Understanding Memory in Neural Networks through Fisher Information Diffusion

Anonymous authorsPaper under double-blind review

000

001

002003004

010 011

012

013

014

016

017

018

019

021

025

026027028

029

031

033

034

037

038

040

041 042

043

044

046

047

048

051

052

ABSTRACT

Information retention and transmission are fundamental to both artificial and biological neural networks. We present a general theoretical framework showing how information can be maintained on dynamically stable manifolds that evolve over time while preserving the geometry of inputs. In contrast to classical memory models such as Hopfield networks, which rely on static attractors, our approach highlights evolving stable subspaces as the substrate of memory. A central contribution of our work is the use of dynamic mean-field theory to uncover a new principle: operating at criticality (spectral radius ≈ 1) is necessary but not sufficient for reliable information retention. Equally crucial—yet overlooked in prior studies—is the alignment between the input structure and the stable subspace. The theory leads to simple initialization rules that guarantee stable dynamics at the edge of chaos. We validate these rules in basic recurrent networks, showing that Fisher information-optimized initialization accelerates convergence and improves accuracy in sequential memory tasks, including the copy task and sequential MNIST compared to standard random initialization. Together, these results provide both principled design guidelines for recurrent networks and new theoretical insight into how information can be preserved over time.

1 Introduction

Recurrent neural networks (RNNs) are fundamental models for processing sequential data, and their dynamics have been a longstanding focus in both neuroscience and machine learning. Early work on random networks established that criticality and chaos play central roles in determining memory lifetime and information retention (Sompolinsky et al., 1988; Ganguli et al., 2008). Building on this foundation, subsequent approaches have sought to stabilize recurrent dynamics through architectural constraints, such as unitary and orthogonal parameterizations (Arjovsky et al., 2016; Jing et al., 2017), or through adaptive state-space models with learnable dynamics (Gu et al.). Other theoretical directions have explored modular assemblies of RNNs (Kozachkov et al.), traveling-wave dynamics as carriers of short-term memory (Ermentrout et al.), and input-driven circuit reconfiguration near criticality (Mastrogiuseppe et al.). Together, these works underscore that memory and stability emerge not from isolated units, but from the interplay between structured connectivity, dynamical regimes, and input geometry.

Traditional analyses of information capacity, such as Hopfield networks, assume that information is stored in stationary fixed points of the network dynamics. While these models have been influential, they are insufficient for modeling working memory in recurrent architectures more broadly. Unlike long-term memory, which can tolerate compression or abstraction, working memory requires preserving the fine-grained distinctions and relational geometry of inputs—maintaining not only class-level information but also the detailed differences between stimuli over short timescales. A static attractor framework would imply that memory corresponds to retrieval from a finite set of stored states, invariant across trials, which fails to capture this geometry-preserving requirement. Notably, recent neurophysiological evidence shows that working memory in the primate cortex is supported by stable dynamic manifolds with rotational dynamics (Ritter & Chadwick, 2025b).

Despite these advances, the theoretical understanding of information dynamics in recurrent networks remains limited. Most analyses assume either fully dense i.i.d. connectivity or a single structured population, oversimplifying both modern architectures and biological circuits. What is missing

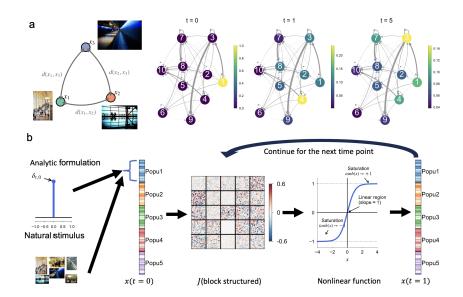


Figure 1: Illustration of Fisher information diffusion and experimental setup. (a) Unlike traditional graph diffusion models, which track the spread of neural activities, we study the diffusion of Fisher information, quantifying how well each subpopulation retains the geometry of the data (i.e., pairwise sample distances $(d(x_i, x_j))$ where x_i, x_j are input samples) over time in a recurrent multi-subpopulation network. (b) Schematic of the network architecture and its time evolution. The network consists of multiple subpopulations with recurrent connections drawn from a zero-mean Gaussian distribution with specified variance, representing connection strength. Inputs can be flexibly configured; for clarity, we show the case where either an impulse (for analytic derivation) or natural images (for testing) are provided to the first subpopulation.

is a general framework for describing how information propagates and is preserved in networks composed of multiple interacting subpopulations with distinct connectivity statistics. Filling this gap is essential for understanding the mechanisms of working memory and for deriving principled design rules that optimize information retention in recurrent architectures.

A Framework for Information Dynamics To this end, we introduce a framework that models general networks as block-structured systems of interacting subpopulations. Layers or modules are concatenated into a one-dimensional vector, and their interactions appear as blocks in the overall connectivity matrix (Fig. 1b). Feedforward architectures emerge as a special case with only adjacent connections, while more general configurations—including feedback loops and skip connections—are naturally encoded through the recurrent interactions among subpopulations. Within this unified representation, we analyze Fisher information dynamics directly in the space of block-structured connectivity. Drawing an analogy to thermal physics, we treat the propagation of Fisher information as a diffusion process across subpopulations (Fig. 1a). The resulting *Fisher diffusion operator* provides an analytic, Markovian characterization of how information about inputs is retained and transmitted.

In this dynamic view of the information in terms of diffusion, the theory shows that stimulus representations in recurrent networks are not fixed but evolve continuously on low-dimensional stable manifolds. While these representations change as the network unfolds in time, the geometry of the input space—defined as pairwise distinctions between stimuli—remains preserved. This dynamic perspective provides a more flexible notion of memory, where information is retained not as static states but as evolving trajectories that maintain the relational structure of inputs.

We validate the theoretical predictions in networks with block-structured connectivity matrices, where the variance of Gaussian weights in each block is controlled. We show that the Fisher diffusion operator accurately captures both the magnitude and temporal dynamics of Fisher information across subpopulations. Extending to natural image inputs, the theory also predicts how information about input geometry—pairwise distances between images—is preserved.

Finally, because representations in our framework are not fixed-point attractors, we ask whether such networks can support memory for sequences of stimuli. From our theory, we derive simple initialization rules that place the network at the edge of chaos while maximizing Fisher information retention. When tested on simple recurrent architectures, these Fisher information—optimized initializations consistently yield faster convergence and higher accuracy on sequential memory tasks—including the copy problem and sequential MNIST—compared to standard random initialization.

In summary, our framework generalizes beyond i.i.d. or single-population analyses, provides an analytic tool for studying information dynamics in recurrent networks, and offers principled initialization rules that directly connect network connectivity, criticality, and Fisher information flow.

2 RECURRENT NETWORKS WITH BLOCK-STRUCTURED CONNECTIVITY

We begin with the general dynamics of recurrent networks, then introduce a block-structured generalization that enables analysis at the subpopulation level.

Recurrent dynamics. We consider a recurrent network of N neurons with discrete-time dynamics

$$h_i(t) = \sum_{j=1}^{N} J_{ij} S_j(t) + \eta_i(t), \qquad S_j(t) = \phi[w_j x(t) + h_j(t-1)],$$
 (1)

where $h_i(t)$ is the internal state of neuron i, $S_i(t)$ its output, and J_{ij} is the connectivity matrix connecting the neuron j to neuron i. The activation function is $\phi(x) = \tanh(x)$ with $\phi'(0) = 1$. The network receives an external input x(t) through weights w_i , and each neuron is driven by independent Gaussian noise $\eta_i(t)$ with zero mean and covariance $\langle \eta_i(t) \eta_j(s) \rangle = \sigma^2 \delta_{ij} \delta_{ts}$.

Subpopulation structure. Classical analyses typically assume that the connectivity matrix J is i.i.d. Gaussian (Toyoizumi & Abbott, 2011), corresponding to a single homogeneous population. To capture more general network structures, we partition the network state $\mathbf{h}(t) \in \mathbb{R}^N$ into M subpopulations (Fig. 1b). Each neuron i is assigned a label $m(i) \in \{1,\ldots,M\}$, with subpopulation m containing a fraction f_m of the neurons such that $\sum_{m=1}^M f_m = 1$. The resulting connectivity matrix J acquires a block structure, where each block encodes connections between two subpopulations. This formulation generalizes standard feedforward or layered networks: purely feedforward connectivity appears as a special case, while feedback and skip connections are naturally represented by off-diagonal blocks.

Block-structured connectivity. Within the mean-field approximation, weights are modeled as independent Gaussians with zero mean and block-dependent variances: $\langle J_{ij} \rangle_J = 0, \langle J_{ij}^2 \rangle_J = \frac{1}{N} g_{m(i)n(j)}^2$. Here m(i) and n(j) denote the subpopulations of neurons i and j. Each block encodes connections from subpopulation n to m, with variance parameter g_{mn}^2 controlling its strength. We will refer to the block-gain matrix G with entries $G_{mn} \equiv g_{mn}^2 f_n$.

Fisher information. We probe memory with an impulse input $x(t) = \theta \, \delta_{t,0}$ and study how well its amplitude θ is preserved across time and populations. The Fisher information (FI) about θ is

$$\mathcal{I}(\theta, t) = \mathbb{E}_{p(\mathbf{h}(t)|\theta)} \left[\frac{\partial^2}{\partial \theta^2} \log p(\mathbf{h}(t) \mid \theta) \right], \tag{2}$$

which defines the Fisher memory curve (Ganguli et al., 2008).

For a fixed J, $p(\mathbf{h}|\theta)$ is Gaussian, the FI simplifies to (see Appendix A.1):

$$\mathcal{I}(\theta, t) = \left\langle \frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta} \log p(\mathbf{h}(t)|\theta) \right\rangle_{p(\mathbf{h}(t)|\theta)}, \tag{3}$$

Mean-field theory yields a block-diagonal covariance $\left\langle \Sigma_{ij}(t) \right\rangle_J = \delta_{ij}(q_{m(j)})$ with population-specific variances $q_{m(j)}$, giving (Appendix A.1)

$$\mathcal{I}(\theta, t) = N \sum_{m} \frac{f_m}{q_m} \left\langle \left(\frac{\partial \mu_m(t)}{\partial \theta}\right)^2 \right\rangle_J, \qquad q_m = \sigma^2 + \sum_{n} G_{mn} \left\langle S_n^2 \right\rangle_n, \tag{4}$$

where $\mu_m(t) \equiv \langle h_m(t) \mid J \rangle$. Although $\mu_m(t) = 0$ on average, FI remains nonzero since it depends on the variance of the sensitivity $\partial \mu_m(t)/\partial \theta$, which is shaped by both the nonlinearity ϕ and the inter-population connectivity. Importantly, this sensitivity is not constant over time and differs across populations.

Fisher information diffusion. We now summarize the key result (derivation in Appendix A.2). We have derived the Fisher diffusion operator A that propagates sensitivities $\langle (\partial \mu_m(t)/\partial \theta)^2 \rangle_J$ across subpopulations from one time step to the next. For two subpopulations, it factorizes into a connectivity and a sensitivity term:

$$A = \underbrace{\begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}}_{\text{Connectivity}} \cdot \underbrace{\begin{pmatrix} \langle (S')^2 \rangle_1 & 0 \\ 0 & \langle (S')^2 \rangle_2 \end{pmatrix}}_{\text{Sensitivity}},$$
 (5)

where $\langle (S')^2 \rangle_n$ denotes the mean squared derivative of the activation function in subpopulation n. The connectivity block captures how information is routed between groups, while the sensitivity block captures how nonlinear responses modulate this transfer. In Appendix A.2 we provide an analytic expression for $\langle (S')^2 \rangle_n$ that depends solely on the block-gain matrix G, making the operator fully determined by network structure. In the linear limit—when the activation function is purely linear—one has $\langle (S')^2 \rangle_n = 1$ for all subpopulations. In this case the sensitivity block reduces to the identity, and the Fisher information operator A coincides with the block-gain matrix G itself. Repeated application of A describes how information flows across populations. The total FI at time t is then $\frac{\mathcal{I}(t)}{N} = \sum_{mn} \frac{f_m}{g_m} (A^t)_{mn} w_n^2$, and $\mathcal{I}(0) = 0$.

Crucially, although individual neural activities evolve nonlinearly, the collective statistics of subpopulations can be expressed as evolving linearly under the diffusion operator. The nonlinearity is absorbed into the term $\langle (S')^2 \rangle_n$, which itself is a nonlinear function of the block gains G. In this way, the operator acts analogously to a transfer matrix in graph diffusion, providing a linear structure that enables a tractable analytic description of how connectivity shapes the encoding and preservation of information over time. We present two tests to evaluate our analytic characterization of Fisher information.

Direct Fisher quantification. First, we directly estimate Fisher information from network simulations. We simulate a network with $N=10{,}000$ neurons, $f_1=f_2=0.5,\,\sigma=0.1,$ and input weights $w_1=1,w_2=0$, so that the impulse is applied only to the first subpopulation. For input amplitudes $\theta\in\{0,-0.1,0.1\}$, we compute the sensitivity term $\left\langle (\partial\mu_m(t)/\partial\theta)^2\right\rangle_J$ in Eq. equation 4 (see Appendix A.3). To test how connectivity motifs affect information dynamics, we consider three configurations: (i) self-recurrence only, (ii) feedforward coupling, and (iii) feedback coupling (Fig. 2a–c). Across all cases, the analytic solution based on the diffusion operator accurately matches simulations—capturing not only the magnitude of Fisher information in each population, but also the temporal dynamics, including oscillatory flow between subpopulations. Finally, we examine how the agreement scales with network size N: the mean-squared error between simulated and analytic trajectories decreases rapidly and becomes negligible for $N \geq 1000$ (Fig. S3).

Preservation of input geometry. Second, we tested whether Fisher information predicts how well the network preserves the geometry of natural inputs. Appendix A.5 provides an intuitive proof linking Fisher optimization to isometry preservation. Concretely, we presented 15,619 CIFAR-10 images (flattened to dimension 7500) as inputs to the first subpopulation of a network with N=15,000, $f_1=f_2=0.5$, and $\sigma=0.1$. Each image was processed individually, and at time t we recorded the activity vectors of both subpopulations as the network's internal representations. To quantify geometry preservation, we computed all pairwise Euclidean distances between the original images, and likewise all pairwise distances between their corresponding neural representations (see Appendix A.4). We then measured the correlation between these two distance matrices: a correlation of 1 would indicate perfect isometry (geometry preserved exactly), while lower correlations reflect increasing distortion. This correlation therefore serves as a direct measure of how faithfully the network preserves the relational structure of its inputs (Fig. 2d–f). Although this metric is distinct from Fisher information, it produces the same qualitative conclusions: the analytic framework accurately predicts both the oscillatory dynamics of information flow and the relative ability of different network motifs to preserve input geometry.

This empirical test highlights a key difference from Hopfield networks. In Hopfield models, memory is implemented by fixed-point attractors, and capacity is limited by the number of such stable states that can be stored. In our framework, by contrast, memory is defined by how well the differences between stimuli are preserved as activity evolves. This capacity does not depend on the number of stimuli presented, but instead on whether the network size N is sufficiently large relative to the sparsity of the input space—a condition closely analogous to the Restricted Isometry Property (RIP) for Gaussian matrices, which links the number of measurements to input sparsity (Foucart & Rauhut, 2013).

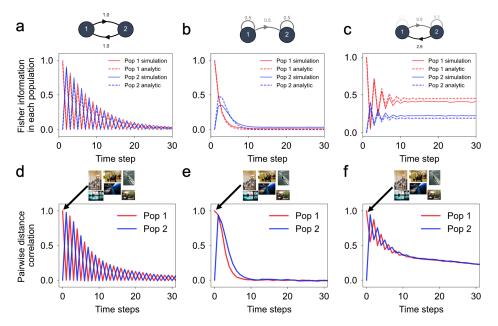


Figure 2: **Empirical validation of the Fisher diffusion framework.** (a–c) Time evolution of Fisher information in two subpopulations after an input pulse to Population 1 at t=0, under three motifs: (a) self-recurrence only, (b) feedforward only, and (c) a recurrent architecture close to the analytic optimum. Solid lines show simulation results; dashed lines show analytic predictions from the FI diffusion operator. The analytic framework accurately captures both the magnitude and temporal dynamics, including oscillatory information flow. (d–f) Geometry preservation for the same networks for CIFAR-10 images. Pairwise correlations between input distances and network representations quantify how well each architecture maintains input geometry. The results confirm that architectures predicted to optimize Fisher retention also preserve stimulus geometry more effectively.

3 CONDITIONS FOR OPTIMAL FISHER INFORMATION

In the context of information diffusion, achieving maximal long-term retention of an input requires two conditions on the diffusion operator A:

- 1. **Criticality.** The spectral radius of A must satisfy $\rho(A) = \max_i |\lambda_i| = 1$. If $\rho(A) < 1$, Fisher information decays exponentially; if $\rho(A) > 1$, it diverges uncontrollably. Criticality therefore guarantees the dynamic stability Kadmon & Sompolinsky (2015) such that information does not vanish at long times, but on its own it is not sufficient for optimal retention.
- 2. Eigenvector alignment and transient information. Let v denote the normalized right eigenvector associated with the leading eigenvalue $\lambda_{\max} = 1$. Asymptotically, $\lim_{t \to \infty} \mathcal{I}(t) \propto \|(\mathbf{w}^\top v)v\|_1$, where \mathbf{w} is the input configuration. Thus, only the input component aligned with v is retained indefinitely. From the perspective of network design, however, transient information carried by other modes should not decay too quickly. A complication is that A is generally nonnormal (not symmetric), so its eigenvalues and eigenvectors may be complex. Since the input configuration \mathbf{w} is real, the effective alignment with complex eigenvectors can be small, limiting

long-term retention. In practice, explicitly computing eigenvalues and eigenvectors is costly for networks with many subpopulations. By contrast, computing Fisher information over time only requires iterated multiplication by A, which is more scalable. Moreover, by choosing the integration horizon T, one can tune the emphasis between transient retention and long-term stability. A practical design objective is therefore to maximize the average Fisher information over a finite horizon: $\overline{\mathcal{I}} = \frac{1}{T} \sum_{t=1}^{T} \mathcal{I}(t)$. This metric balances stability at criticality with the preservation of transient information.

To illustrate these conditions, we analyze a simple two-population recurrent network. The block gain matrix is $G = \begin{pmatrix} G_{11} & G_{12} \\ G_{21} & G_{22} \end{pmatrix}$, so that the Fisher diffusion operator A depends on the four parameters G_{mn} . Because A can be summarized in terms of its trace and determinant, the parameter space can be reduced from four to two dimensions, enabling a clear visualization. We perform a dense grid search over all $G_{mn} \in [0,3]$. For each parameter setting, we compute: The spectral radius $\rho(A)$, used to identify the critical boundary $\rho(A) = 1$ (Fig. 3a), and the average Fisher information $\overline{\mathcal{I}}$ across 100 timesteps, aggregated over both populations (Fig. 3b). We observe that:

- 1. The critical boundary (grey dashed line) extends across the full range of Tr(G), confirming that criticality is a necessary condition for sustained information flow.
- 2. Optimal information retention occurs only along the critical boundary but is restricted to a narrower band of Tr(G), showing that criticality alone is not sufficient.

This demonstrates that while criticality is required for sustained information flow, alignment of the stable diffusion direction v with the input configuration \mathbf{w} is additionally necessary to achieve optimal Fisher information retention.

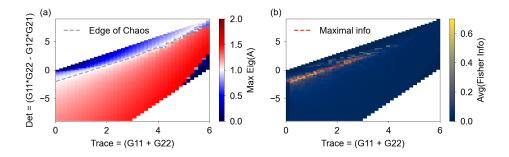


Figure 3: **Edge of chaos: necessary but not sufficient for optimal Fisher information.** (a) Phase diagram of the spectral radius $\rho(A)$ across network parameters $\mathrm{Tr}(G)$ and $\det(G)$. The grey dashed line marks the critical boundary $\rho(A)=1$. (b) Average Fisher information per timestep over 100 steps, showing that optimal retention occurs only within a restricted band of $\mathrm{Tr}(G)$ values.

4 OPTIMAL STRUCTURE FOR LONGER CHAINS

Our analytical framework extends naturally to networks of many subpopulations or deeper recurrent structures. For clarity and tractability, we focus on sequential chains in which only adjacent subpopulations are connected on through adjustable self-recurrent, feedforward, and feedback links (Fig. 4). The connectivity is captured by a generalized Toeplitz-like gain matrix G where only G_{mm} , $G_{m,m+1}$, and $G_{m+1,m}$ are nonzero, preserving the chain structure while allowing parameter flexibility. All subpopulations are equal in size $(f_m = 1/M)$, and input is applied only to the first subpopulation $(w_i = \delta_{i,1})$.

Within this architecture we build the Fisher information functional in terms of the allowed block gains G_{mn} and maximize the time- and population-averaged Fisher information $\overline{\mathcal{I}}$ over T=100 using differential evolution. The optimization reveals clear design principles: strong feedforward connections propagate signals efficiently, while carefully placed feedback stabilizes and modulates

this flow. Indiscriminate feedback is detrimental; instead, optimal networks exhibit sparse, strategically positioned feedback links that break the chain into nested loops for robust information retention. The characteristic broken-feedback pattern can be intuitively justified in the linear limit (Appendix A.6).

Finally, we find a striking scaling law (Fig. 4): keeping the total number of neurons fixed—yet large enough for mean-field theory—the network's total Fisher information grows approximately linearly with the number of subpopulations. Thus, deeper or more finely partitioned chains intrinsically possess greater information capacity when their connectivity is properly optimized.

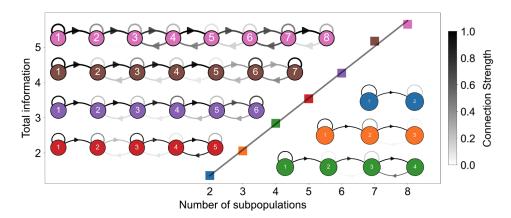


Figure 4: **Connectivity structures and information retention across subpopulations.** Inset diagrams show optimized connectivity patterns for networks with 2–8 subpopulations, revealing strong feedforward pathways, moderate self-recurrence, and sparse feedback loops. The main panel demonstrates that total Fisher information retention scales linearly with the number of subpopulations, indicating that greater modular depth enhances memory capacity.

5 SEQUENTIAL STIMULUS

For an optimal network that satisfies the necessary condition of dynamic stability, the neural activity evolves on a stable manifold rather than settling into a fixed-point attractor. When an input is projected into the network, the internal representation changes continuously over time instead of remaining static. As a result, identical stimuli injected at different times can lead to distinct downstream representations at later time. This continual drift of stimulus representations provides a natural mechanism for encoding the order of sequential inputs, akin to the function of working memory.

Once a block-gain matrix G optimal for Fisher information is found for a given input configuration, a corresponding block-structured connectivity matrix J can be readily constructed. Specifically, we sample the elements of J from zero-mean Gaussians with variances derived from the corresponding entries of G. This simple procedure initializes a recurrent network to operate near the edge of chaos, providing a principled starting point for training.

To test the principles of our initialization and its relevance for sequential stimulus processing, we tested the method on two benchmarks on simple RNNs: the copy task and sequential MNIST. Our hypothesis is that initializing a network to operate near the edge of chaos facilitates more efficient training.

Copy task The copy task follows the standard setup of Graves et al. (2014), Arjovsky et al. (2016), and Gu et al. (2021). Each input sequence has length $T_{\rm delay} + 20$. The first 10 tokens are random one-hot vectors in categories $\{1,\ldots,8\}$, followed by $T_{\rm delay}$ zeros, a single delimiter token (category 9), and finally 9 more zeros. The target output has the same length but remains zero until the final 10 steps, where it reproduces the initial random sequence (Fig. 5). This task probes a network's ability to encode categorical information and maintain it in memory for $T_{\rm delay}$ time steps before recall.

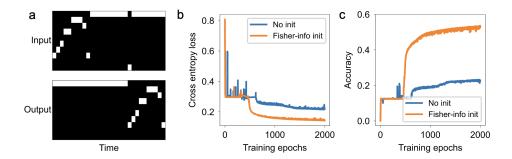


Figure 5: Sequential memory test on the copy task. (a) Illustration of the copy task with $T_{\rm wait}=10$. (b) Networks initialized with Fisher information–optimized weights converge significantly faster and achieve higher accuracy than those with standard random initialization.

We trained a simple RNN of 100 neurons with a tanh nonlinearity and $T_{\rm delay}=50$. For Fisher-information-optimized initialization, the network is partitioned into 10 equal subpopulations (matching the input dimension), with purely feedforward connections $G_{m+1,m}=1$ ($m=1,\ldots,9$) and a single feedback link $G_{1,10}=1$. The corresponding Fisher diffusion matrix A thus has spectral radius one. Weights from input to RNN are sent such that input stimulus are passed to the first subpopulation. Compared to standard random initialization, the Fisher-optimized initialization yields substantially faster convergence and higher final accuracy under otherwise identical training conditions (learning rate $=10^{-3}$; Fig. 5).

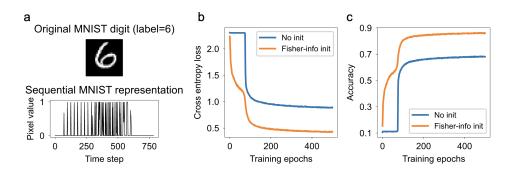


Figure 6: **Sequential memory on Sequential MNIST.** (a) Pixel values of each MNIST image are presented as a 784-step input sequence. (b) Networks initialized with Fisher information—optimized weights converge substantially faster, and (c) achieve higher final accuracy compared to networks with standard random initialization.

Sequential MNIST In the sequential MNIST task, the 784 pixels of each image are fed one by one into the RNN, and the network must classify the digit using only the final hidden state (Fig. 6). To emphasize the advantage of operating at the edge of chaos, we fixed the recurrent weights to the same circular subpopulation structure used in the copy task and trained only the readout layer. In practice, using a one-layer MLP with tanh nonlinearity as the readout significantly improved performance. As with the copy task, Fisher-information-optimized initialization led to both faster convergence and higher final accuracy than random initialization (Fig. 6).

Together, these findings show that initializing the network with Fisher-information-optimized weights—i.e., operating at the edge of chaos—naturally creates a stable manifold along which input representations can evolve. This initialization effectively equips the recurrent network with an intrinsic encoder that both preserves the geometry of the input stimulus and supports flexible movement of representations. As a result, training can focus on learning an appropriate decoder, leading to faster convergence and higher final accuracy (Figs. 5, 6).

6 DISCUSSION

Our contributions are threefold: (i) We introduced a block-structured, mean-field framework in which the Fisher diffusion operator analytically tracks how information flows across interacting subpopulations in recurrent networks. (ii) Criticality (spectral radius ≈ 1) is necessary but not sufficient for long-term retention, and alignment between input structure and the stable subspace is equally essential. (iii) From these principles we derived simple, Fisher-information-optimized initializations that (empirically) accelerate training and improve accuracy on sequential memory tasks.

We bridge multiple perspectives—Fisher information, geometry preservation, and dynamical stability—under a single operator formulation. This connection explains why preserving local geometry, maintaining stability at criticality, and ensuring Fisher information flow are mathematically equivalent conditions. Importantly, the block-structured formulation extends classical one-population mean-field theory to arbitrary modular architectures, making it possible to study realistic networks with feedback, skip connections, and heterogeneous subpopulations.

Fisher information optimized initialization offers theory driven design rules for recurrent networks: initialize near criticality, align inputs with stable diffusion modes, and regulate feedback to balance flux. These rules are architecture-agnostic, compatible with modern RNNs and state-space models, and can be incorporated as regularizer that explicitly encourage balanced Fisher flow. Notably, initialization schemes in RNN and SSMs have been shown to critically influence performance (e.g. Gu et al. (2022) Keller et al. (2024)), analogous in spirit to our approach.

The dynamic, geometry-preserving memory described by our framework provides a principled alternative to classical attractor models. Instead of storing fixed points, recurrent networks maintain evolving trajectories that conserve the relative geometry of inputs—consistent with recent neurophysiological observations of stable manifolds and rotational dynamics in cortex (Ritter & Chadwick, 2025a). Operating at the subpopulation level makes it naturally suited to multi-area circuits, offering predictions for how inter-areal connectivity supports information retention and traveling-wave—like activity patterns. Relatedly, recent theoretical work shows that hidden traveling waves in trained RNNs can bind working memory variables to wave-like representations (Karuvally et al., 2024), suggesting that our framework for information dynamics can also provide a foundation for understanding wave-based mechanisms of memory.

Limitations While our framework provides a principled and interpretable theory of information dynamics in recurrent networks, several limitations remain. Most importantly, our analysis is focused on the encoding of information: how network connectivity structures shape the retention and propagation of Fisher information across subpopulations. We do not address the decoding stage, where task-specific outputs are read out from the evolving internal representations. Thus, our framework should not be viewed as a method for directly discovering architectures that maximize task performance, but rather as a way to endow a given architecture with theoretically grounded initializations that improve training efficiency and stability. In this sense, our work is complementary to performance-oriented models such as unitary RNNs, orthogonal networks, and structured state-space models, which achieve superior accuracy on demanding sequence benchmarks. Our goal has not been to compete with such models, but to provide a general theoretical foundation that explains how information is preserved, and to derive simple initialization rules that translate this theory into practice. By doing so, we highlight principles—criticality, alignment, and balanced Fisher flow—that may also inform the design of future high-performance architectures.

REFERENCES

- Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of the 33rd International Conference on Machine Learning (ICML)*, volume 48, pp. 1120–1128. PMLR, 2016. URL https://proceedings.mlr.press/v48/arjovsky16.html.
- Bard Ermentrout et al. Traveling waves encode the recent past and enhance sequence learning.
- Simon Foucart and Holger Rauhut. *An Introduction to Compressive Sensing*. Birkhäuser, Basel, Boston, 2013. ISBN 978-0-8176-4948-7.
 - Surya Ganguli, Dongsung Huh, and Haim Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences*, 105(48):18970–18975, 2008. doi: 10.1073/pnas.0813041105.
 - Alex Graves, Greg Wayne, and Ivo Danihelka. Neural turing machines. *arXiv preprint*, (arXiv:1410.5401), 2014. URL https://arxiv.org/abs/1410.5401.
 - Albert Gu, Ankit Gupta, Karan Goel, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *NeurIPS*, 2022. Shows how initialization of diagonal SSMs is critical for performance and recovers S4-like behavior.
 - Boyuan Gu et al. Bridging expressivity and scalability with adaptive unitary ssms.
 - Yameng Gu, Lucas E. Sainburg, Sizhe Kuang, Feng Han, Jack W. Williams, Yikang Liu, Nanyin Zhang, Xiang Zhang, David A. Leopold, and Xiao Liu. Brain activity fluctuations propagate as waves traversing the cortical hierarchy. *Cerebral Cortex*, 31(9):3986–4005, 2021. doi: 10.1093/cercor/bhab062. URL https://doi.org/10.1093/cercor/bhab062.
 - Li Jing, Zikai Shen, Wei Peng, Guangyu Chen, Yunhao Liu, and Bo Song. Tunable efficient unitary neural networks (eunn) and their application to rnns. *Proceedings of the 34th International Conference on Machine Learning*, pp. 1733–1741, 2017. arXiv:1706.02761.
 - Jonathan Kadmon and Haim Sompolinsky. Transition to chaos in random neuronal networks. *Physical Review X*, 5(4):041030, 2015. doi: 10.1103/PhysRevX.5.041030.
 - Arjun Karuvally, Terrence J. Sejnowski, and Hava T. Siegelmann. Hidden traveling waves bind working memory variables in recurrent neural networks. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 23266–23290, 2024. URL https://proceedings.mlr.press/v235/karuvally24a.html.
 - T. Anderson Keller, Lyle Muller, Terrence J. Sejnowski, and Max Welling. Traveling waves encode the recent past and enhance sequence learning. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2309.08045. Preprint arXiv:2309.08045.
 - Leo Kozachkov, Michaela Ennis, and Jean-Jacques Slotine. Rnns of rnns: Recursive construction of stable assemblies of recurrent neural networks. In *NeurIPS 2020*, pp. –.
 - Francesca Mastrogiuseppe et al. Input-driven circuit reconfiguration in critical recurrent neural networks.
 - Laura Ritter and Angus Chadwick. Efficient working memory maintenance via high-dimensional rotational dynamics. *bioRxiv preprint*, 2025a. doi: 10.1101/2025.09.08.674838. URL https://www.biorxiv.org/content/10.1101/2025.09.08.674838v1. Preprint: bioRxiv, version 1, September 12, 2025.
- Laura Ritter and Angus Chadwick. Efficient working memory maintenance via high-dimensional rotational dynamics. *bioRxiv preprint*, 2025b. doi: 10.1101/2025.09.08.674838. URL https://www.biorxiv.org/content/10.1101/2025.09.08.674838v1. Preprint: bioRxiv, version 1, September 12, 2025.

Haim Sompolinsky, Andrea Crisanti, and H. J. Sommers. Chaos in random neural networks. *Physical Review Letters*, 61(3):259–262, 1988. doi: 10.1103/PhysRevLett.61.259.

T. Toyoizumi and L. F. Abbott. Beyond the edge of chaos: Amplification and temporal integration by recurrent networks in the chaotic regime. *Physical Review E*, 84(5):051908, 2011. doi: 10.1103/PhysRevE.84.051908. URL https://doi.org/10.1103/PhysRevE.84.051908.

A APPENDIX

A.1 MEMORY CAPACITY VIA FISHER INFORMATION

In this Appendix, we quantify a recurrent network's ability to retain past inputs by computing the Fisher information of its activity with respect to an input parameter. Intuitively, Fisher information measures how small changes in the input θ affect the distribution of network states; larger Fisher information indicates higher sensitivity and thus greater memory capacity.

A.1.1

Fisher Information and KL Divergence Let $\mathbf{h}(t) \in \mathbb{R}^n$ denote the network's vector of neural activities at time t, driven by an input scalar θ . We compare the distributions

$$p(\mathbf{h} \mid \theta)$$
 and $p(\mathbf{h} \mid \theta + \delta \theta)$ (6)

using the Kullback-Leibler (KL) divergence:

$$D_{\mathrm{KL}}(p(\mathbf{h} \mid \theta) \parallel p(\mathbf{h} \mid \theta + \delta\theta)) = \int p(\mathbf{h} \mid \theta) \log \frac{p(\mathbf{h} \mid \theta)}{p(\mathbf{h} \mid \theta + \delta\theta)} d\mathbf{h}.$$
 (7)

A second-order Taylor expansion of $\log p(\mathbf{h} \mid \theta + \delta \theta)$ shows that the leading term in the KL divergence is

$$D_{\mathrm{KL}} \approx -\frac{1}{2} \, \delta \theta^{2} \, \underbrace{\int p(\mathbf{h} \mid \theta) \left[\frac{\partial^{2}}{\partial \theta^{2}} \log p(\mathbf{h} \mid \theta) \right] d\mathbf{h},}_{\mathcal{I}(\theta, t)} \tag{8}$$

where the first-order term vanishes because

$$\int p \,\partial_{\theta} \log p = \partial_{\theta} \int p = 0 \,. \tag{9}$$

We therefore identify the Fisher information at time t as

$$\mathcal{I}(\theta, t) = \mathbb{E}_{p(\mathbf{h}|\theta)} \left[\partial_{\theta}^{2} \log p(\mathbf{h} \mid \theta) \right]. \tag{10}$$

Larger \mathcal{I} implies that small input perturbations produce more distinguishable changes in the network state.

A.1.2 GAUSSIAN APPROXIMATION

To obtain a closed-form expression, we invoke the Maximum Entropy (?) principle and assume that, at each t, the network's activity conditional on θ is Gaussian:

$$p(\mathbf{h} \mid \theta) = \frac{1}{(2\pi)^{n/2} \sqrt{\det \Sigma(t)}} \exp\left[-\frac{1}{2} \left(\mathbf{h} - \boldsymbol{\mu}(\theta, t)\right)^{\top} \Sigma(t)^{-1} \left(\mathbf{h} - \boldsymbol{\mu}(\theta, t)\right)\right]. \tag{11}$$

Here:

- μ(θ, t) is the mean activity, which we assume depends smoothly (approximately linearly) on θ.
- Σ(t) is the covariance matrix, which we treat as input-independent and determined by the network's intrinsic dynamics.

Under these assumptions, one finds the Fisher Memory Curve Ganguli et al. (2008):

$$\mathcal{I}(\theta, t) = \left\langle \frac{\partial}{\partial \theta} \frac{\partial}{\partial \theta} \log p(\mathbf{h}(t)|\theta) \right\rangle_{p(\mathbf{h}(t)|\theta)} = \left\langle \left(\frac{\partial \boldsymbol{\mu}(t)}{\partial \theta}\right)^T \Sigma^{-1}(t) \frac{\partial \boldsymbol{\mu}(t)}{\partial \theta} \right\rangle_{p(\mathbf{h}(t)|\theta)}. \tag{12}$$

so that memory capacity is governed by how sensitive of the the mean activity shifts with input (the "signal") weighted by the inverse noise covariance.

A.2 ANALYTIC FISHER INFORMATION FOR MULTIPLE SUB-POPULATIONS

We consider a recurrent network divided into multiple subpopulations. For each sub-population m, let

$$\mu_m(t) = \langle h_m(t) | J \rangle. \tag{13}$$

denote the mean activity at time t, averaged over dynamic noise but at fixed synaptic matrix J. Our goal is to compute the Fisher information $\mathcal{I}_m(t)$ of $\mu_m(t)$ with respect to an input parameter θ :

$$\mathcal{I}(\theta, t) = K \sum_{m} \frac{f_m}{q_m} \left\langle \left(\frac{\partial \mu_m(t)}{\partial \theta} \right)^2 \right\rangle_J. \tag{14}$$

The key is to derive the analytic formula for $\left\langle \left(\frac{\partial \mu_m(t)}{\partial \theta} \right)^2 \right\rangle_J$.

A.2.1 REPLICA TRICK FOR THE DERIVATIVE OF THE MEAN

Introducing replicas a, b, we have

$$\left\langle \left(\frac{\partial \mu_m(t)}{\partial \theta} \right)^2 \right\rangle_J = \left\langle \left[\frac{\partial}{\partial \theta} \langle h_m(t) | J \rangle \right]^2 \right\rangle_J = \left\langle \frac{\partial}{\partial \theta^a} \langle h_m^a(t) | J \rangle \frac{\partial}{\partial \theta^b} \langle h_m^b(t) | J \rangle \right\rangle_J
= \frac{\partial^2}{\partial \theta^a \partial \theta^b} \langle h_m^a(t) h_m^b(t) \rangle = \frac{\partial^2}{\partial \theta^a \partial \theta^b} q_m^{ab}(t) ,$$
(15)

where $q_m^{ab}(t) \equiv \langle h_m^a(t) h_m^b(t) \rangle_m$.

A.2.2 MEAN-FIELD EXPRESSION FOR THE CORRELATION

In mean-field, the correlation splits into an i.i.d. noise term and a term generated by recurrent inputs:

$$q_m^{ab}(t,s) = \sigma^2 \delta_{ab} \delta_{ts} + \sum_n g_{mn}^2 f_n \langle S_n^a(t) S_n^b(s) \rangle_n = \sigma^2 + \sum_n G_{mn} C_n^{ab}(t,s) , \qquad (16)$$

with $G_{mn}\equiv g_{mn}^2f_n$ and $C_n^{ab}(t,s)=\langle S_n^a(t)S_n^b(s)\rangle$, the firing rate correlation in sub-population n.

A.2.3 DIFFERENTIATING C_n^{ab}

Applying $\partial_{\theta}^{a}\partial_{\theta}^{b}$ to $C_{n}^{ab}(t,s)$ and using the chain rule for the nonlinearity $\phi(\,\cdot\,)$ (with $\phi'=\frac{d\phi}{dz}$) yields

$$\partial^{ab}C_{n}^{ab}(t) = \partial^{ab}\langle\phi^{a}(t)\phi^{b}(t)\rangle_{n}
= \partial^{ab}\langle\phi(w_{n}\Theta^{a}(t-1) + x^{a}(t-1))\phi(w_{n}\Theta^{b}(t-1) + x^{b}(t-1))\rangle_{n}
= \langle\phi'^{a}\cdot(w_{n} + \partial_{a}x^{a}(t-1))\phi'^{b}\cdot(w_{n} + \partial_{b}x^{b}(t-1))\rangle_{n}
= \langle\phi'^{a}\phi'^{b}\rangle_{n}\langle(w_{n}^{2} + (\partial_{a}x^{a}(t-1) + \partial_{b}x^{b}(t-1))w_{n} + \partial_{a}x^{a}(t-1)\partial_{b}x^{b}(t-1)\rangle_{n}
= \langle\phi'^{a}\phi'^{b}\rangle_{n}(w_{n}^{2} + \partial_{a}\partial_{b}\langle x^{a}(t-1)x^{b}(t-1)\rangle_{n}
= \langle\phi'^{a}\phi'^{b}\rangle_{n}(w_{n}^{2} + \partial_{a}\partial_{b}q_{n}^{ab}(t-1)).$$
(17)

The first order in line 4 with terms $\langle \partial_a x^a \rangle_n = \partial_a \langle x^a \rangle_n = \partial_a 0 = 0$.

A.2.4 RECURRENCE FOR THE SECOND DERIVATIVE OF a

Combining equation 16 and equation 17 produces a linear recurrence:

$$\partial^{a}\partial^{b}q_{m}^{ab}(t+1) = \sum_{n} G_{mn} \langle \phi'^{a}\phi'^{b} \rangle_{n} (\partial^{a}\partial^{b}q_{n}^{ab}(t) + w_{n}^{2}\delta_{t,0}), \qquad A_{mn} \equiv G_{mn} \langle \phi'^{a}\phi'^{b} \rangle_{n}.$$
(18)

Because q^{ab} depends only on earlier inputs, the initial condition is $\partial_{\theta}^{a} \partial_{\theta}^{b} q_{n}^{ab}(t) = 0$ for $t \leq 0$. Iterating equation 18 once at t = 0 gives

$$\partial_{\theta}^{a} \partial_{\theta}^{b} q_{m}^{ab}(1) = \sum_{n=1}^{M} A_{mn} w_{n}^{2}, \tag{19}$$

Since the network only receive the input at t=0, each successive iteration amounts to matrix multiplication by $A_{mn} \equiv G_{mn} \langle \phi'^a \phi'^b \rangle_n$. Repeating the recurrence t times results:

$$\partial^a \partial^b q_{m,t+1}^{ab} = \sum_n (\mathbf{A}^{t+1})_{mn} w_n^2 . {20}$$

A.2.5 CLOSED-FORM FISHER INFORMATION

Finally, substituting Eq. (20) into equation 15 gives

$$\left\langle \left(\frac{\partial \mu_m(t)}{\partial \theta} \right)^2 \right\rangle_J = \sum_n (\mathbf{A}^t)_{mn} w_n^2, \quad A_{mn} \equiv G_{mn} \left\langle (\phi')^2 \right\rangle_n.$$
 (21)

Eq. equation 21 shows that the propagation of the Fisher information through the network can be effectively captured by the Fisher information diffusion operator A.

A.2.6 Analytic derivation of the Fisher information diffusion operator

To obtain an analytical expression for the Fisher information in relation to the optimal connectivity parameters G_{mn} and to gain an intuitive understanding of Fisher information from a network perspective, two key tasks are essential:

- 1. Analytically resolve the self-consistent equations for q_1 and q_2 to understand the dynamics in the system. These solutions also allow us to construct Gaussian probability distributions with variances q_1 and q_2 , respectively. From these distributions, we compute the second-order moments $\langle S^2 \rangle$ and $\langle (S')^2 \rangle$, which are essential for determining the Fisher information diffusion operator.
- 2. Derive an analytical formula for Fisher information that elucidates the relationship between network characteristics, the connection between populations of neurons, and optimal information capacity.

A.2.7 ANALYTIC CALCULATION OF THE $\langle (\tanh(x))^2 \rangle$

Since both the self consistent equations

$$q_{1} = \sigma_{1}^{2} = \sigma^{2} + G_{11}\langle S^{2}\rangle_{1} + G_{12}\langle S^{2}\rangle_{2} ,$$

$$q_{2} = \sigma_{2}^{2} = \sigma^{2} + G_{22}\langle S^{2}\rangle_{2} + G_{21}\langle S^{2}\rangle_{1} .$$
(22)

and the Fisher information diffusion operator $A_{mn}=G_{mn}\langle(S')^2\rangle\rangle_n$. With $S=\tanh(x)$, S^2 and S' are highly nonlinear and non local, the values are not close to 0 or 1. As a result, using Taylor expansion of the $\tanh(x)$ produces both poor approximation and analytic challenge when calculating the Gaussian average $\langle S^2\rangle$ and $\langle (S')^2\rangle$. We notice that both expressions S^2 and S' only relate to the some form of Gaussian average $\langle \tanh^2(x)\rangle$, and we can approximate the $\tanh^2(x)$ with $1-\exp\frac{x^2}{2\epsilon^2}$ (See Fig. S1). Note that this expression insures $(1-\exp\frac{x^2}{2\epsilon^2})|_{x=0}=0$. The optimal parameter $\epsilon=0.7784$ can be derived from the minimizing the integral difference $\int_{-\infty}^{\infty}|1-\exp\frac{x^2}{2\epsilon^2}-\tanh^2(x)|$. The Gaussian average $\langle f(x)\rangle_{\sigma}=\int_{-\infty}^{\infty}f(x)N(0,\sigma^2)$ can be calculated easily with a simple form:

$$\langle \tanh^2(x) \rangle_i = 1 - \frac{1}{\sqrt{1 + (\frac{\sigma_i}{\epsilon})^2}} = 1 - \frac{1}{\sqrt{1 + \mu_i^2}}, \mu_i = \frac{\sigma_i}{\epsilon} ,$$

$$\langle \tanh'(x)^2 \rangle_i = \langle (1 - \tanh^2(x))^2 \rangle_i = \frac{1}{\sqrt{1 + 2\mu_i^2}}.$$
(23)

For numeric solutions, Eq. (equation 22) (equation 27) extends naturally to an arbitrary number of sub-populations, yield a closed system of nonlinear equations for the variance q_i . This system can be efficiently solved using the 'fsolve' function from the 'scipy.optimize' package. Once the q_i are obtained, they are substituted into Eq. (equation 27) to evaluate the Fisher-information diffusion operator.

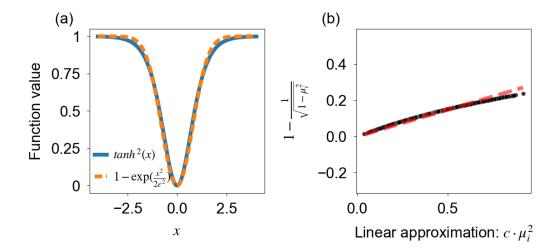


Figure S1: (a) Approximation of the nonlinear function $\tanh^2(x)$ using the surrogate form $f(x,\epsilon)=1-\exp(-x^2/2\epsilon^2)$, with the optimal parameter $\epsilon=0.7784$. (b) Linear approximation of the expression $1-\frac{1}{\sqrt{1+\mu_i^2}}$ using $c\cdot\mu_i^2$, where the optimal slope c=0.2948 is determined via least-squares fitting. Each point μ_i is obtained from a grid search over network configurations in the two-population case, constrained to operate near the edge of chaos $(|\rho(A)-1|<0.1)$, where $\rho(A)$ denotes the spectral radius of the effective connectivity matrix.

A.2.8 Analytic calculation of the order parameters q_1 and q_2

With Eq. (equation 23), we can rewrite the self consistent equations:

$$\mu_1^2 = \mu^2 + M_{11} \left(1 - \frac{1}{\sqrt{1 + \mu_1^2}}\right) + M_{12} \left(1 - \frac{1}{\sqrt{1 + \mu_2^2}}\right),$$

$$\mu_2^2 = \mu^2 + M_{21} \left(1 - \frac{1}{\sqrt{1 + \mu_1^2}}\right) + M_{22} \left(1 - \frac{1}{\sqrt{1 + \mu_2^2}}\right),$$

$$M_{mn} = G_{mn} / \epsilon^2, \ \mu_i^2 = \sigma_i^2 / \epsilon^2, \ \mu^2 = \sigma^2 / \epsilon^2.$$
(24)

Solving the Eq. (equation 24) directly is difficult and will lead to unintuitive expression since this is a system of cubic equations with non uniform power in each term. For systems operating near the edge of chaos—characterized by a spectral radius close to one ($|\rho(A)-1|<0.1$)—the variable μ_i remains small (Fig S1). In this regime, we can approximate the nonlinear expression $1-\frac{1}{\sqrt{1+\mu_i^2}}$ using a linearized form. Specifically, we use a least-squares fit to determine the optimal slope c in the following approximation:

$$1 - \frac{1}{\sqrt{1 + \mu_i^2}} \approx c\mu_i^2, \quad i \in \{1, 2\}, \quad c = 0.2948.$$
 (25)

This approximation simplifies further analysis while preserving accuracy in the small- μ_i limit.

With the linear approximation Eq. (equation 25), the solutions to the self consistent equations are:

$$\mu_1^2 \approx \frac{\mu^2 + c\mu^2(M_{12} - M_{22})}{1 - c\operatorname{Tr}(M) + c^2\det(M)},$$

$$\mu_2^2 \approx \frac{\mu^2 + c\mu^2(M_{21} - M_{11})}{1 - c\operatorname{Tr}(M) + c^2\det(M)},$$

$$\det(M) = M_{11}M_{22} - M_{12}M_{21}, \operatorname{Tr}(M) = M_{11} + M_{22}.$$
(26)

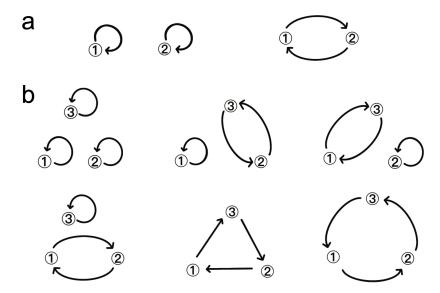


Figure S2: A illustration of the determinant of the weight matrix for 2 populations(a) and 3 populations (b).

Under the meanfield, subpopulation 1 and subpopulation 2 are considered as nodes with weight matrix as $M_{mn} = G_{mn}/\epsilon^2$. This mapping makes every term in the analytic expression of Eq. (equation 26) a familiar graph invariant. The trace, $\text{Tr}(M) = \sum_m M_{mm}$, equals the total weight of all self-loops in the network ??. Meanwhile, the determinant, $\det(M)$, via the Leibniz expansion, becomes a signed sum over all cycle covers (loop configurations), each monomial corresponding to a distinct set of loops weighted by the product of edge weights ??. These loop configurations are illustrated in Fig. S2.

A.2.9 Analytic expression of the Fisher information diffusion operator A

The Fisher information diffusion operator $A_{mn}=G_{mn}\langle(S')^2\rangle_n=\epsilon^2 M_{mn}\langle(S')^2\rangle_n$. In Eq. (equation 26), we have derived analytic expression in terms of the scaled conductivities $M_{mn}=G_{mn}/\epsilon^2$ and plugging in Eq. (equation 23). Directly substitute the Eq. (equation 26) into Eq. (equation 23), we get:

$$\langle (S')^{2} \rangle_{1} = \frac{1}{\sqrt{1 + 2\mu_{1}^{2}}} \approx f(M)((1 + c(M_{12} - M_{22}))),$$

$$\langle (S')^{2} \rangle_{2} = \frac{1}{\sqrt{1 + 2\mu_{2}^{2}}} \approx f(M)((1 + c(M_{21} - M_{11}))),$$

$$f(M) = \frac{\mu^{2}}{1 - c\operatorname{Tr}(M) + c^{2}\operatorname{det}(M)},$$

$$\tilde{f}(M) = 1 - 2c\frac{\mu^{2}}{1 - c\operatorname{Tr}(M) + c^{2}\operatorname{det}(M)} = 1 - 2cf(M),$$

$$c = 0.2948.$$

$$(27)$$

$$A = \epsilon^2 \begin{pmatrix} M_{11} \langle (S')^2 \rangle_1 & M_{12} \langle (S')^2 \rangle_2 \\ M_{21} \langle (S')^2 \rangle_1 & M_{22} \langle (S')^2 \rangle_2 \end{pmatrix} = \epsilon^2 \begin{pmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{pmatrix} \begin{pmatrix} \langle (S')^2 \rangle_1 & 0 \\ 0 & \langle (S')^2 \rangle_2 \end{pmatrix} , \quad (28)$$

$$\epsilon = 0.7784 .$$

At criticality—i.e. on the "edge of chaos"—the Fisher-information diffusion operator A acquires an eigenvalue exactly equal to unity. Equivalently:

$$\det(I - A_{\text{opt}}) = 0. (29)$$

By expanding det(I - A) for our two-population system and grouping terms, we obtain the expression for the condition for the edge of chaos in a fully symmetric form with respect to subpopulations:

$$0 = \det(I - A_{opt})$$

$$= 1 - \epsilon^{2} (\langle (S')^{2} \rangle_{1} M_{11} + \langle (S')^{2} \rangle_{2} M_{22}) + \epsilon^{4} \langle (S')^{2} \rangle_{1} \langle (S')^{2} \rangle_{2} \det(M)$$

$$= 1 - \epsilon^{2} L_{1}(M) + \epsilon^{4} \det(M) L_{2}(M) ,$$

$$L_{1}(M) = [\tilde{f}(M) - 2c^{2} f(M) \operatorname{Tr}_{off}(M)] \operatorname{Tr}(M) + 2c^{2} f(M)[M_{11}d_{2} + M_{22}d_{1}] ,$$

$$L_{2}(M) = \tilde{f}^{2}(M) + 2c^{2} f(M)[\operatorname{Tr}(M) - \operatorname{Tr}_{off}(M)]$$

$$+ 4c^{4} f^{2}(M)(d_{1}d_{2} - \operatorname{Tr}(M) \operatorname{Tr}_{off}(M)) ,$$

$$f(M) = \frac{\mu^{2}}{1 - c \operatorname{Tr}(M) + c^{2} \det(M)} , \quad \tilde{f}(M) = 1 - 2c f(M) ,$$

$$d_{i} = \sum_{k} M_{ik} = M_{i1} + M_{i2}, \quad \operatorname{Tr}_{off}(M) = M_{12} + M_{21}, ,$$

$$c = 0.2948, \quad \epsilon = 0.7784 .$$

$$(30)$$

where we recognize:

- 1. **Trace**, $Tr(M) = M_{11} + M_{22}$. The *total self-loop weight* (sum of length-1 cycles), which sets first-order feedback gain.
- 2. Off-diagonal trace, $\mathbf{Tr}_{off}(M) = M_{12} + M_{21}$. The *total cross-population coupling*, measuring the strength of two-node interactions ?.
- 3. **Determinant,** $det(M) = M_{11}M_{22} M_{12}M_{21}$. A signed sum over all 2-cycle covers:
 - $M_{11}M_{22}$ counts two independent self-loops,
 - $M_{12}M_{21}$ counts the reciprocal 2-node cycle.

The determinant provides insights into the connectivity and spanning trees of a graph, as detailed in the Matrix-Tree Theorem ?.

4. Weighted in-degrees.

$$d_i = \sum_{k} M_{ik} = M_{i1} + M_{i2},$$

the total incoming weight to subpopulation i. The concept of in-degree is a basic measure in graph theory, indicating the number of edges arriving at a node ?.

Here, through algebraic manipulation and careful rearrangement, we derive a form of the edge-of-chaos condition that is symmetric across subpopulations and expressed entirely in terms of familiar graph-theoretic quantities—such as trace, off-diagonal trace, determinant, and in-degree of the connectivity matrix. This reformulation reveals how the topology of structured neural networks directly shapes the onset of criticality.

A.3 NUMERIC CALCULATION OF FISHER INFORMATION FROM MONTE-CARLO

In the main text, we benchmark the analytic expression of the Fisher information against a direct Monte-Carlo estimate obtained from explicit simulations of the recurrent neural network (RNN). The numerical procedure consists of three main stages: (i) initialization of the random block-structured connectivity, (ii) simulation of neural trajectories under baseline and perturbed inputs, and (iii) estimation of derivatives via symmetric finite differences.

A.3.1 NETWORK INITIALIZATION

The network consists of N neurons partitioned into M subpopulations with sizes $n_m = f_m N$. Synaptic connectivity is represented by a block-structured random matrix $J \in \mathbb{R}^{N \times N}$. Each block J_{mn} is sampled i.i.d. from a Gaussian distribution

$$J_{kl} \sim \mathcal{N}\left(0, \frac{g_{mn}^2}{N}\right), \qquad k \in \text{pop}_m, \ l \in \text{pop}_n,$$

where g_{mn} denotes the population-dependent gain parameter. This choice controls the effective recurrent gain while ensuring that connectivity statistics remain stable as N increases. Here k indexes a postsynaptic neuron in population m and l indexes a presynaptic neuron in population n. This blockwise construction ensures that the recurrent connectivity statistics are determined by the gain matrix g while preserving the correct population sizes.

Dynamical simulation. Neural activity is described by pre-activations $x_t \in \mathbb{R}^N$ and firing rates $S_t = \tanh(x_t)$. The recurrent dynamics evolve according to

$$x_{t+1} = JS_t + \sigma \xi_t, \qquad \xi_t \sim \mathcal{N}(0, I),$$

with additive Gaussian noise of variance σ^2 . At initialization, an external input θ is injected into the first population, implemented by setting $x_{1:n_1} \leftarrow \theta$. Multiple trajectories are simulated in parallel to estimate ensemble averages.

A.3.2 PERTURBATION PROTOCOL

To estimate the Fisher information with respect to the input parameter θ , we simulate network dynamics under three input conditions: baseline θ , positively perturbed $\theta + \Delta \theta$, and negatively perturbed $\theta - \Delta \theta$. For each condition, we record the full trajectory of neural activities $\{\mathbf{h}(t)\}_{t=1}^T$. Throughout the simulations, we set the baseline input to $\theta = 0$.

A.3.3 FISHER INFORMATION ESTIMATION

The sensitivity of mean activity to θ is approximated via symmetric finite differences:

$$\frac{\partial \mu_t}{\partial \theta} \approx \frac{\mu_t(\theta + \Delta \theta) - \mu_t(\theta - \Delta \theta)}{2\Delta \theta},$$

where $\mu_t(\theta)$ is the average firing rate at time t across trajectories. Squaring and averaging these derivatives over neurons within population m yields a time-resolved Fisher information stored in each subpopulation about the input stimulus over time as in Eq equation 4

Here, we also show the MSE between the simulated fisher information and the analytic prediction of the fisher information in Fig. S3

A.4 TEST OF FISHER INFORMATION WITH NATURAL IMAGES AS INPUT

Experimental setup. To evaluate whether Fisher information predicts geometry preservation, we tested the framework using natural inputs. We used 15,619 CIFAR-10 images, each flattened to a 7,500-dimensional vector, and presented them to the first subpopulation of a two-population recurrent network with N=15,000 neurons, $f_1=f_2=0.5$, and noise variance $\sigma=0.1$. Each image was processed independently, and at each time step t we recorded the activities of both subpopulations.

Measuring geometry preservation. Geometry preservation was quantified by comparing pairwise distances between images in the input space to pairwise distances between their corresponding neural representations. Specifically: 1. For the input set, we computed all pairwise Euclidean distances $D_{\text{input}}(i,j) = \|x_i - x_j\|_2$ between the flattened image vectors. 2. For the network representations at time t, we computed analogous pairwise distances $D_{\text{rep}}(i,j) = \|h_i(t) - h_j(t)\|_2$ for each subpopulation. 3. To assess how faithfully the network preserved geometry, we calculated the Pearson correlation coefficient between the upper triangular entries of the two distance matrices,

$$\rho(t) = \operatorname{corr}(\operatorname{vec}(D_{\operatorname{input}}), \operatorname{vec}(D_{\operatorname{rep}}(t))),$$

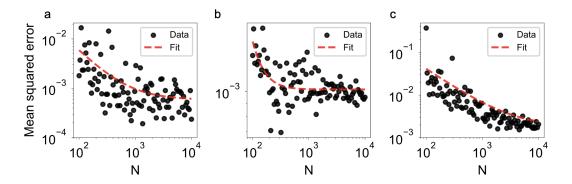


Figure S3: Convergence of simulated Fisher information to analytic predictions. Mean-squared error (MSE) between Fisher information trajectories obtained from simulations and from the analytic diffusion operator, across the three connectivity configurations in Fig. 2a–c. The MSE decreases rapidly with network size and becomes negligible for $N \geq 1000$. An exponential fit (dashed line) is shown to highlight the convergence trend.

where $\rho=1$ indicates perfect isometry (exact geometry preservation) and lower values indicate increasing distortion.

This procedure yields a time series $\rho_m(t)$ for each subpopulation m, quantifying how input geometry is preserved over time as information diffuses through the network.

Interpretation. Although this metric differs from Fisher information, it recovers the same qualitative behavior. In particular, the analytic framework predicts both (i) the oscillatory dynamics of information flow across subpopulations and (ii) the relative ability of different connectivity motifs to preserve the geometry of natural images.

A.5 CONNECTING FISHER INFORMATION WITH RESTRICTED ISOMETRY PROPERTY

Theorem 1 (Optimal network structure and information retention). Consider a recurrent network with block gain matrix G and nonlinear activation ϕ . Under the mean-field approximation, optimal information retention—defined as preservation of local geometry between stimulus representations—is achieved when $G\langle \phi'^2 \rangle = 1$.

This condition coincides with the Fisher information criterion for non-vanishing memory.

Sketch proof. The result connects two perspectives:

- (1) Compressed sensing. For linear systems f(x) = Jx, the Restricted Isometry Property (RIP) ensures approximate distance preservation between sparse vectors: $||f(u) f(v)||^2 \approx ||u v||^2$. For Gaussian $J \sim \mathcal{N}(0, g/\sqrt{N})$, RIP holds when $J^{\top}J \approx I$, requiring average squared gain $g^2 = 1$.
- (2) Nonlinear extension. With nonlinearity ϕ , distances transform as

$$||f(x) - f(x')||^2 \approx ||\phi'(x) J(x - x')||^2.$$

Mean-field theory replaces $\phi'(x)^2$ by its population average $\langle \phi'^2 \rangle$, yielding an effective gain condition

$$G\langle \phi'^2 \rangle \approx 1.$$

(3) Fisher information connection. From the Fisher diffusion framework, long-term non-vanishing memory requires exactly the same balance: largest eigenvalue of $G\langle\phi'^2\rangle$ equals 1. Thus, preserving local geometry (via RIP/Johnson–Lindenstrauss arguments) and preserving Fisher information lead to the same criterion.

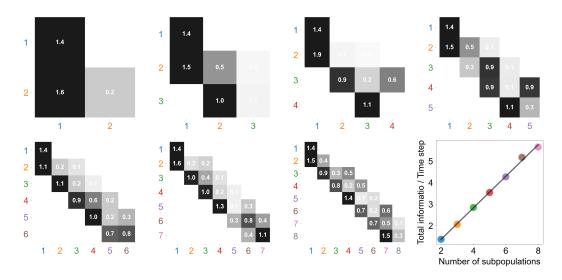


Figure S4: Optimized connectivity matrices G for networks composed of 2 to 8 neuronal subpopulations (insets). Each tick color corresponds to a different number of subpopulations. Matrix element G_{mn} represents the strength of the connection from subpopulation n to subpopulation m. The final panel illustrates that the total Fisher information $\overline{\mathcal{I}}$ scales approximately linearly with the number of subpopulations in the optimally connected network.

A.6 PROOF OF OPTIMAL STRUCTURE FOR FISHER INFORMATION IN A CHAINED LINEAR NETWORK

In the linear limit—when the activation function is purely linear—one has $\langle (S')^2 \rangle_n = 1$ for all sub-populations. The sensitivity block in the Fisher diffusion operator therefore reduces to the identity, so that

$$A_{\text{linear}} = G.$$

Without loss of generality, consider a chain of four subpopulations with unit self-recurrence:

$$A_{\text{linear}} = G = \begin{pmatrix} 1 & G_{12} & 0 & 0 \\ G_{21} & 1 & G_{23} & 0 \\ 0 & G_{32} & 1 & G_{34} \\ 0 & 0 & G_{43} & 1 \end{pmatrix}.$$
(31)

Optimal Fisher information requires that the spectral radius of A equals one, i.e. the largest eigenvalue satisfies $\lambda_{\max} = 1$. Equivalently,

$$0 = \det(A - I) = \det\begin{pmatrix} 0 & G_{12} & 0 & 0 \\ G_{21} & 0 & G_{23} & 0 \\ 0 & G_{32} & 0 & G_{34} \\ 0 & 0 & G_{43} & 0 \end{pmatrix} = G_{12}G_{21}G_{34}G_{43}.$$

For efficient information transmission, the input must enter the first subpopulation and propagate forward through the chain. Hence the forward gains G_{21} and G_{43} cannot vanish. To satisfy $\det(A-I)=0$, we therefore require

$$G_{12} = 0$$
 or $G_{34} = 0$.

This condition eliminates the global feedback loops that would otherwise close the chain, giving rise to the broken-loop optimal structure illustrated in Fig. 4. The exact values for the block gain matrix G for each optimized network is shown in Fig. S4