d \in \mathbb{R}$. Let $B = \|\boldsymbol{B}\|_F$ and $v = \|\boldsymbol{v}\|_F$. Defining $\boldsymbol{h}(\tau) = \int_0^\tau e^{(-\boldsymbol{I}_{N-1}+\boldsymbol{B})(\tau-t)}\boldsymbol{v}e^{(d-1)t}dt$, we have that

$$\left\|e^{(-I+\boldsymbol{T}^R)\tau}\right\|_2 = \left\|\begin{pmatrix} e^{(-\boldsymbol{I}_{N-1}+\boldsymbol{B})\tau} & \boldsymbol{h}(\tau) \\ \boldsymbol{0}^\top & e^{(d-1)\tau} \end{pmatrix}\right\| \tag{89}$$

$$\leq \left\|\begin{pmatrix} \left\|e^{(-\boldsymbol{I}_{N-1}+\boldsymbol{B})\tau}\right\|_2 & \|\boldsymbol{h}(\tau)\|_2 \\ 0 & e^{(d-1)\tau} \end{pmatrix}\right\|. \tag{90}$$

We now consider two cases. If $B < 1$, $\|e^{(-\boldsymbol{I}_{N-1}+\boldsymbol{B})\tau}\|_2 \leq e^{(B-1)\tau}$. Thus, $\|\boldsymbol{h}(\tau)\| \leq v\int_0^\tau e^{(B-1)(\tau-t)}e^{(d-1)t}dt$, thus yielding the bound

$$\left\|e^{(-I+\boldsymbol{T}^R)\tau}\right\|_2 \leq \left\|\begin{pmatrix} e^{(B-1)\tau} & v\int_0^\tau e^{(B-1)(\tau-t)}e^{(d-1)t}dt \\ 0 & e^{(d-1)\tau} \end{pmatrix}\right\| \tag{91}$$

$$= \left\|\exp\left(\left(-\boldsymbol{I}_2 + \begin{pmatrix} B & v \\ 0 & d \end{pmatrix}\right)\tau\right)\right\| \tag{92}$$

$$\leq \frac{1}{\sqrt{2 - \|\boldsymbol{T}^R\|_F^2}}, \tag{93}$$

where the last inequality follows from the 2x2 upper triangular bound. In the case $B > 1$, we have (from 90)

$$\left\|e^{(-I+\boldsymbol{T}^R)\tau}\right\|_2 \leq \left\|\begin{pmatrix} \frac{1}{\sqrt{2-B^2}} & \frac{v}{\sqrt{2-B^2}}\int_0^\tau e^{(d-1)t}dt \\ 0 & e^{(d-1)\tau} \end{pmatrix}\right\|, \tag{94}$$

where we have applied the induction hypothesis on $\boldsymbol{B}$, as well as the bound

$$\|\boldsymbol{h}(\tau)\| \leq \int_0^\tau \|e^{(-\boldsymbol{I}_{N-1}+\boldsymbol{B})(\tau-t)}\|_2\|\boldsymbol{v}\|e^{(d-1)t}dt \tag{95}$$

$$\leq \frac{v}{\sqrt{2-B^2}}\int_0^\tau e^{(d-1)t}dt, \tag{96}$$

where for $\tau - t \geq 1$, we use the induction hypothesis directly, whereas for $0 \leq \tau - t < 1$, we use $\|e^{(-\boldsymbol{I}+\boldsymbol{B})(\tau-t)}\| \leq e^{(B-1)(\tau-t)} \leq e^{B-1} \leq \frac{1}{\sqrt{2-B^2}}$. Observe that we can factor the matrix on the RHS of 94 as

$$\begin{pmatrix} \frac{1}{\sqrt{2-B^2}} & 0 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} 1 & v\int_0^\tau e^{(d-1)t}dt \\ 0 & e^{(d-1)\tau} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{2-B^2}} & 0 \\ 0 & 1 \end{pmatrix}\exp\left(\left(-\boldsymbol{I}_2 + \begin{pmatrix} 1 & v \\ 0 & d \end{pmatrix}\right)\tau\right). \tag{97}$$

Thus, for $B > 1$, we have the bound

$$\left\| e^{(-I+\boldsymbol{T}^R)\tau} \right\|_2^2 \leq \frac{1}{2-B^2} \cdot \frac{1}{2-(1+v^2+d^2)}, \tag{98}$$

which we would like to upper bound by $\frac{1}{2-\|\boldsymbol{T}^R\|_F^2} = \frac{1}{2-B^2-v^2-d^2}$.

Observe that

$$(2-B^2)(1-v^2-d^2) - (2-B^2-v^2-d^2) = (v^2+d^2)(B^2-1) \tag{99}$$

$$\geq 0, \tag{100}$$

and thus $(2-B^2)(1-v^2-d^2) \geq (2-B^2-v^2-d^2)$, proving the desired claim.

B.4 Activity-constrained regime

The limit $\lambda \to 0$ is a bit more difficult to analyze in full generality. Our approach will be to show that a particular structured choice of W (but not necessarily the only choice) can achieve arbitrarily small cost $\mathcal{L}$ when $\lambda = 0$, and then derive how the optimal W within this family of solutions adjusts to small but nonzero λ. For $\lambda = 0$, the objective reduces to

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{\boldsymbol{q}^\top \boldsymbol{R}^{-1} \boldsymbol{q}}. \tag{101}$$

We first establish a lower bound on this term, and then show by construction a network that can achieve the lower bound, up to a constant factor.

Observe that $\boldsymbol{q}^\top \boldsymbol{R}^{-1} \boldsymbol{q} = \boldsymbol{c}^\top e^{\boldsymbol{A}\tau} R^{-1} e^{\boldsymbol{A}^\top \tau} \boldsymbol{c} = \boldsymbol{c}^\top \boldsymbol{M}(\tau) \boldsymbol{c}$, where $\boldsymbol{M}(\tau) = e^{\boldsymbol{A}\tau} \boldsymbol{R}^{-1} e^{\boldsymbol{A}^\top \tau}$. Note that $\boldsymbol{M}(\tau) \succeq 0$. We can also show that $\boldsymbol{M}(\tau)$ is monotonically decreasing in τ in the Loewner order:

$$\frac{d}{d\tau} \boldsymbol{M} = e^{\boldsymbol{A}\tau} (\boldsymbol{A}\boldsymbol{R}^{-1} + \boldsymbol{R}^{-1}\boldsymbol{A}^\top) e^{\boldsymbol{A}^\top \tau}. \tag{102}$$

Since $\boldsymbol{R}$ satisfies the Lyapunov equation $\boldsymbol{A}^\top \boldsymbol{R} + \boldsymbol{R}\boldsymbol{A} = -\boldsymbol{I}$, left and right multiplying by $\boldsymbol{R}^{-1}$ yields the identity $\boldsymbol{A}\boldsymbol{R}^{-1} + \boldsymbol{R}^{-1}\boldsymbol{A}^\top = -\boldsymbol{R}^{-2}$. Thus, we have that

$$\frac{d}{d\tau} \boldsymbol{M} = -e^{\boldsymbol{A}\tau} \boldsymbol{R}^{-2} e^{\boldsymbol{A}^\top \tau} \preceq 0. \tag{103}$$

Moreover, we have the exact identity $\int_0^\infty \mathrm{Tr}(\boldsymbol{M}(t))dt = \int_0^\infty \mathrm{Tr}(\boldsymbol{R}^{-1} e^{\boldsymbol{A}^\top t} e^{\boldsymbol{A}t})dt = \mathrm{Tr}(\boldsymbol{I}) = N$.

Thus, we have the bounds

$$\tau \,\mathrm{Tr}(\boldsymbol{M}(\tau)) \leq \int_0^\tau \mathrm{Tr}(\boldsymbol{M}(\tau))dt \leq N, \tag{104}$$

yielding $\mathrm{Tr}(\boldsymbol{M}(\tau)) \leq \frac{N}{\tau}$. Finally, this yields

$$\boldsymbol{c}^\top \boldsymbol{M}(\tau) \boldsymbol{c} \leq \lambda_{max}(\boldsymbol{M}(\tau)) \leq \frac{N}{\tau}. \tag{105}$$

Thus,

$$\frac{1}{\boldsymbol{q}^\top \boldsymbol{R}^{-1} \boldsymbol{q}} \geq \frac{\tau}{N}. \tag{106}$$

The idealized solution in this limit is to produce a spike of activity along $\boldsymbol{c}$ precisely at the readout time τ. Since this kind of sudden transient amplification requires extremely nonnormal dynamics to execute, we consider a choice of recurrent weights of the homogeneous feedforward chain

$$\boldsymbol{A} = \boldsymbol{W} - \boldsymbol{I} = -a\boldsymbol{I} + k\boldsymbol{\Gamma}, \tag{107}$$

where $\boldsymbol{\Gamma}_{ij} = \delta_{i+1,j}$, and $a, k > 0$ represent the decay rate and nonnormal amplification strength, respectively.

It is convenient to write $\boldsymbol{A} = \boldsymbol{D}\boldsymbol{J}\boldsymbol{D}^{-1}$, where

$$\boldsymbol{D} = \mathrm{Diag}(1, \rho^{-1}, \rho^{-2}, \ldots, \rho^{-(N-1)}), \quad \boldsymbol{J} = a(-\boldsymbol{I} + \boldsymbol{\Gamma}). \tag{108}$$

Indeed, since $\boldsymbol{D}\boldsymbol{\Gamma}\boldsymbol{D}^{-1} = \rho\boldsymbol{\Gamma}$, we have

$$\boldsymbol{D}\boldsymbol{J}\boldsymbol{D}^{-1} = -a\boldsymbol{I} + a\rho\boldsymbol{\Gamma} = -a\boldsymbol{I} + k\boldsymbol{\Gamma} = \boldsymbol{A}.$$

Then,

$$\boldsymbol{q}^\top(\boldsymbol{R}+\lambda\boldsymbol{I})^{-1}\boldsymbol{q} = \tilde{\boldsymbol{q}}^\top(\tilde{\boldsymbol{R}}+\lambda\boldsymbol{D}^2)^{-1}\tilde{\boldsymbol{q}},$$

where

$$\tilde{\boldsymbol{q}} = e^{\boldsymbol{J}^\top\tau}\boldsymbol{e}_1, \quad \tilde{\boldsymbol{R}} = \int_0^\infty e^{\boldsymbol{J}^\top t} D^2 e^{\boldsymbol{J}t}\, dt.$$

Taking ρ to be large and λ to 0, we have

$$\begin{aligned}\tilde{\boldsymbol{R}} &= \int_0^\infty e^{\boldsymbol{J}^\top t}\left(\boldsymbol{e}_1\boldsymbol{e}_1^\top + \rho^{-2}\boldsymbol{e}_2\boldsymbol{e}_2^\top + \rho^{-4}\boldsymbol{e}_3\boldsymbol{e}_3^\top + \cdots\right) e^{\boldsymbol{J}t}\, dt \\ &= \int_0^\infty e^{-2at}\boldsymbol{f}(t)\boldsymbol{f}(t)^\top\, dt + O(\rho^{-2}),\end{aligned}$$

where

$$\boldsymbol{f}(t) = \left(1,\, at,\, \frac{(at)^2}{2!},\, \ldots,\, \frac{(at)^{N-1}}{(N-1)!}\right)^\top.$$

Similarly,

$$\tilde{\boldsymbol{q}} = e^{-a\tau}\boldsymbol{f}(\tau).$$

Observe that, to leading order in ρ^{-2}, $\tilde{\boldsymbol{R}}_{ij} = \langle f_i(t), f_j(t)\rangle_a$ is a Gram matrix of polynomials $(at)^i/i!$ up to degree $N-1$ under the weighted inner product

$$\langle f, g\rangle_a = \int_0^\infty e^{-2at} f(t)g(t)\, dt.$$

Thus, letting

$$\boldsymbol{\Phi}(t) = (\phi_0, \phi_1, \ldots, \phi_{N-1})^\top$$

be the orthogonal basis for this inner product space, we have that $\tilde{\boldsymbol{R}}^{-1}$

$$\tilde{\boldsymbol{q}}^\top\tilde{\boldsymbol{R}}^{-1}\tilde{\boldsymbol{q}} = e^{-2a\tau}\sum_{i=0}^{N-1}\phi_i(\tau)^2 + O(\rho^{-2}).$$

For the weight e^{-2at}, this basis is given by the Laguerre polynomials

$$\phi_n(t) = \sqrt{2a}\, L_n(2at).$$

Taking N to be proportional to $a\tau$ as

$$x = 2a\tau = Nz,$$

we have from asymptotics of Laguerre kernels that (Forrester et al., 2005)

$$e^{-x}\sum_{n=0}^{N-1} L_n(x)^2 = \frac{1}{2\pi}\sqrt{\frac{4-z}{z}} + O(1/N).$$

Thus,

$$\tilde{\boldsymbol{q}}^\top\tilde{\boldsymbol{R}}^{-1}\tilde{\boldsymbol{q}} \approx 2a\cdot\frac{1}{2\pi}\sqrt{\frac{4-z}{z}} = \frac{N}{2\pi\tau}\sqrt{z(4-z)}.$$

This is maximized at $z=2$, yielding

$$L(a^*)\big|_{\lambda=0,\rho\to\infty} = \frac{\pi\tau}{N},$$

with

$$a^* = \frac{N}{\tau}.$$

We now justify that taking $\rho\to\infty$ is indeed optimal for minimizing this activity cost term. We can rewrite the denominator as

$$\tilde{\boldsymbol{q}}^\top(\tilde{\boldsymbol{R}}+\lambda\boldsymbol{D}^2)^{-1}\tilde{\boldsymbol{q}} = \sup_{\boldsymbol{u}\in\mathbb{R}^N}\frac{(\boldsymbol{u}^\top\tilde{\boldsymbol{q}})^2}{\boldsymbol{u}^\top(\tilde{\boldsymbol{R}}+\lambda\boldsymbol{D}^2)\boldsymbol{u}}. \tag{109}$$

Observing that $\boldsymbol{u}^\top\tilde{\boldsymbol{q}} = e^{-a\tau}\sum_{i=0}^{N-1} u_i\frac{(a\tau)^i}{i!}$, we can thus define a polynomial $p_u(s)$ of degree $N-1$ such that $\boldsymbol{u}^\top\tilde{\boldsymbol{q}} = e^{-a\tau}p_u(a\tau)$. Using the expansion $\tilde{\boldsymbol{R}} = \sum_{i=0}^{N-1}\int_0^\infty e^{-2at}\rho^{-2i}\boldsymbol{f}^{(i)}(t)\boldsymbol{f}^{(i)}(t)^\top dt$,

where $\cdot^{(i)}$ denotes the ith derivative, we can then pass to an optimization over polynomials p of degree at most $N-1$:

$$\tilde{\boldsymbol{q}}^\top(\tilde{\boldsymbol{R}}+\lambda\boldsymbol{D}^2)^{-1}\tilde{\boldsymbol{q}} = \sup_{p:deg(p)\leq N-1} \frac{e^{-2a\tau}p(a\tau)^2}{\sum_{i=0}^{N-1}\int_0^\infty \rho^{-2i}p^{(i)}(at)^2dt + \lambda\sum_{i=0}^{N-1}\rho^{-2i}p^{(i)}(0)^2}. \quad (110)$$

Since the denominator is nonincreasing in ρ, it follows that the activity cost term $(\tilde{\boldsymbol{q}}^\top(\tilde{\boldsymbol{R}}+\lambda\boldsymbol{D}^2)^{-1}\tilde{\boldsymbol{q}})^{-1}$ is nonincreasing in ρ. Thus, taking $\rho\to\infty$ is optimal when $\lambda\to 0$.

The resulting objective cost improves towards zero as $1/N$; an infinitely long chain with infinitely large decay rate and amplification factor therefore achieves zero activity cost. This is pathological, so we are interested in understanding what the optimal amplifying chain is when λ is small but nonzero. In particular, there should be an optimal chain length $N^*<\infty$ which should decay with λ.

Heuristically, for small but nonzero λ, $\rho\to\infty$ is disallowed, but $k>a$ must still hold for transient growth in state norm to occur. Assuming the optimal decay rate retains the scaling $a\sim N/\tau$ found above, k^* must therefore scale at least as large as N/τ, yielding that the Frobenius penalty scales as $\|\boldsymbol{W}\|_F^2 = N((1-a)^2+k^2)\sim\frac{N^3}{\tau^2}$. This yields

$$\mathcal{L}\sim C_1\frac{\tau}{N}+\lambda C_2\frac{N^3}{\tau^2}, \quad (111)$$

for some $\mathcal{O}(1)$ constants C_1 and C_2. Thus, for small but nonzero λ, there is a finite optimal number of modes N that minimizes the loss, which scales like $N^*\sim\lambda^{-1/4}\tau^{3/4}$. We emphasize that the exact exponents derived here are not universal, and are likely specific to the particular ansatz for W analyzed here.

B.5 Alternative activity-efficient construction based on oscillatory dynamics

Another natural way of constructing a highly activity-efficient solution is to construct a linear dynamical system that implements a Dirichlet kernel approximation of a temporally narrow spike around the readout time. In particular, we seek to use

$$D_n(x) = \sum_{k=-n}^{n} e^{ikx} = 1+2\sum_{k=1}^{n}\cos(kx), \quad (112)$$

which approaches a periodic delta function as $D_\infty(x) = 2\pi\sum_{k=-\infty}^{\infty}\delta(x+2\pi k)$. Thus, choosing weights such that the filter satisfies $g(t) = \frac{1}{N}e^{-a(t-\tau)}D_n(\omega(t-\tau))$ with frequency $\omega=\frac{\pi}{\tau}$ yields a spike of magnitude 1 at $t=\tau$, with periodic (but exponentially damped) spikes reoccurring at $t=3\tau, 5\tau$, etc.

Let $N=2n+1$. If we choose recurrent weights $\boldsymbol{W}_0$ such that $\boldsymbol{A}_0=-\boldsymbol{I}+\boldsymbol{W}_0$ takes the form

$$\boldsymbol{A}_0 = \begin{pmatrix} -a & & & & & & & \\ & -a & -\omega & & & & & \\ & \omega & -a & & & & & \\ & & & -a & -2\omega & & & \\ & & & 2\omega & -a & & & \\ & & & & & \ddots & & \\ & & & & & & -a & -n\omega \\ & & & & & & n\omega & -a \end{pmatrix}, \quad (113)$$

where a is the decay rate, and ω is the base frequency.

Let $\boldsymbol{A}_0^k$ denote the k^{th} 2x2 diagonal block.

With readout weights $\boldsymbol{c} = \begin{bmatrix} c^{(0)} & (\boldsymbol{c}^{(1)})^\top & \dots & (\boldsymbol{c}^{(n)})^\top \end{bmatrix}^\top$ where $\boldsymbol{c}^{(k)}\in\mathbb{R}^2$ for $1\leq k\leq n$, and $c^{(0)}\in\mathbb{R}$

and input weights $\boldsymbol{b}_0 = \begin{bmatrix} b^{(0)} & (\boldsymbol{b}^{(1)})^\top & \dots & (\boldsymbol{b}^{(n)})^\top \end{bmatrix}^\top$ with components defined likewise.

Then, the filter implemented by the system is given by

$$g(t) = \boldsymbol{c}^\top e^{\boldsymbol{A}_0 t}\boldsymbol{b}_0 = e^{-at}c^{(0)}b^{(0)} + \sum_{k=1}^{n}(\boldsymbol{c}^{(k)})^\top e^{\boldsymbol{A}_0^k t}\boldsymbol{b}^{(k)} \tag{114}$$

$$= e^{-at}\left(c^{(0)}b^{(0)} + \sum_{k=1}^{n}(\boldsymbol{c}^{(k)})^\top \begin{pmatrix}\cos(k\omega t) & \sin(k\omega t)\\ -\sin(k\omega t) & \cos(k\omega t)\end{pmatrix}\boldsymbol{b}^{(k)}\right) \tag{115}$$

Then, it is easy to see that choosing $b^{(0)} = \frac{e^{a\tau}}{\sqrt{N}}$, $\boldsymbol{b}^{(k)} = \frac{e^{a\tau}(-1)^k}{\sqrt{N}}\begin{pmatrix}\sqrt{2} & 0\end{pmatrix}^\top$ and $c^{(0)} = \frac{1}{\sqrt{N}}$, $\boldsymbol{c}^{(k)} = \frac{1}{\sqrt{N}}\begin{pmatrix}\sqrt{2} & 0\end{pmatrix}^\top$ yields

$$g(t) = \frac{e^{-a(t-\tau)}}{N}\left(1 + 2\sum_{k=1}^{n}\cos(k\omega t)(-1)^k\right). \tag{116}$$

Choosing $\omega = \frac{\pi}{\tau}$ then yields

$$g(t) = \frac{e^{-a(t-\tau)}}{N}\left(1 + 2\sum_{k=1}^{n}\cos(k\omega(t-\tau))\right) \tag{117}$$

$$= \frac{e^{-a(t-\tau)}}{N}D_n(\omega(t-\tau)), \tag{118}$$

which has the desired property that $g(\tau) = 1$.

However, as stated, this construction uses a normal dynamics matrix W, and thus cannot achieve transient amplification necessary to stay activity efficient. In particular, the total activity is given by

$$E_x = \int_0^\infty \|\widetilde{\boldsymbol{x}}_0(t)\|^2 dt = \int_0^\infty \|e^{\boldsymbol{A}_0 t}\boldsymbol{b}_0\|^2 dt \tag{119}$$

$$= \frac{1}{N}\int_0^\infty e^{-2a(t-\tau)}\left(1 + 2\sum_{k=1}^{n}(\cos^2(k\omega t) + \sin^2(k\omega t))\right) \tag{120}$$

$$= \frac{e^{2a\tau}}{2a}. \tag{121}$$

This quantity is minimized at $a^* = \frac{1}{2\tau}$, thus yielding a total activity of $E_x^* = e\tau$, which does not improve with N.

To improve the activity efficiency of this solution, we apply a similarity transform $\boldsymbol{S} = I + (\kappa - 1)\boldsymbol{c}\boldsymbol{c}^\top$, such that $\boldsymbol{A} = \boldsymbol{S}\boldsymbol{A}_0\boldsymbol{S}^{-1}$, together with $\boldsymbol{b} = \frac{1}{\kappa}\boldsymbol{S}\boldsymbol{b}_0$. This stretches activity along the readout direction, relative to other orthogonal directions, by a factor κ, yielding a filter $g(t)$ identical to that implemented by $(\boldsymbol{c}, \boldsymbol{W}_0, \boldsymbol{b}_0)$. The total activity under these transformations then becomes

$$E_x = \frac{1}{\kappa^2}\int_0^\infty \boldsymbol{b}_0^\top e^{\boldsymbol{A}_0^\top t}\boldsymbol{S}^\top\boldsymbol{S}e^{\boldsymbol{A}_0 t}\boldsymbol{b}_0 dt \tag{122}$$

$$= \frac{1}{\kappa^2}\int_0^\infty \boldsymbol{b}_0^\top e^{\boldsymbol{A}_0^\top t}(\boldsymbol{I} + (\kappa^2 - 1)\boldsymbol{c}\boldsymbol{c}^\top)e^{\boldsymbol{A}_0 t}\boldsymbol{b}_0 dt \tag{123}$$

$$= \frac{1}{\kappa^2}\int_0^\infty \|\widetilde{\boldsymbol{x}}_0(t)\|^2 dt + \left(1 - \frac{1}{\kappa^2}\right)\int_0^\infty g(t)^2 dt \tag{124}$$

$$= \frac{1}{\kappa^2}\frac{e^{2a\tau}}{2a} + \left(1 - \frac{1}{\kappa^2}\right)\int_0^\infty g(t)^2 dt. \tag{125}$$

Thus, for large κ, the total activity under this transformation approaches $\int_0^\infty g(t)^2 dt$, and is therefore concentrated along the readout direction.

The readout-aligned activity integral can then be simplified as (Mathematica)

$$\int_0^\infty g(t)^2 dt = \frac{1}{N^2}\int_0^\infty e^{-2a(t-\tau)} D_n^2(\omega(t-\tau))dt \tag{126}$$

$$= \frac{\tau e^{2a\tau}}{N}\left[\frac{1}{2a\tau} + 4a\tau \sum_{m=1}^{N-1} \frac{\left(1-\frac{m}{N}\right)(-1)^m}{4(a\tau)^2 + m^2\pi^2}\right] \tag{127}$$

$$= \frac{\tau e^{2a\tau}}{N}\left[A_N - \frac{1}{N}B_N\right], \tag{128}$$

where $A_N = \frac{1}{2a\tau} + 4a\tau \sum_{m=1}^{N-1} \frac{(-1)^m}{4(a\tau)^2+m^2\pi^2}$ and $B_N = 4a\tau \sum_{m=1}^{N-1} \frac{m(-1)^m}{4(a\tau)^2+m^2\pi^2}$.

Observe that $\lim_{N\to\infty} A_N = \frac{1}{\sinh(2a\tau)}$ exactly, and thus $A_N = \frac{1}{\sinh(2a\tau)} + \mathcal{O}(N^{-2})$. In addition, we have that $B_N = \mathcal{O}(1)$ as B_N is convergent. Thus, we have the leading asymptotics

$$\int_0^\infty g(t)^2 dt = \frac{\tau e^{2a\tau}}{N\sinh(2a\tau)} + \mathcal{O}\left(\frac{1}{N^2}\right) \tag{129}$$

$$= \frac{2\tau}{N(1-e^{-4a\tau})} + \mathcal{O}\left(\frac{1}{N^2}\right). \tag{130}$$

Estimating the optimal $a^\star$ would require more careful treatment of the subleading terms. For illustrative purposes however, setting $a = \frac{1}{2\tau}$ yields $\int_0^\infty g(t)^2 dt \approx \frac{2\tau}{N(1-e^{-2})}$.

Thus, in the limit $\kappa \to \infty$, this nonnormal oscillatory construction also achieves an activity cost $E_x = \mathcal{O}\left(\frac{\tau}{N}\right)$, demonstrating that the activity-efficient scaling derived above does not rely specifically on feedforward chain connectivity. Nonetheless, like the feedforward chain, it relies on transient amplification arising from high-rank, nonnormal recurrent weights.

C Noise robustness

As in Sec. 2.3, we consider isotropic process noise added to the dynamics of variance σ^2 injected into the dynamics:

$$\frac{d\boldsymbol{x}}{dt} = -\boldsymbol{x} + \boldsymbol{W}\boldsymbol{x} + \boldsymbol{b}u(t) + \boldsymbol{\eta}(t), \tag{131}$$

where $\mathbb{E}[\boldsymbol{\eta}(t)] = 0$ and $\mathbb{E}[\boldsymbol{\eta}(t)\boldsymbol{\eta}(t')^\top] = \sigma^2 I\delta(t-t')$. Then, the readout $y(\tau)$ satisfies

$$\mathbb{E}_\eta[y(\tau)] = (\boldsymbol{c}^\top e^{\boldsymbol{A}\tau}\boldsymbol{b})z \tag{132}$$

$$\mathrm{Var}_\eta[y(\tau)] = \sigma^2 \boldsymbol{c}^\top \left(\int_0^\tau e^{\boldsymbol{A}t} e^{\boldsymbol{A}^\top t} dt\right)\boldsymbol{c} = \sigma^2 \int_0^\tau \|\boldsymbol{q}(t)\|^2 dt \tag{133}$$

If we define $\widetilde{\boldsymbol{q}}(t) = \boldsymbol{q}(\tau - t)$ for $t \in [0,\tau]$, the more standard convention for the adjoint dynamics variable that evolves backwards in time, it is easy to see that

$$\frac{d}{dt}\widetilde{\boldsymbol{q}}(t)^\top \boldsymbol{x}(t) = \frac{d}{dt}\boldsymbol{c}^\top e^{\boldsymbol{A}(\tau-t)} e^{\boldsymbol{A}t}\boldsymbol{b} = \frac{d}{dt}\boldsymbol{c}^\top e^{\boldsymbol{A}\tau}\boldsymbol{b} = 0. \tag{134}$$

Thus, combining this with the readout gain constraint $\boldsymbol{c}^\top e^{\boldsymbol{A}\tau}\boldsymbol{b} = \widetilde{\boldsymbol{q}}(0)^\top \widetilde{\boldsymbol{x}}(0) = 1$, we have the conservation law that

$$\widetilde{\boldsymbol{q}}(t)^\top \widetilde{\boldsymbol{x}}(t) = 1 \tag{135}$$

for all $t \in [0,\tau]$. Thus, minimizing activity norm necessarily requires $\widetilde{\boldsymbol{q}}$ to grow in norm to compensate. More concretely, from Cauchy-Schwarz, we have:

$$\left(\int_0^\tau \widetilde{\boldsymbol{q}}(t)^\top \widetilde{\boldsymbol{x}}(t)dt\right)^2 \le \left(\int_0^\tau \|\widetilde{\boldsymbol{x}}(t)\|^2 dt\right)\left(\int_0^\tau \|\widetilde{\boldsymbol{q}}(t)\|^2 dt\right) \le E_x E_q. \tag{136}$$

Thus, $E_x E_q \ge \tau^2$.

C.1 Noise robustness of the amplifying chain

For the concrete ansatz of the homogeneous feedforward chain $\boldsymbol{A} = -a\boldsymbol{I} + k\boldsymbol{\Gamma}$, the asymptotics of E_q can be explicitly computed:

$$E_q = \int_0^\infty \|\boldsymbol{q}(t)\|^2 dt \tag{137}$$

Using that $q_i(t) = e^{-at}\frac{(kt)^i}{i!}$, we have that

$$\|\boldsymbol{q}(t)\|^2 = e^{-2at}\sum_{j=0}^{N-1}\frac{(kt)^{2j}}{(j!)^2} \tag{138}$$

Performing the integral then yields

$$E_q = \frac{1}{2a}\sum_{j=0}^{N-1}\binom{2j}{j}\left(\frac{\rho^2}{4}\right)^j. \tag{139}$$

Using the approximation that $\binom{2j}{j} \sim \frac{4^j}{\sqrt{\pi j}}$, we have that

$$E_q \sim \frac{1}{2a}\sum_{j=0}^{N-1}\frac{\rho^{2j}}{\sqrt{\pi j}}. \tag{140}$$

For $\rho > 1$, this expression diverges exponentially fast in N.

C.2 Adjoint cost sets an activity scale

In the limit of vanishing weight cost $\lambda \to 0$, the activity and adjoint costs approach the lower bound $E_x E_q \approx \tau^2$. Under this assumption, we have that $E_x + \gamma E_q = E_x + \gamma\frac{\tau^2}{E_x}$, which then yields $E_x^* = \gamma^{1/2}\tau$ and $E_q^* = \gamma^{-1/2}\tau$. If we then take $\|\widetilde{\boldsymbol{x}}(t)\|$ and $\|\boldsymbol{q}(t)\|$ to be constant throughout $[0,\tau]$, as is observed numerically, we then obtain $\|\widetilde{\boldsymbol{x}}(t)\| \approx \gamma^{1/4}$ and $\|\boldsymbol{q}(t)\| \approx \gamma^{-1/4}$ for $t \in [0,\tau]$.

D Numerical experiments and training details

Details for running numerical experiments in this paper are stated below. All code will be made publicly available upon acceptance.

D.1 Linear RNN experiments

To numerically optimize the objective in Eq. 4, we used the dissipative Hamiltonian parameterization (Mehl et al., 2016) of the effective recurrent weights $\boldsymbol{A} = -\boldsymbol{I} + \boldsymbol{W}$, given by

$$\boldsymbol{A} = (\boldsymbol{J} - \boldsymbol{R}_0)\boldsymbol{Q},$$

where $\boldsymbol{J}$ is skew-symmetric, $\boldsymbol{R}_0 \succ \boldsymbol{0}$, and $\boldsymbol{Q} \succ \boldsymbol{0}$, thus enforcing Hurwitz stability by construction. This was necessary to avoid numerical issues when solving the Lyapunov equation $\boldsymbol{A}^\top\boldsymbol{R} + \boldsymbol{R}\boldsymbol{A} = -\boldsymbol{I}$ for the Gramian $\boldsymbol{R}$ appearing in the activity cost using a Lyapunov solver.

Optimization was performed in JAX with the L-BFGS-B optimizer. We use this optimizer because standard first-order methods converged much more slowly. We used networks of size $N = 30$ (Fig. 1) or $N = 50$ (Figs. 2, 3). These network sizes were chosen for efficiency of the optimization process, and were selected to be large enough to accommodate the largest effective rank observed across the λ and τ values considered.

D.2 Nonlinear RNN experiment (continuous readout window task)

As described in the main text, we use $\phi = \tanh$ RNNs of the form

$$\dot{\boldsymbol{x}} = -\boldsymbol{x} + \boldsymbol{W}\phi(\boldsymbol{x}(t)) + \boldsymbol{b}u(t) \tag{141}$$

with readout $y(t) = \boldsymbol{c}^\top\phi(\boldsymbol{x}(t))$. To approximate the constrained optimization problem, we use the loss

$$\mathcal{L} = \alpha\mathbb{E}_z\left[\int_\tau^{\tau+\xi}(y(t)-z)^2dt\right] + \frac{1}{1+\lambda}\mathbb{E}_z\left[\int_0^T\|\phi(\boldsymbol{x}(t))\|^2dt\right] + \frac{\lambda}{1+\lambda}\left(\|\boldsymbol{W}\|_F^2 + \|\boldsymbol{b}\|^2\right), \tag{142}$$

where we set the task term strength as $\alpha = 10^3$, and the weight and activity costs have been normalized by $1 + \lambda$ so that the weight of the constraint-enforcing terms relative to the task term is roughly consistent across λ.

As in the linear case, we fix the readout weights to be unit norm, and simplify the optimization further by fixing it to be along the uniform direction as $\boldsymbol{c} = \frac{1}{\sqrt{N}}\mathbf{1}$. We use networks of $N = 80$ neurons and discretize the dynamics with a time constant timestep $dt = 0.05$. Networks were trained with the Adam optimizer at a learning rate of 1e-3 for 5×10^5 steps.

To compute the adjoint variable $\boldsymbol{q}(t) = \frac{dy(\tau)}{d\boldsymbol{x}(t)}$ for Fig. 4e, we numerically integrated the backwards-in-time adjoint dynamics

$$-\frac{d\boldsymbol{q}(t)}{dt} = \boldsymbol{J}(\boldsymbol{x}(t))^\top \boldsymbol{q}(t) \tag{143}$$

with initial condition $\boldsymbol{q}(\tau) = \text{diag}(\phi'(\boldsymbol{x}(\tau)))\boldsymbol{c}$, where $\boldsymbol{J}(\boldsymbol{x}) = -\boldsymbol{I} + \boldsymbol{W}\text{diag}(\phi'(\boldsymbol{x}))$ is the Jacobian of the forward dynamics.

D.3 ALM MODEL AND TRAINING

We train networks with $N = 128$ neurons, and discretize the dynamics with timestep $dt = 0.25\tau_x$ (units of time constants, with τ_x = 40 ms). For the fixed reservoir used to preprocess the pulse inputs before passing as input to the trained networks, we use the same random reservoir parameters with $\tanh$ nonlinearity as in Pereira-Obilinovic et al. (2026) (described in https://github.com/ulisespereira/alm-coding-residuals-condavg-train), with dynamics

$$\dot{\boldsymbol{z}} = -\boldsymbol{z} + \boldsymbol{W}_{\text{octx}} \tanh(\boldsymbol{z}) + \boldsymbol{I}_{\text{stim}}(t). \tag{144}$$

The state of this reservoir is then passed as input $\boldsymbol{u}(t) = \tanh(\boldsymbol{z}(t))$ to the trained network through weights $\boldsymbol{B}$.

We use the following specification for the timing of task periods:

- $t_{\text{start}} = -2.1$
- sample cue onset at $t_{\text{sample}} = -1.85$ s consisting of three 150 ms pulses with 100 ms gaps in between, occurring over $[-1.85, -1.20]$ s
- delay period onset at -1.2 s
- go cue at 0 s
- Response window: $[t_{\text{resp}} + 0.1, +0.35]$ s
- heldout window: $[+0.1, t_{\text{end}} = 0.6]$ s

We train networks with the objective

$$\mathcal{L} = \alpha\mathcal{L}_{\text{task}} + \frac{1}{1+\lambda}\int_{t_{\text{sample}}}^{t_{\text{end}}} \|\phi(\boldsymbol{x})\|^2 dt + \frac{\lambda}{1+\lambda}\left(\|\boldsymbol{W}\|_F^2 + \|\boldsymbol{B}\|_F^2\right), \tag{145}$$

where $\mathcal{L}_{\text{task}} = \frac{\int_{t_{\text{start}}}^{t_{\text{resp}}} \sum_{c\in\{L,R\}} (y_c(t) - y_c^\star(t))^2 dt}{\int_{t_{\text{start}}}^{t_{\text{resp}}} \sum_{c\in\{L,R\}} y_c^\star(t)^2 dt}$. We choose a task term strength $\alpha = 10^4$, and use the Adam optimizer with learning rate lr=2 $\times$ 10^{-3} and cosine annealing to 0.05 lr over 8000 epochs. We initialized recurrent weights and inputs weights as $\boldsymbol{W}_{ij} \sim (10^{-3}/\sqrt{N})\mathcal{N}(0,1)$, $\boldsymbol{B}_{ia} \sim (1/\sqrt{200})\mathcal{N}(0,1)$, the biases as $\boldsymbol{b}^0 = \mathbf{0}$, and the max firing rates as $r_i^{\max} = 1$. For each combination of $\lambda \in \{0.1, 0.3, 1, 3, 10, 30, 100\}$ and $\sigma \in \{0, 10^{-3}, 10^{-2}\}$, we trained a total of 20 networks with different random seeds.

D.3.1 PERTURBATION AND ANALYSIS PROCEDURES

To estimate task coding directions, we averaged the condition difference direction $\boldsymbol{r}_R(t) - \boldsymbol{r}_L(t)$ over the following periods:

- sample: first 0.6 s of the sample period
- choice: last 0.6 s of the delay period
- response: 0.6 s following the onset of the go cue

Perturbations were applied during the early delay period at a time $t_0 \in \{-1.1, -0.9, -0.7\}$ and in each trial condition as a shift to the state $\boldsymbol{x}(t_0) \to \boldsymbol{x}(t_0) + g\hat{\boldsymbol{v}}$, where $\hat{\boldsymbol{v}}$ is a unit direction representing the perturbed direction with a signed perturbation strength $g \in \pm\{2, 4, 6, 8, 10, 12\}$. To determine whether or not a flip of the choice occurred, we use a similar criterion as in (Pereira-Obilinovic et al., 2026): a trial with condition $c \in \{L, R\}$ was considered flipped by a perturbation if the network output averaged over the last 50 timesteps $\overline{y}_c$ satisfied $|\overline{y}_c - \overline{y}^\star_{\neg c}| < \varepsilon$, where $\varepsilon = 0.25|\overline{y}^\star_R - \overline{y}^\star_L|$. The flip rate was then measured as the fraction of all combinations of g, trial conditions, and perturbation times t_0 that resulted in a flip.

For Fig. 5b, we report nonnormality as Henrici's departure

$$\frac{\sqrt{\|\boldsymbol{W}\|_F^2 - \sum_i |\lambda_i(\boldsymbol{W})|^2}}{\|\boldsymbol{W}\|_F}, \tag{146}$$

where $\lambda_i(\cdot)$ represents the ith eigenvalue.

E SUPPLEMENTARY FIGURES

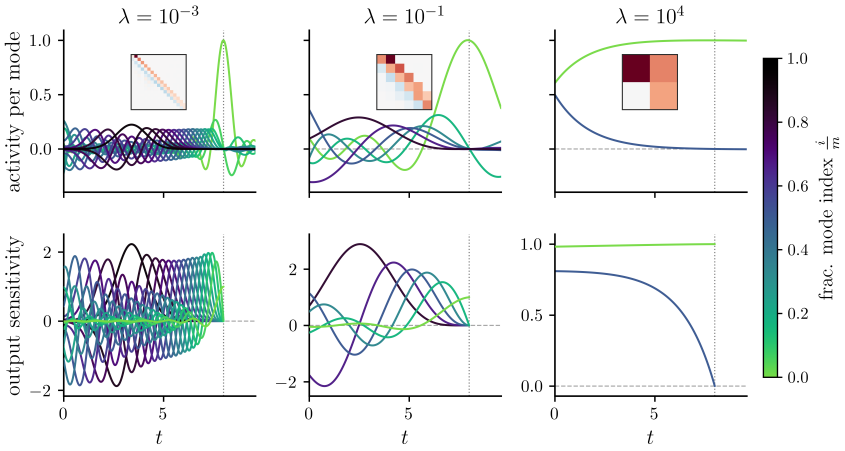


Figure E.1: Analogous to Fig. 2a, but with an adjoint cost $\gamma = 0.01$.

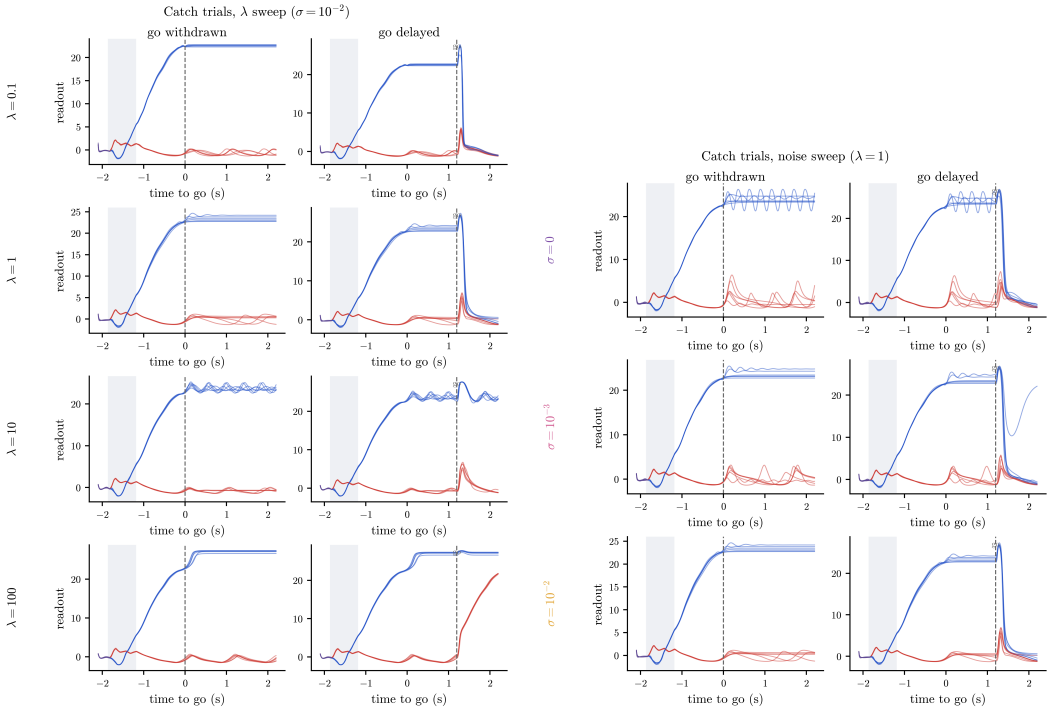


Figure E.2: Catch trials where the go cue is either withdrawn entirely or delayed by 1.2 s, shown across λ at fixed $\sigma = 10^{-2}$ (left) and across σ at fixed $\lambda = 1$ (right).

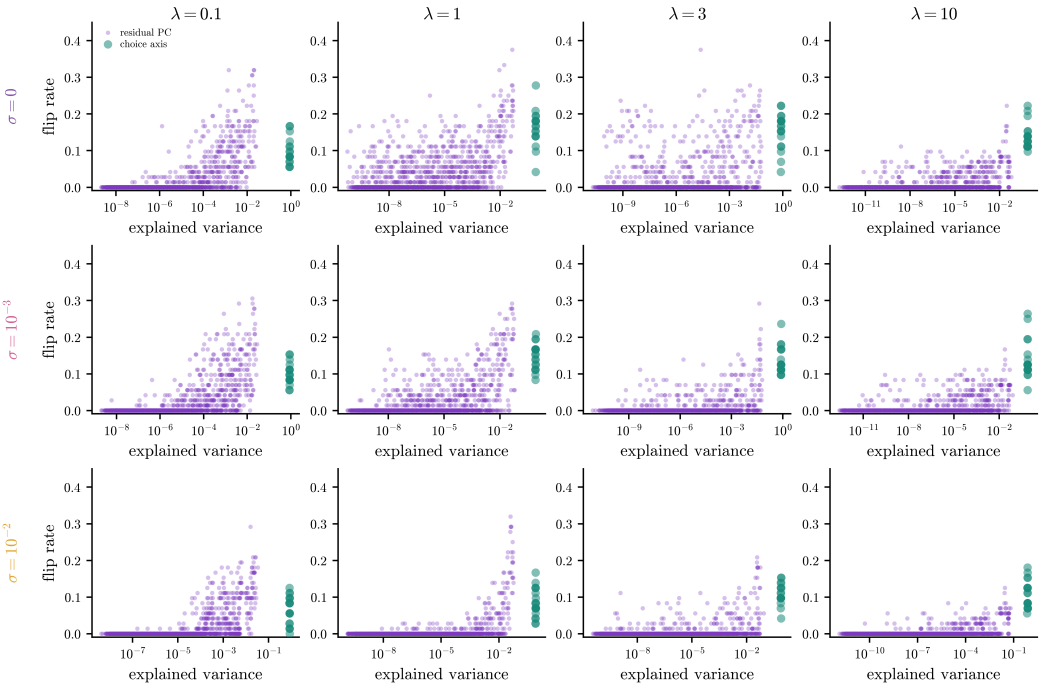


Figure E.3: Sweep of explained variance versus flip rate across σ and λ.