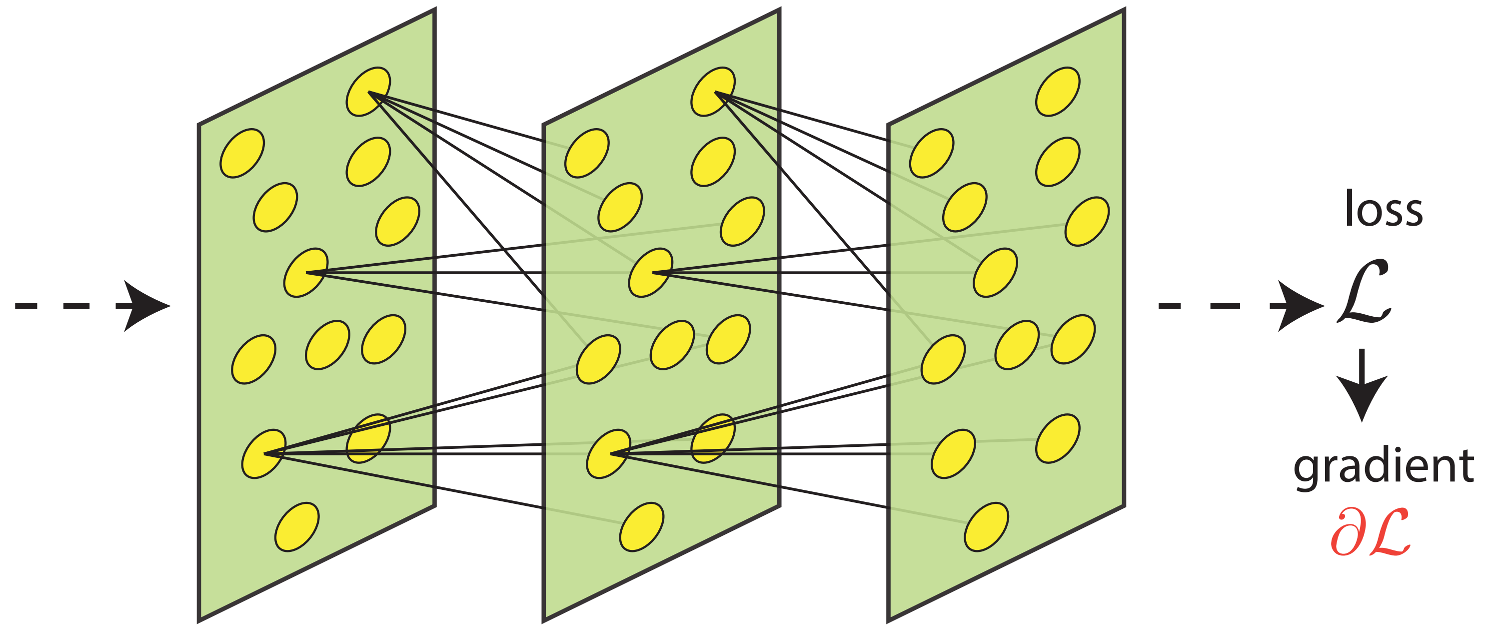
On asymptotic learning signals in recurrent networks

Piotr A. Sokół's
doctoral dissertation
research

Il Memming Park (박일;朴逸)
Group Leader @ Champalimaud Centre for the Unknown
Associate Professor @ Stony Brook University
Computational And Theoretical Neural Information Processing (CATNIP) = Neural Dynamics Lab
https://catniplab.github.io/

Previous contributions to ML theory

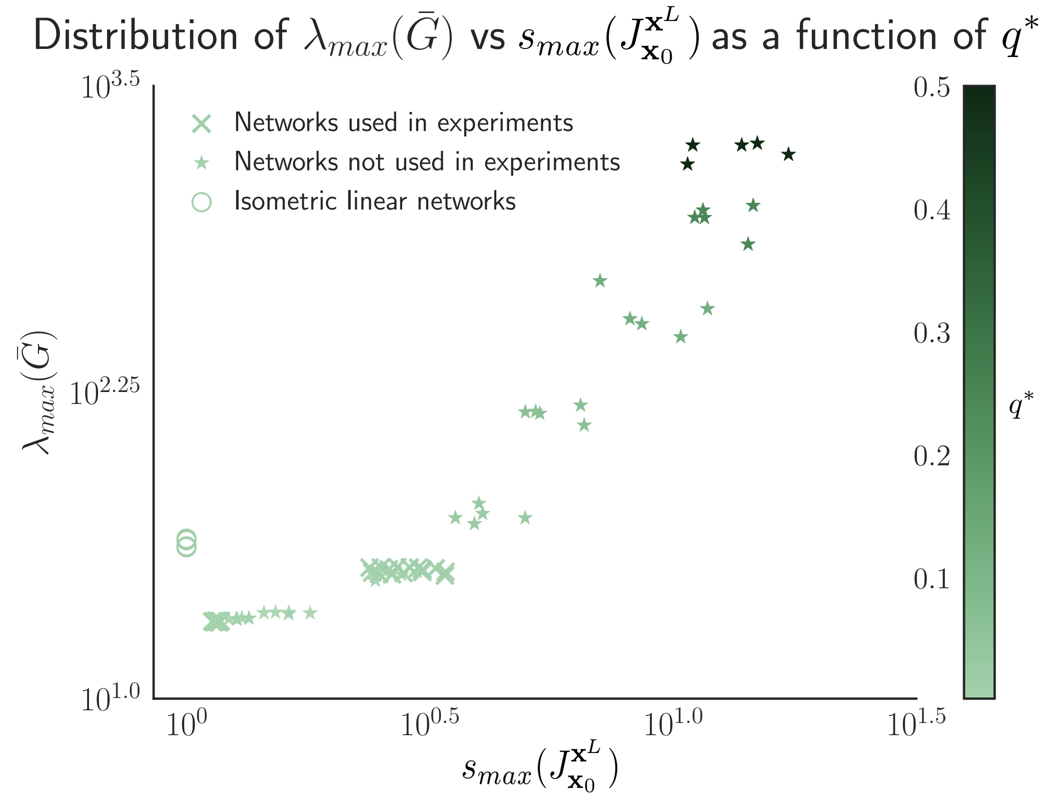
Sokol, P., & Park, I. M. Information geometry of orthogonal initializations and training. ICLR 2020



$$WW^\top = \sigma_w^2 \mathbf{I}$$

scaled random orthogonal weight matrices

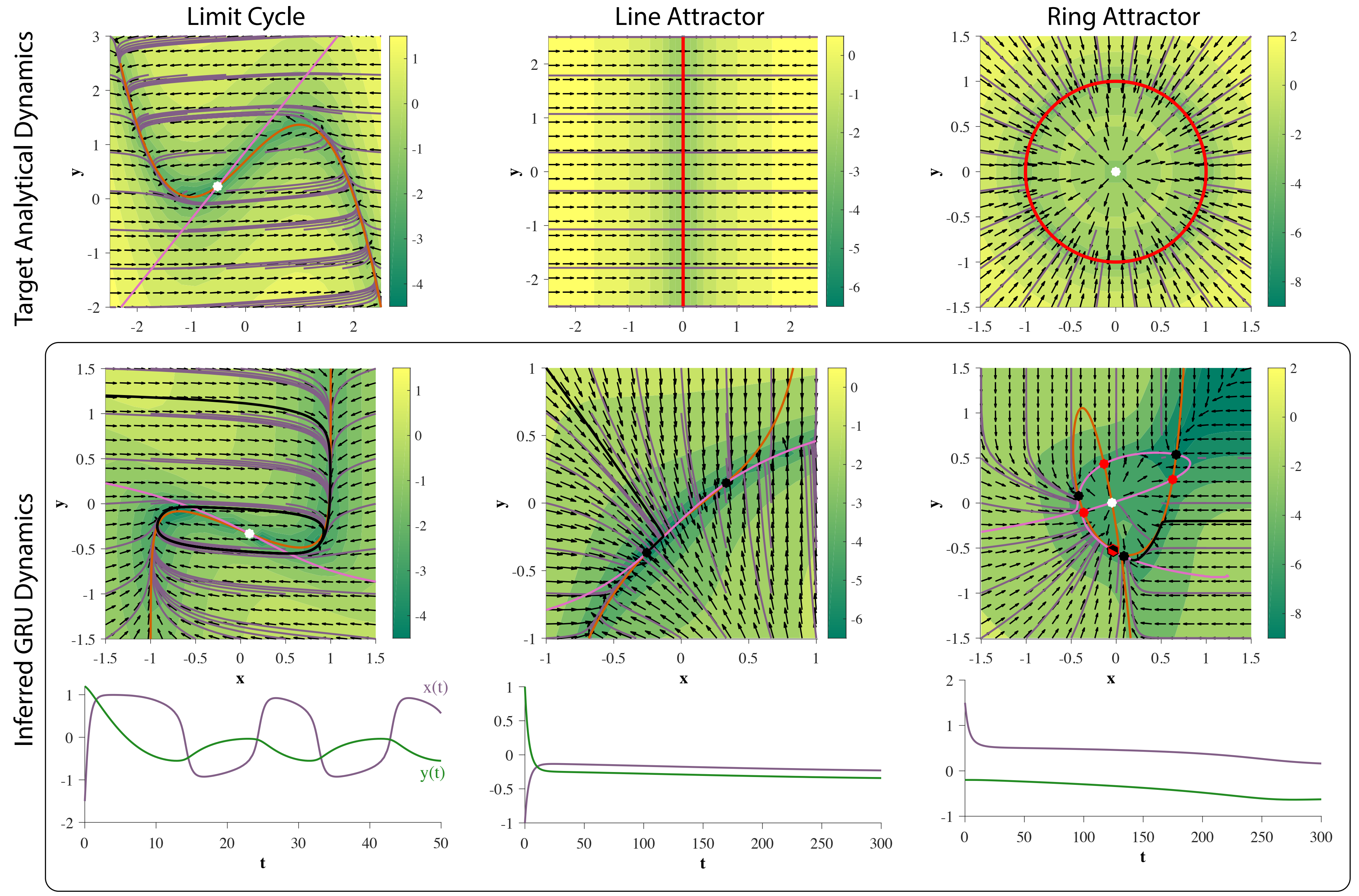
[Pennington *et al.*, NeurIPS 2017]



- Fisher information matrix (FIM) captures the curvature of the gradient. Maximum eigenvalue of FIM tells us how fast the gradient will change directions with gradient descent. (a.k.a. gradient smoothness)
- We bounded the maximum eigenvalue of the FIM by the squared spectral radius of the input-output Jacobian.
- Fast convergence of 200-layer deep networks using manifold constraints.

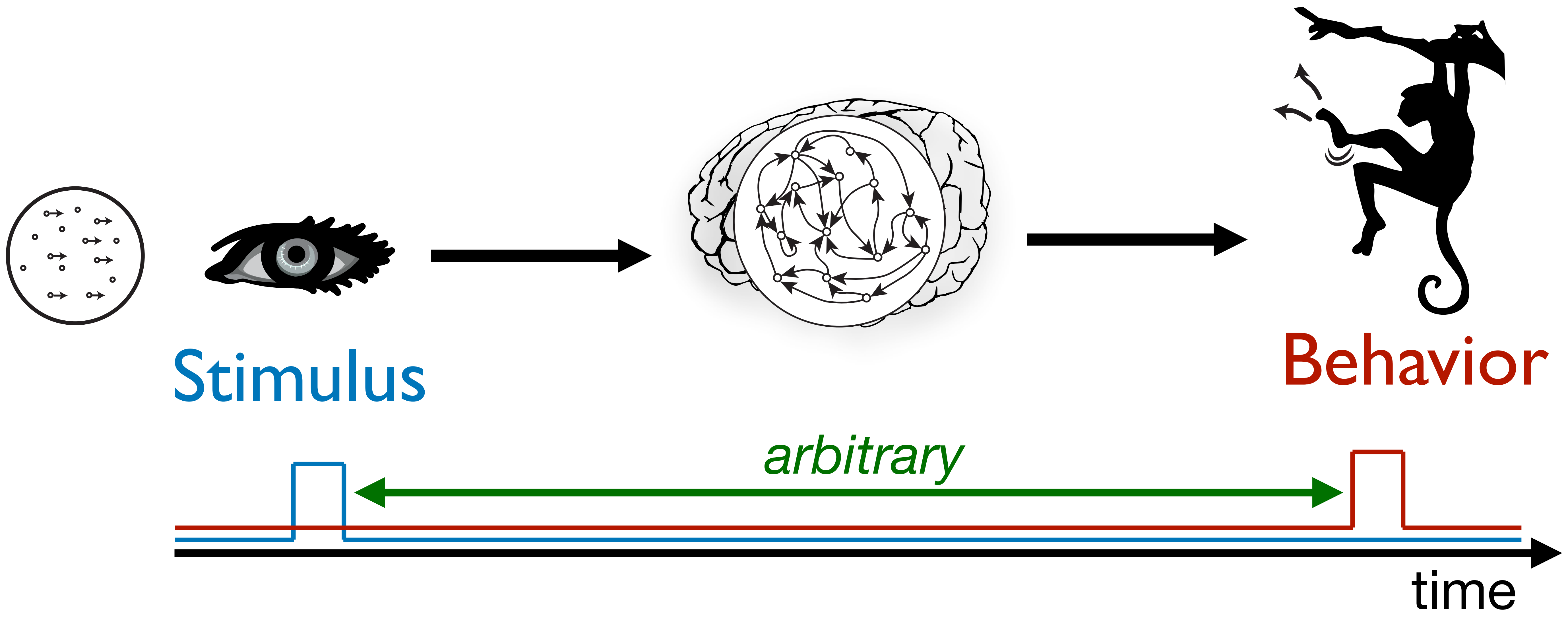
Jordan, I. D., Sokół, P. A., & Park, I. M. (2021). Gated Recurrent Units Viewed Through the Lens of Continuous Time Dynamical Systems. Frontiers in Comp. Neurosci.

- Expressive power of small GRU-RNNs



Arbitrarily long-range temporal dependency

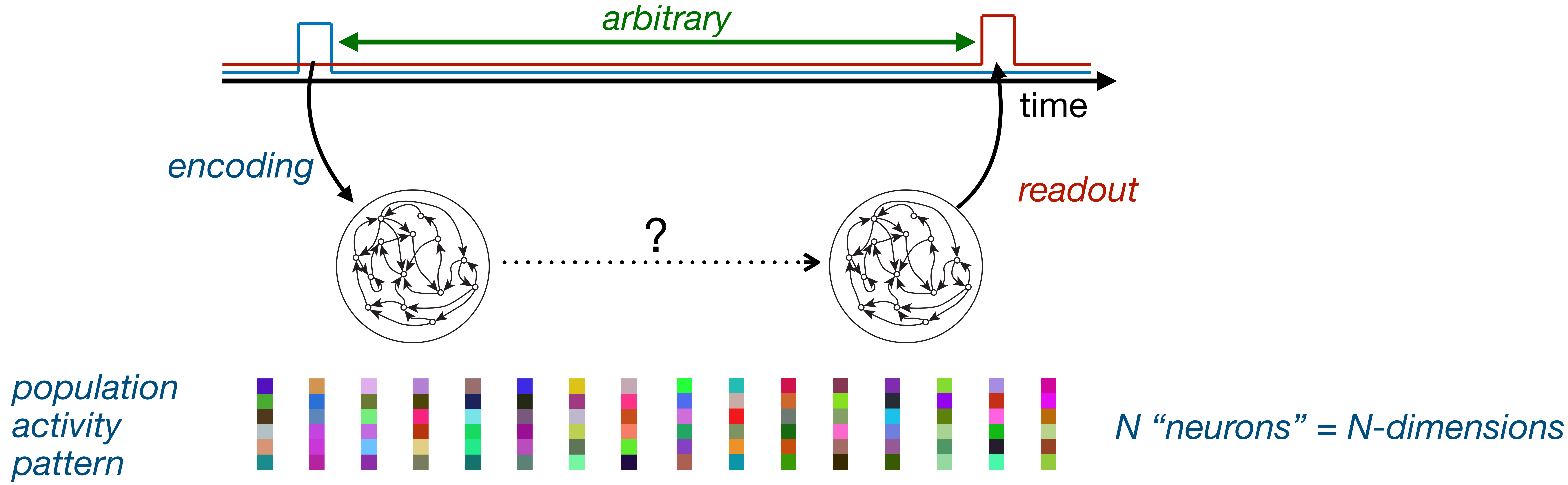
in supervised, unsupervised, reinforcement learning



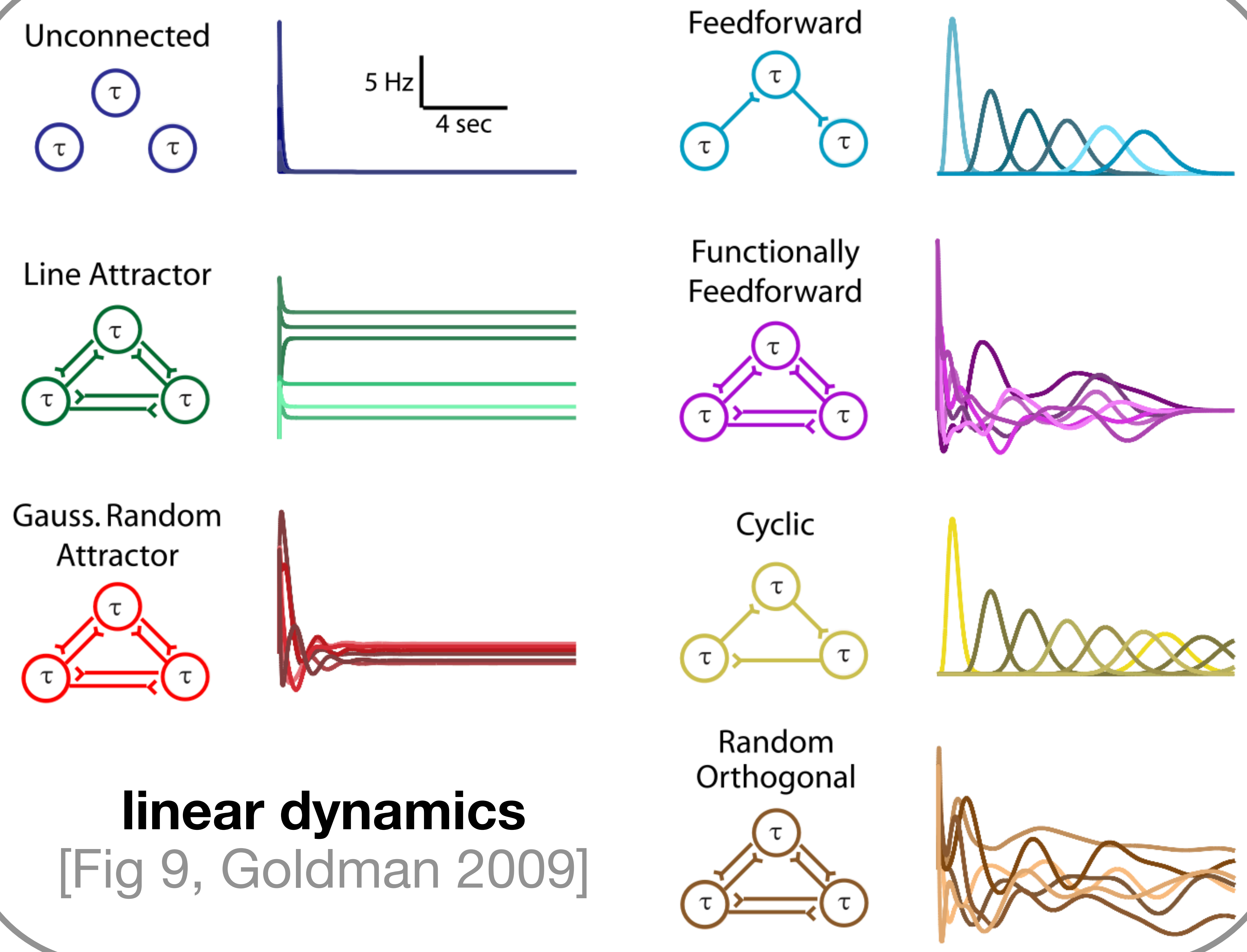
- Interval reproduction
- Delayed discrimination
- Evidence accumulation
- Copy-memory task
- k-bit flip-flop task
- permuted MNIST

No fundamental limit to the maximum temporal separation between the presentation of relevant information to the production of the desired behavior.

What kinds of neural dynamics support long-range temporal dependence learning?



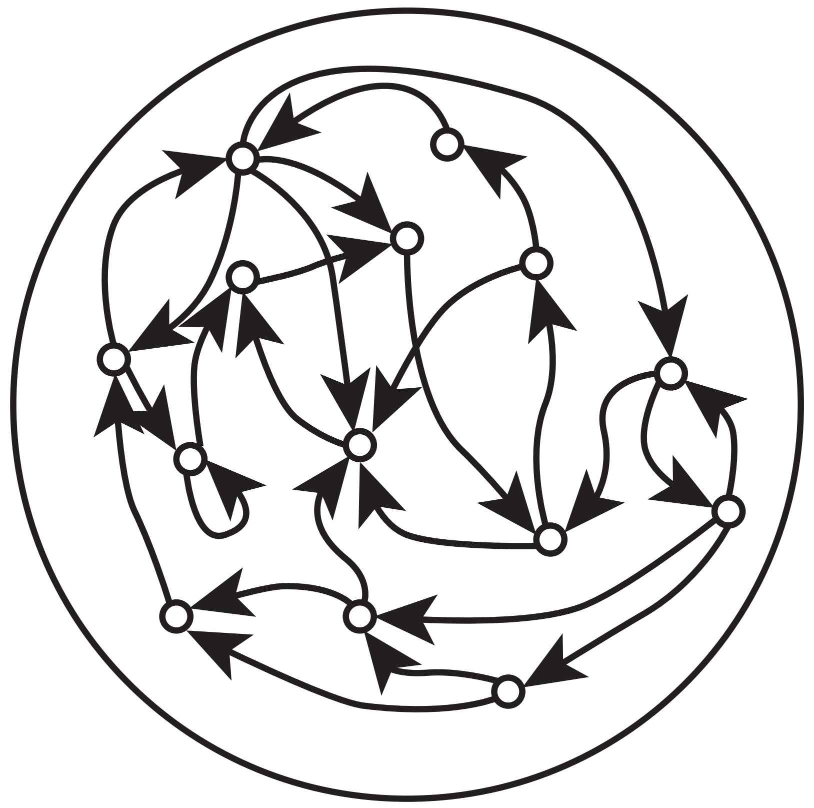
Is a good memory structure enough?



linear dynamics
[Fig 9, Goldman 2009]

nonlinear dynamics

Reservoir computing



Attractor dynamics

Chaos

edge-of-chaos

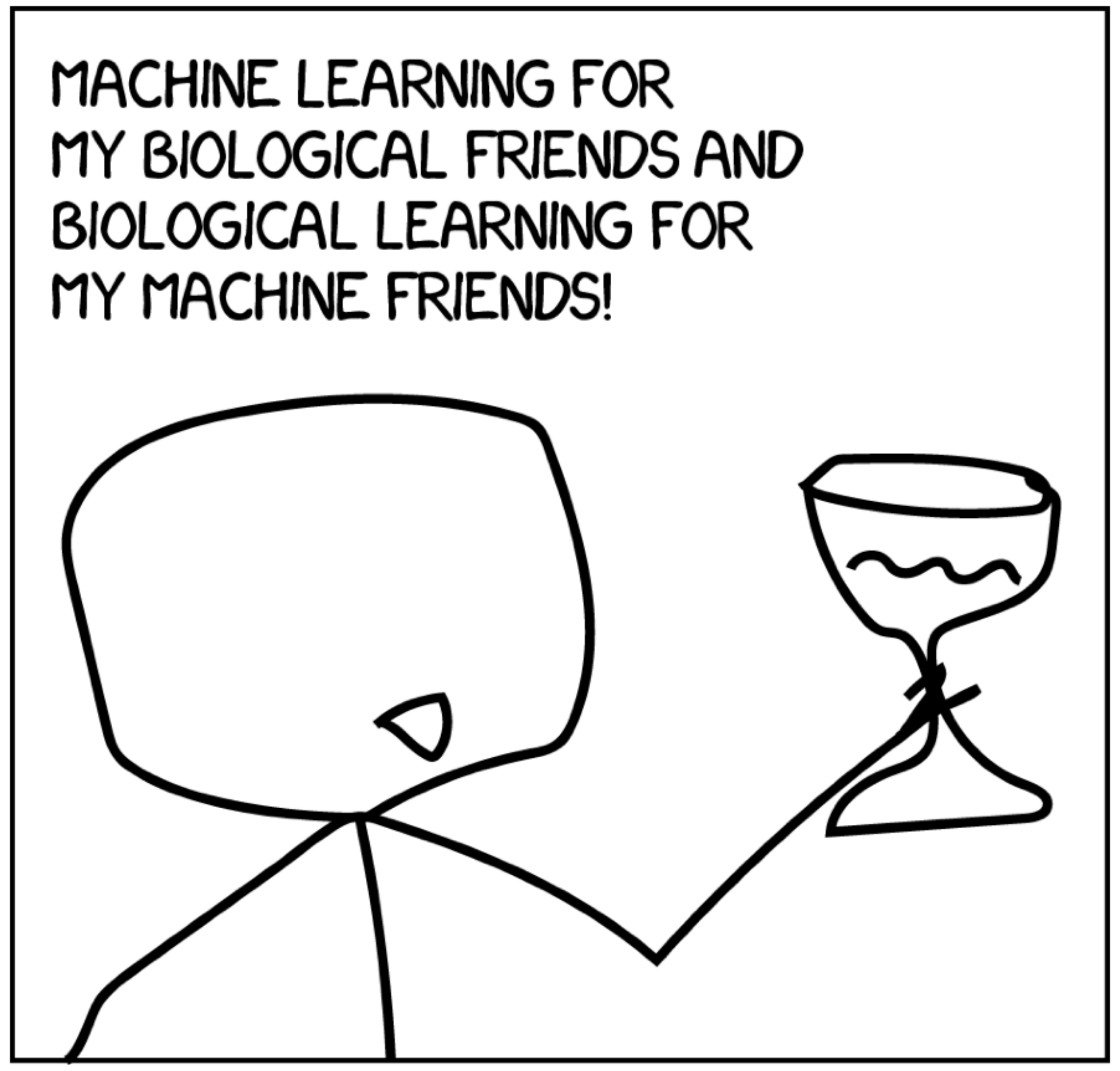
Neural manifolds

Oscillations

Meta-stable dynamics

...

Outline



neuro-AI o'clock

- **Decouple** good memory and good learning signals
- Characterization of **asymptotic** behavior of learning signals
- **Necessary condition** on the dynamics for good learning signal
- **Initialization scheme** for artificial recurrent neural networks
- Implications for biological neural networks

Learning to minimize error

statistical learning theory

Two kinds of strategies:

- Jump between potential solutions to find one with small error.
  - Evolutionary algorithms, logical reasoning
- Use directional learning signal derived from the error to make incremental changes.
  - Gradient descent!

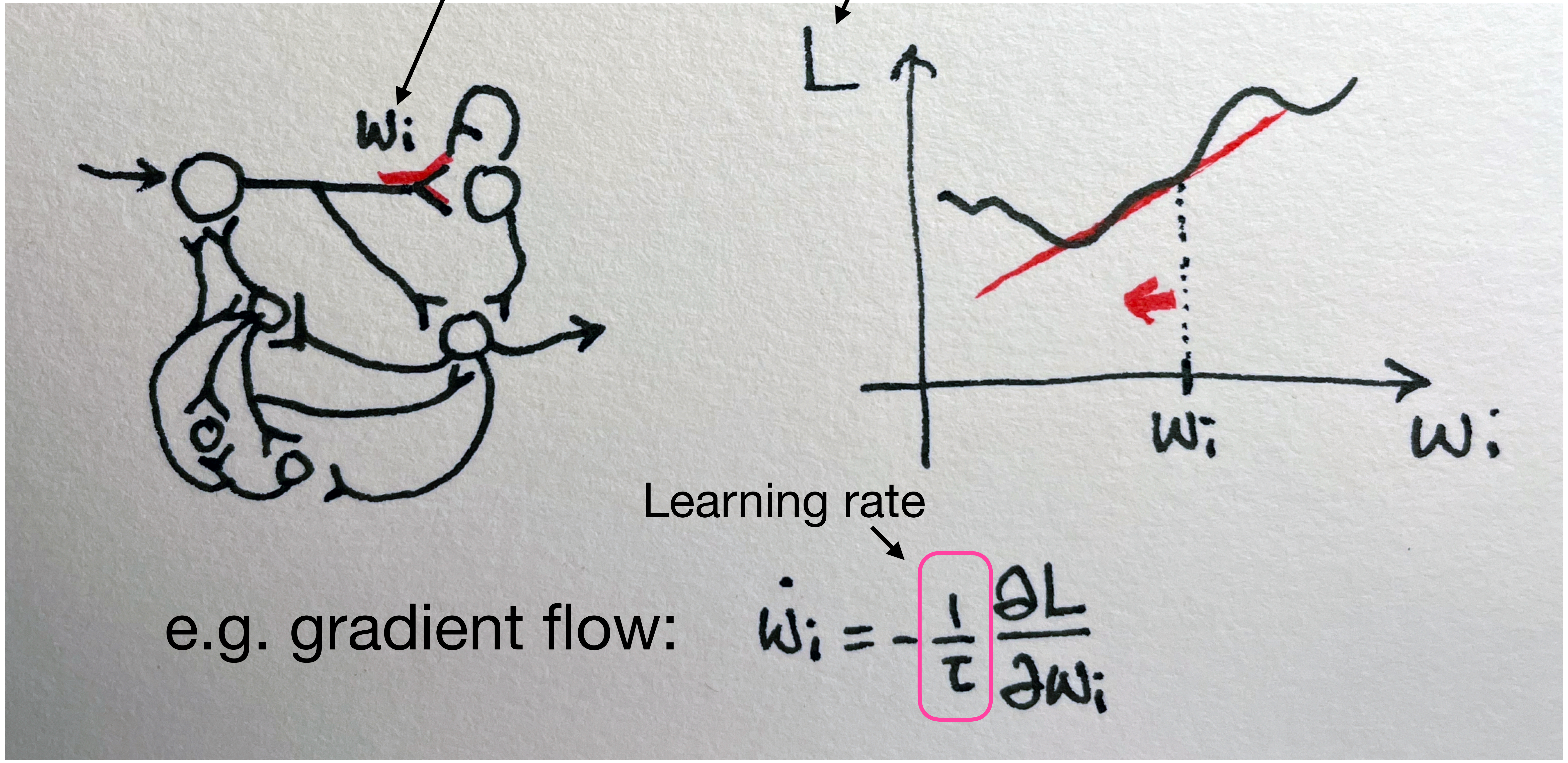
Rosenblatt's Perceptron (1957)

Widrow & Hoff's LMS (1960)

Gradient descent

Adjustable knob

Loss function



Gradient = learning signal

$$\frac{\partial L}{\partial w_i} = \lim_{\Delta \to 0} \frac{L(w_i + \Delta) - L(w_i)}{\Delta}$$

Limits on gradient representation

Dynamic range matters

- Mathematically, as long as the information processing is differentiable, we can use gradient descent to learn.
- However, gradients must be represented biophysically or digitally.
  - Due to noise, small gradients are indistinguishable from zero.
    Due to saturation, large gradients are treated equally.
  - Due to finite precision in floating points, similar numerical issues arise in ANNs.
- Practically, if the gradients are too small or large in magnitude, gradient descent fails.

EVGP

Exploding and Vanishing Gradient Problem

- Unfortunately, gradient signals often diverge or vanish in magnitude in deep neural architectures and recurrent networks as the chain of derivatives gets longer.
- EVGP in machine learning is tackled with various heuristics (next slide).
- EVGP in neuroscience has been discussed in the context of liquid state machines and chaos. [e.g. Mikhaeil et al. 2022; Laje & Buonomano 2013; Maass et al. 2002]
- Theoretical investigations have gaps. [Glorot & Bengio 2010; Bengio et al. 1994]

Approaches to resolve the EVGP

Dynamical systems view

Recurrent dynamics as an ODE

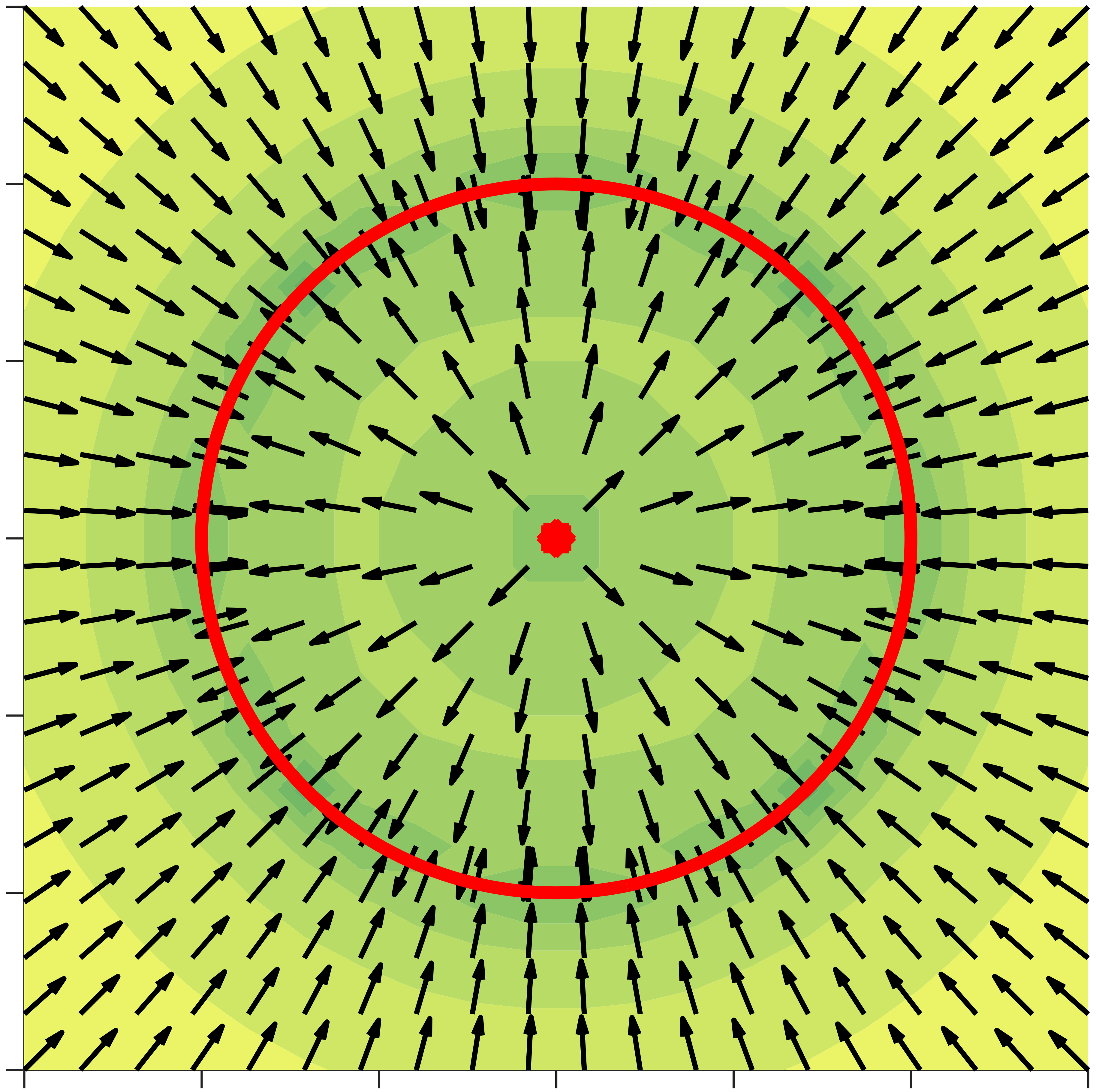
recurrent dynamics

stimulus or input

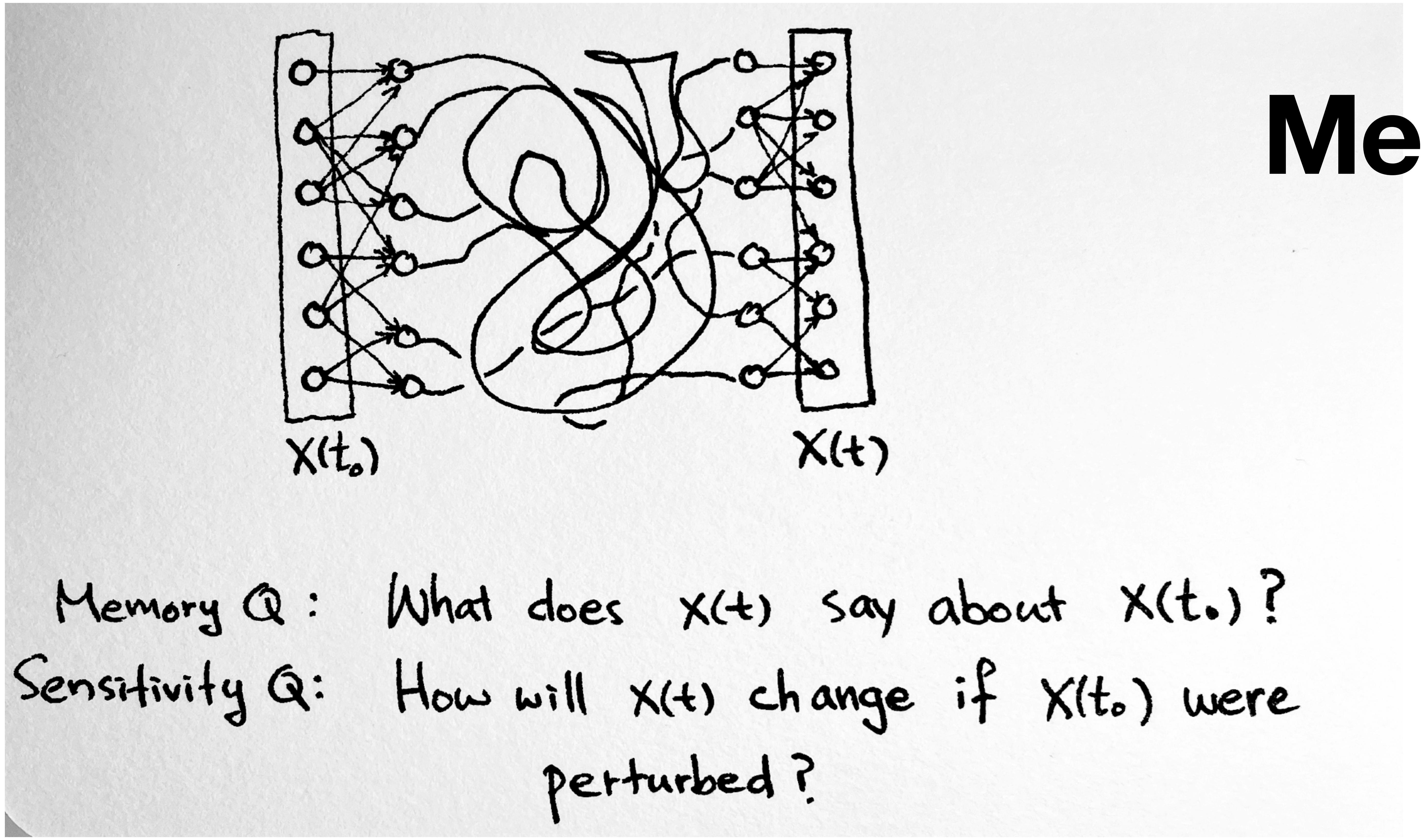
parameter vector

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x(t), u(t), w)$$

N-dim neural activity (hidden state)



Memory & Sensitivity



$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x(t), u(t), w)$$

$$x(t, x(t_0) + \Delta) - x(t, x(t_0))$$

difference between two different stimuli to be stored in memory (not infinitesimal)

Sensitivity: directional information (infinitesimal) ⟶

$$\delta(t) = \frac{\partial x(t)}{\partial x(t_0)} = \lim_{\Delta \to 0} \frac{x(t, x(t_0) + \Delta) - x(t, x(t_0))}{\Delta}$$

Sensitivity, adjoint, and gradient

$$\delta(t) = \frac{\partial x(t)}{\partial x(t_0)}$$

sensitivity

They are reciprocal twins.
They share the same fate!

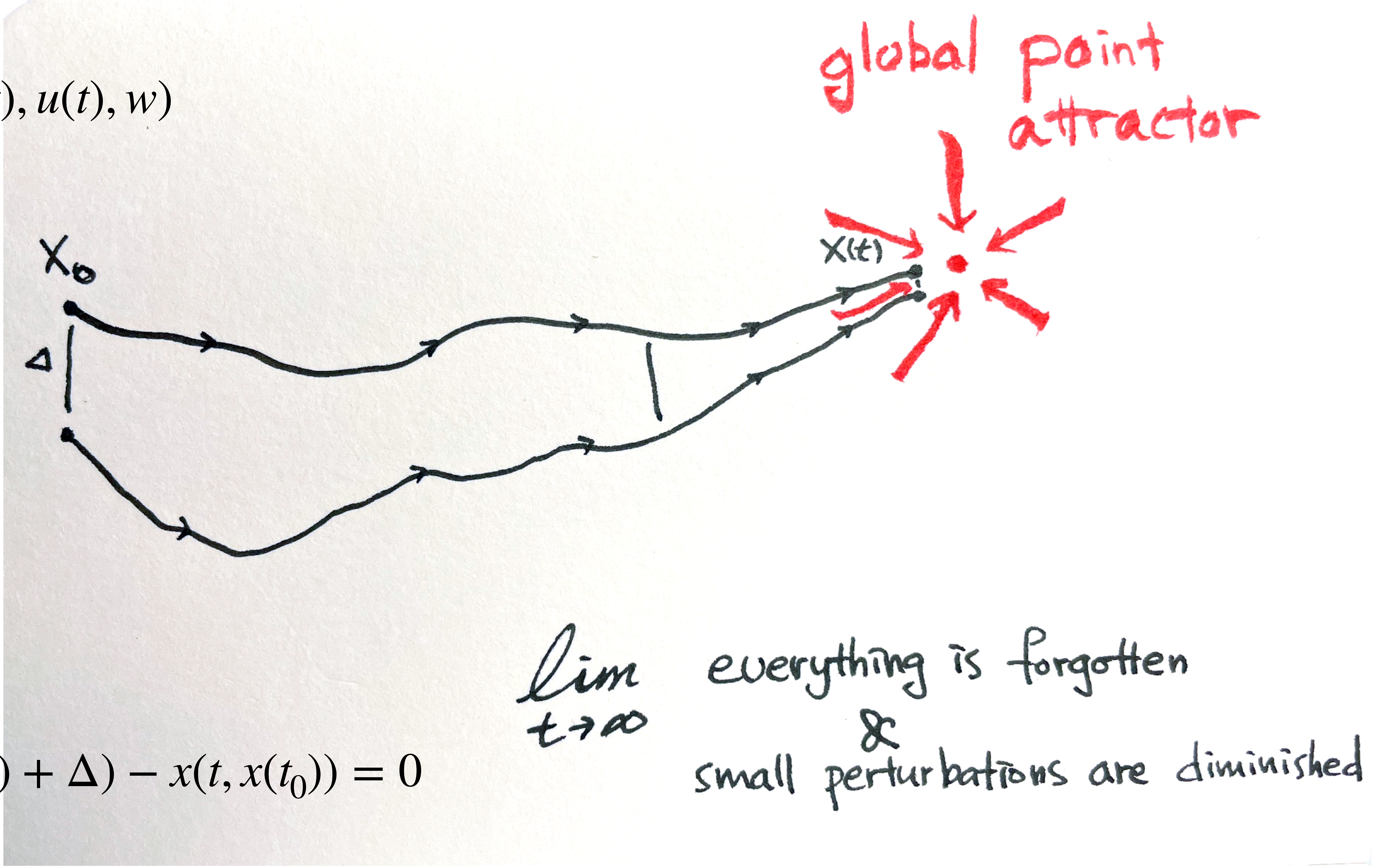
adjoint

gradient

$$\frac{\partial L}{\partial w_i} = \int_{t_0}^{t_1} \frac{\partial L}{\partial x(t_1)} \frac{\partial x(t_1)}{\partial x(t)} \frac{\partial x(t)}{\partial f(u(t))} \frac{\partial f(u(t))}{\partial w_i} \mathrm{d}t$$

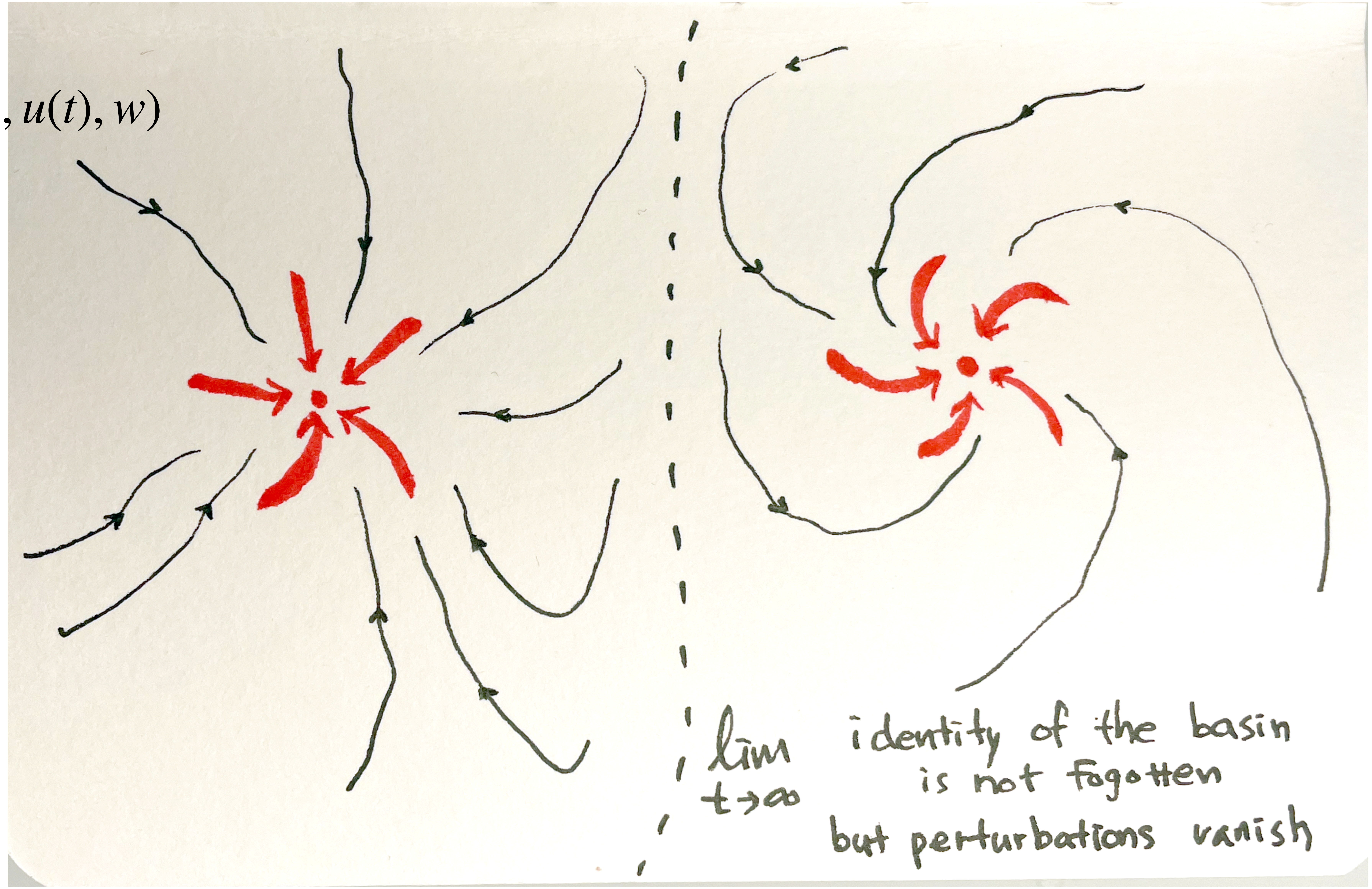
(adjoint=sensitivity) connects gradient over time

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x(t), u(t), w)$$



$$\lim_{t\to\infty} x(t, x(t_0) + \Delta) - x(t, x(t_0)) = 0$$

$$\frac{\mathrm{d}x}{\mathrm{d}t} = f(x(t), u(t), w)$$



Robust memory comes with vanishing gradient

IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 5, NO. 2, MARCH 1994

Learning Long-Term Dependencies with Gradient Descent is Difficult

Yoshua Bengio, Patrice Simard, and Paolo Frasconi, *Student Member, IEEE*

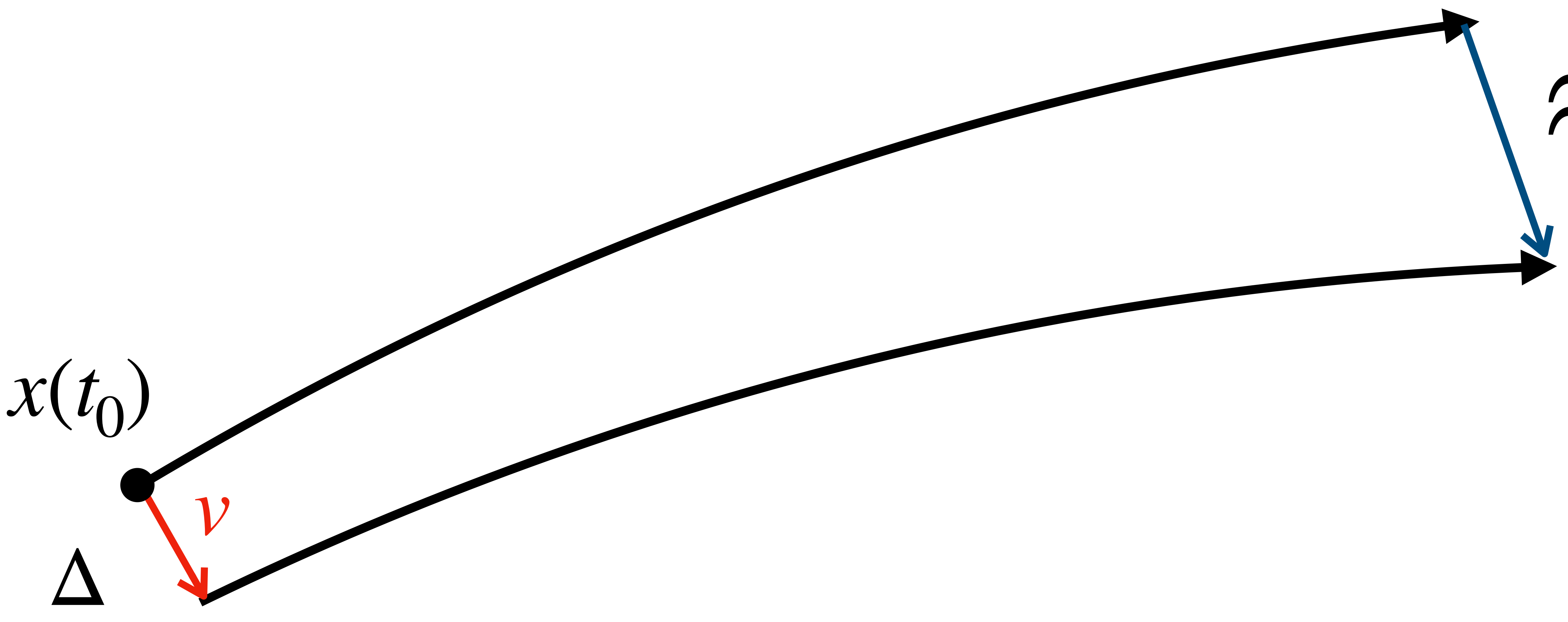
1. Infinitely long memory → • Existence of non-fading states
2. Robust memory content → • Attractor dynamics
3. Non-vanishing/exploding gradient

incompatible

Lyapunov exponents

exponential time constant of perturbed variation

- Characterizes asymptotic behavior of the gradients and sensitivity dynamics



$$\approx \Delta e^{\lambda t}$$

$$\left\| \frac{\partial x(t)}{\partial x(t_0)} v \right\| \approx e^{\lambda t} \cdot \|v\|$$

$$\lambda(x(t_0), \mathbf{v}) := \limsup_{t \to \infty} \frac{1}{t} \log \left\| \frac{\partial x(t)}{\partial x(t_0)} \mathbf{v} \right\|$$

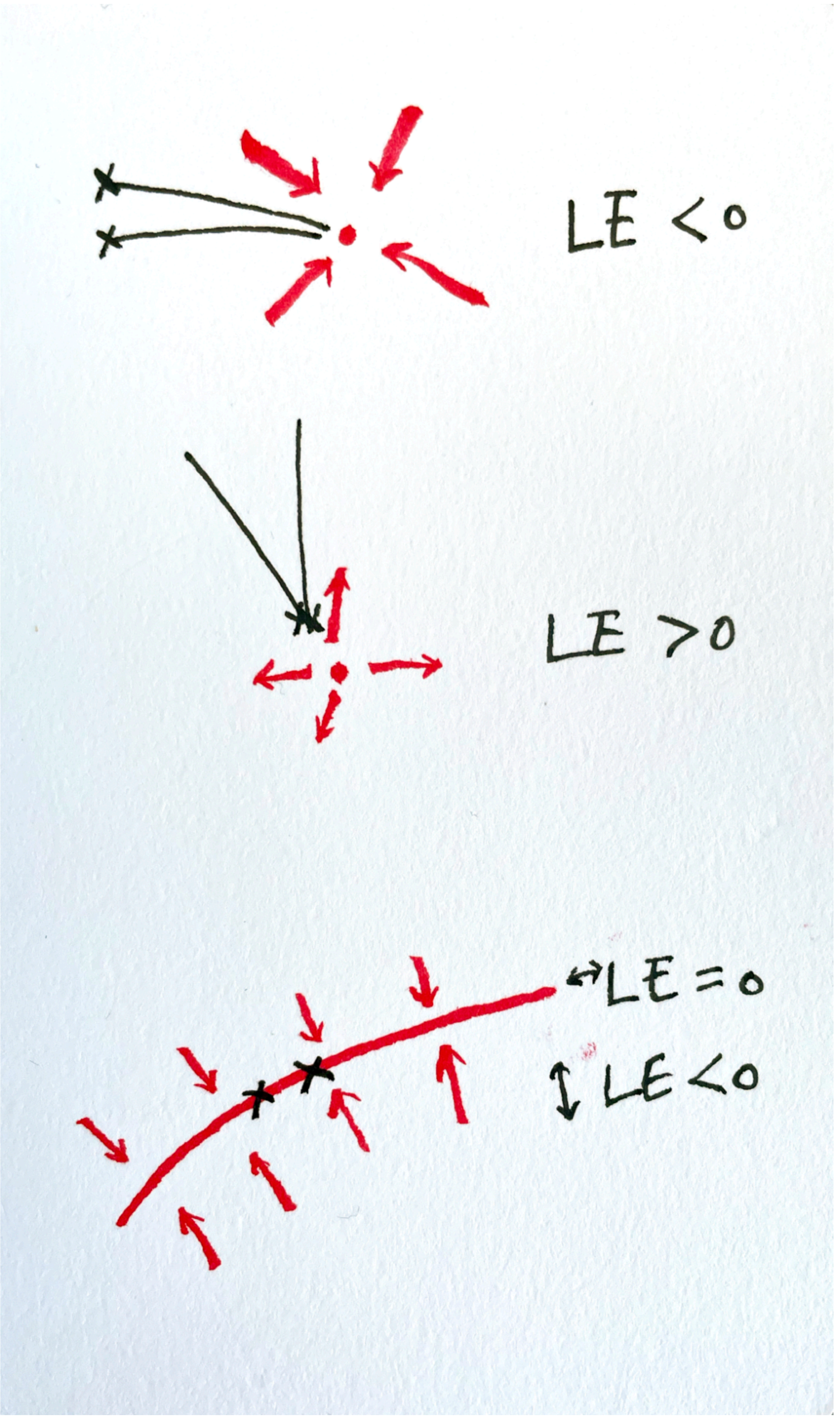
Lyapunov exponent (LE)

Lyapunov exponents

per attractor

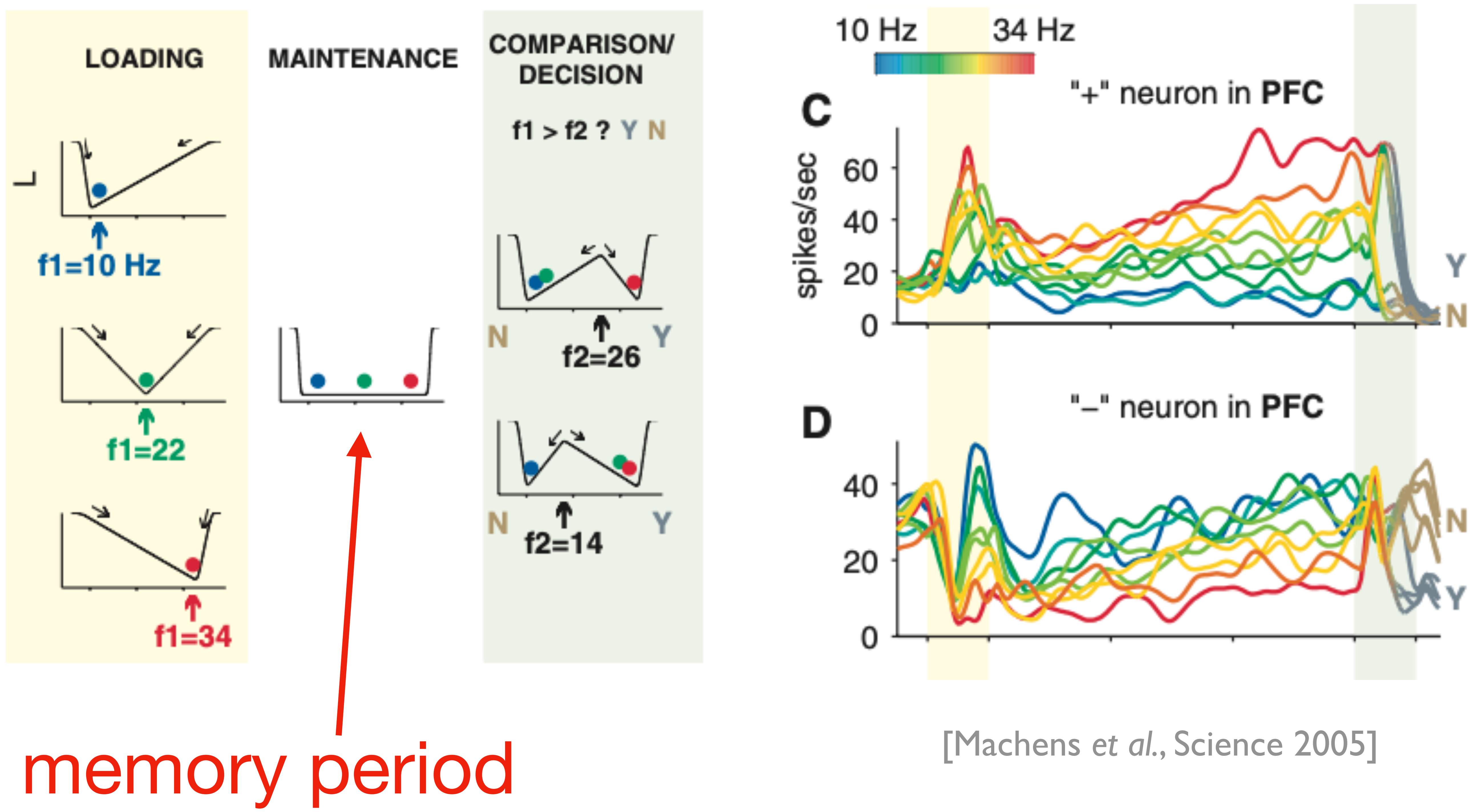
- Within a basin of attraction, all states share the same fate and LE spectrum*. (* under certain assumptions)
- Positive LE: asymptotically exploding gradient
- Negative LE: asymptotically vanishing gradient
- Zero LE: asymptotically marginally stable

What are the systems with many zero LEs?



Continuous attractor dynamics

- No flow within a low-dimensional manifold, attractive flow to the manifold.
- Persistent neural activity while memory content is held.



[Machens *et al.*, Science 2005]

Continuous attractor dynamics

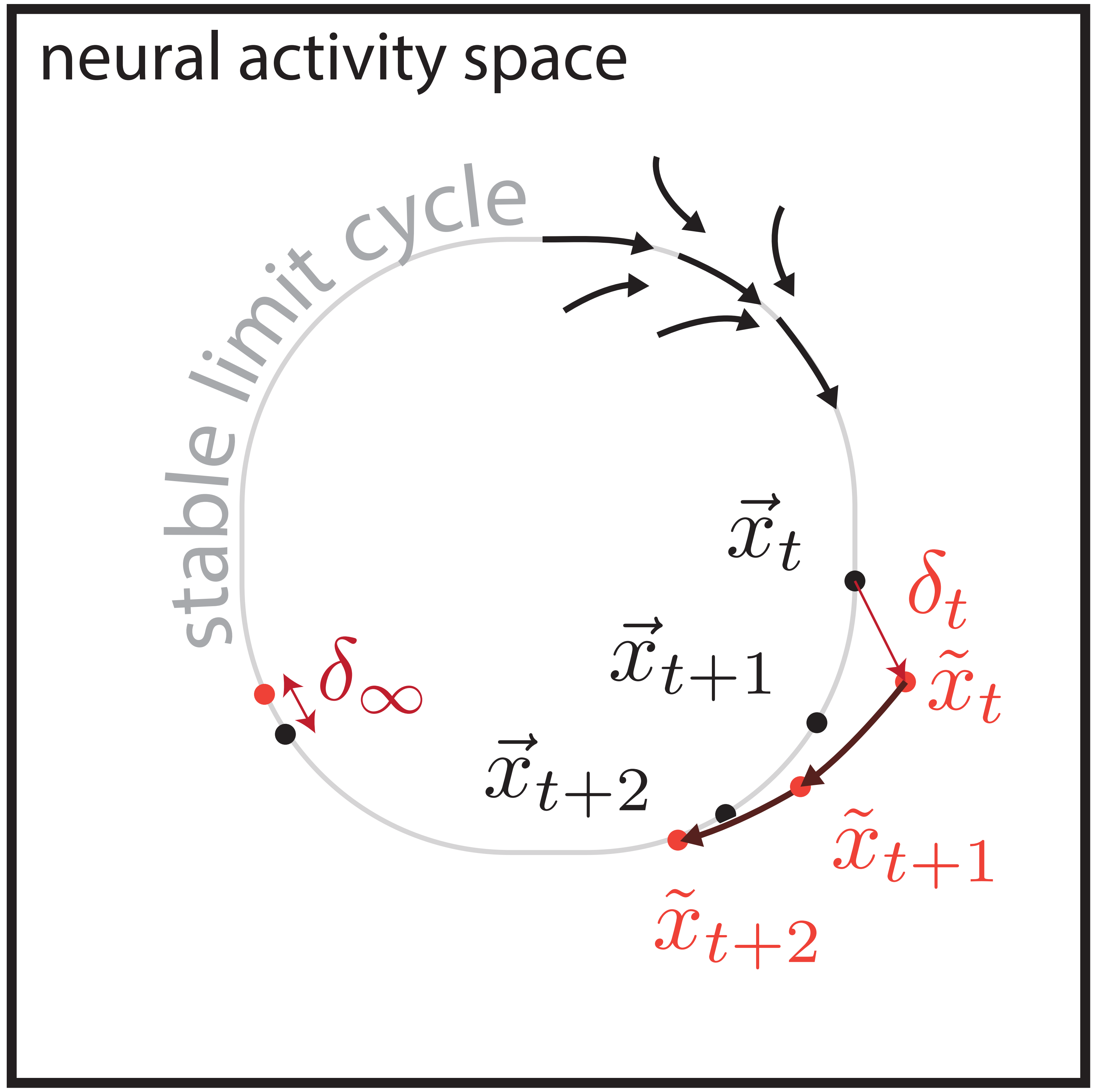
- In general, continuous attractor networks have an attracting manifold with constant (typically zero) flow. The "continuity" refers to the manifold structure which resembles the familiar continuous Euclidean space.

- Issue: **fine tuning problem**

$$\tau \frac{\mathrm{d}x_i}{\mathrm{d}t} = -x_i(t) + \sum_j \boxed{w_{i,j}} x_j(t) + I_i(t)$$

recurrent excitation has to counter the decay precisely

Stable limit cycle dynamics

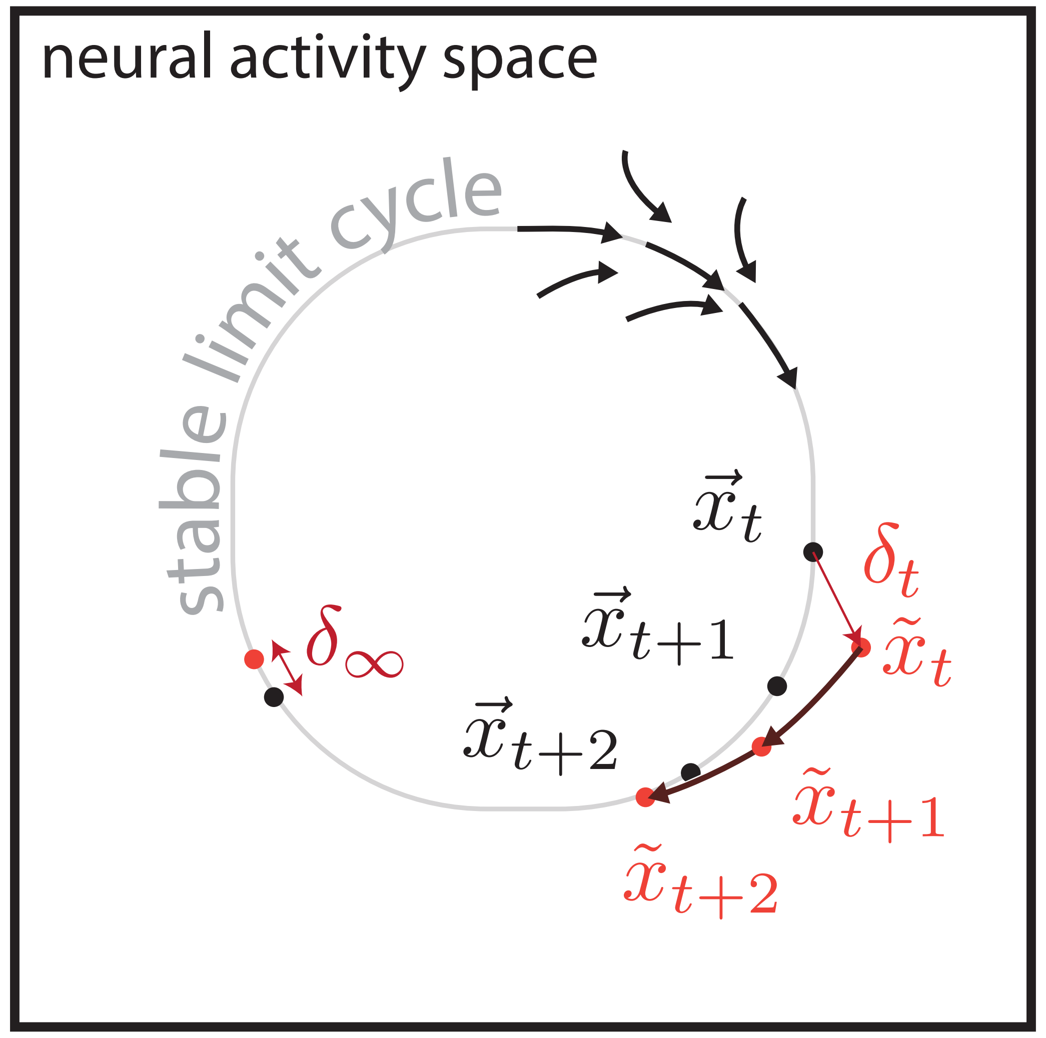


sensitivity remains for infinite time

- infinitesimal and finite perturbations of the *phase* are not forgotten.
  - good sensitivity
- linearized dynamics (thus the sensitivity and adjoint) are asymptotically periodic.
- 1-dimension non-vanishing/non-exploding gradient (1 zero LE)

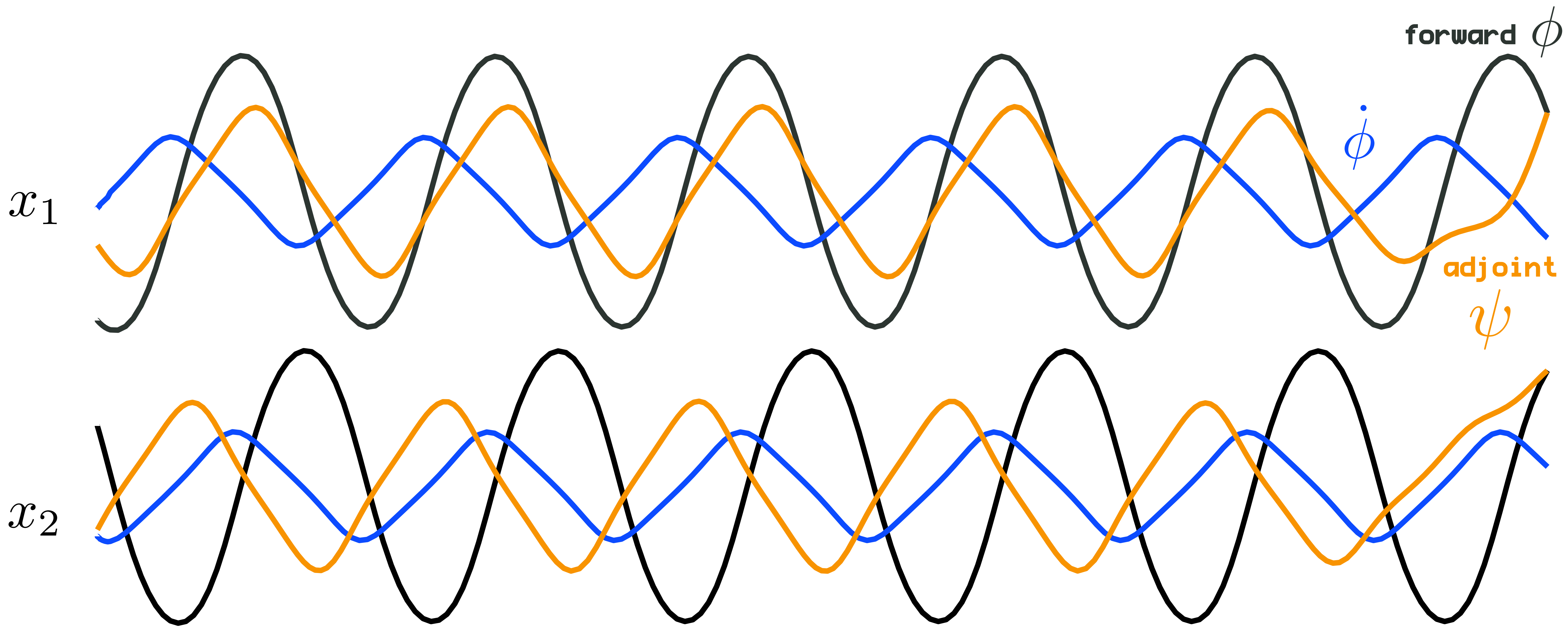
[Sokół et al., Asilomar 2019]

Stable limit cycle dynamics

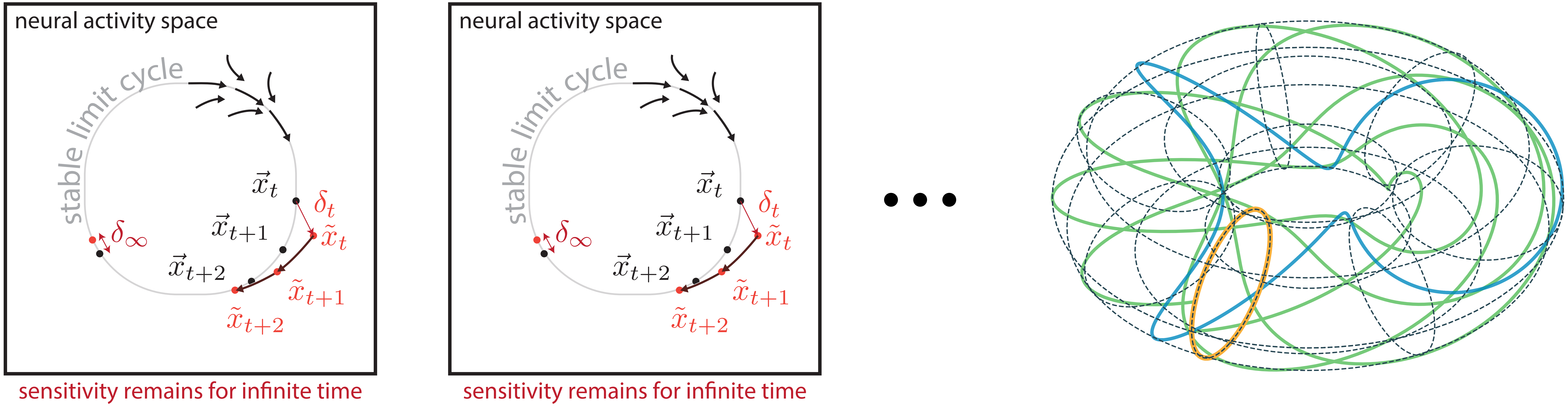


sensitivity remains for infinite time

- adjoint / learning signal is periodic



Quasi-periodic attractor dynamics

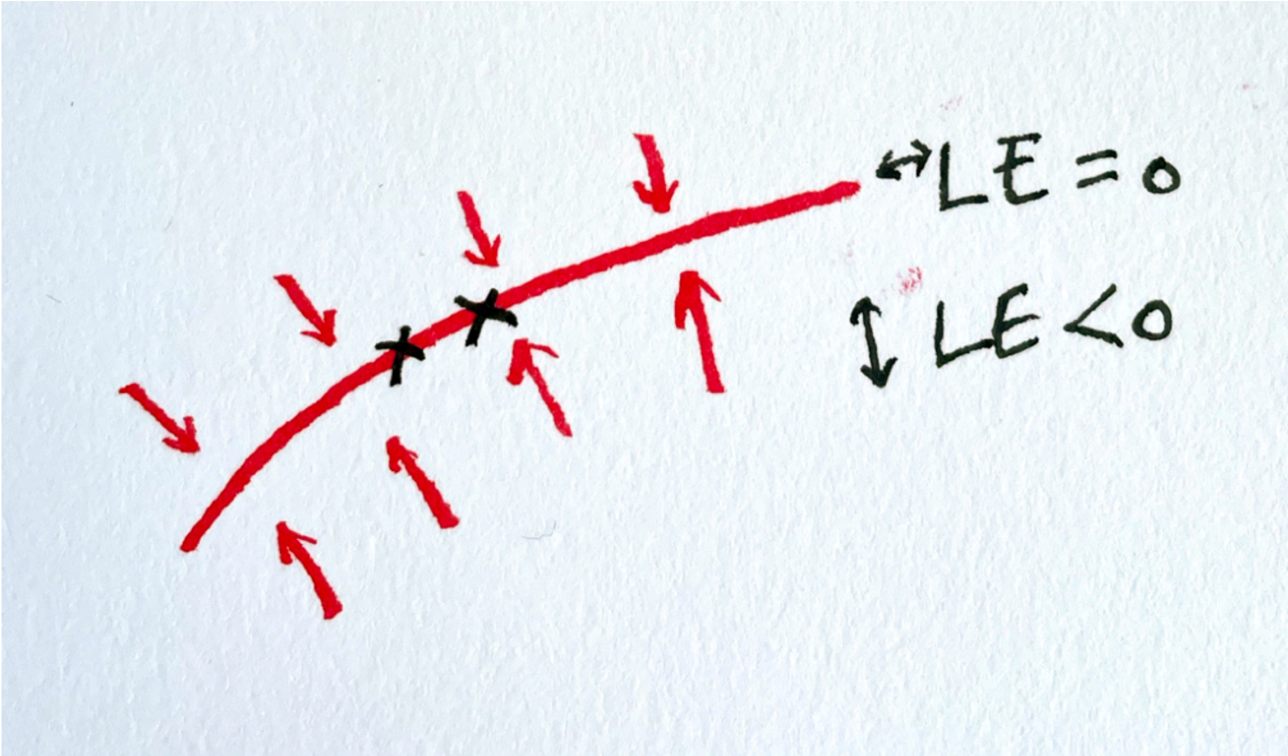


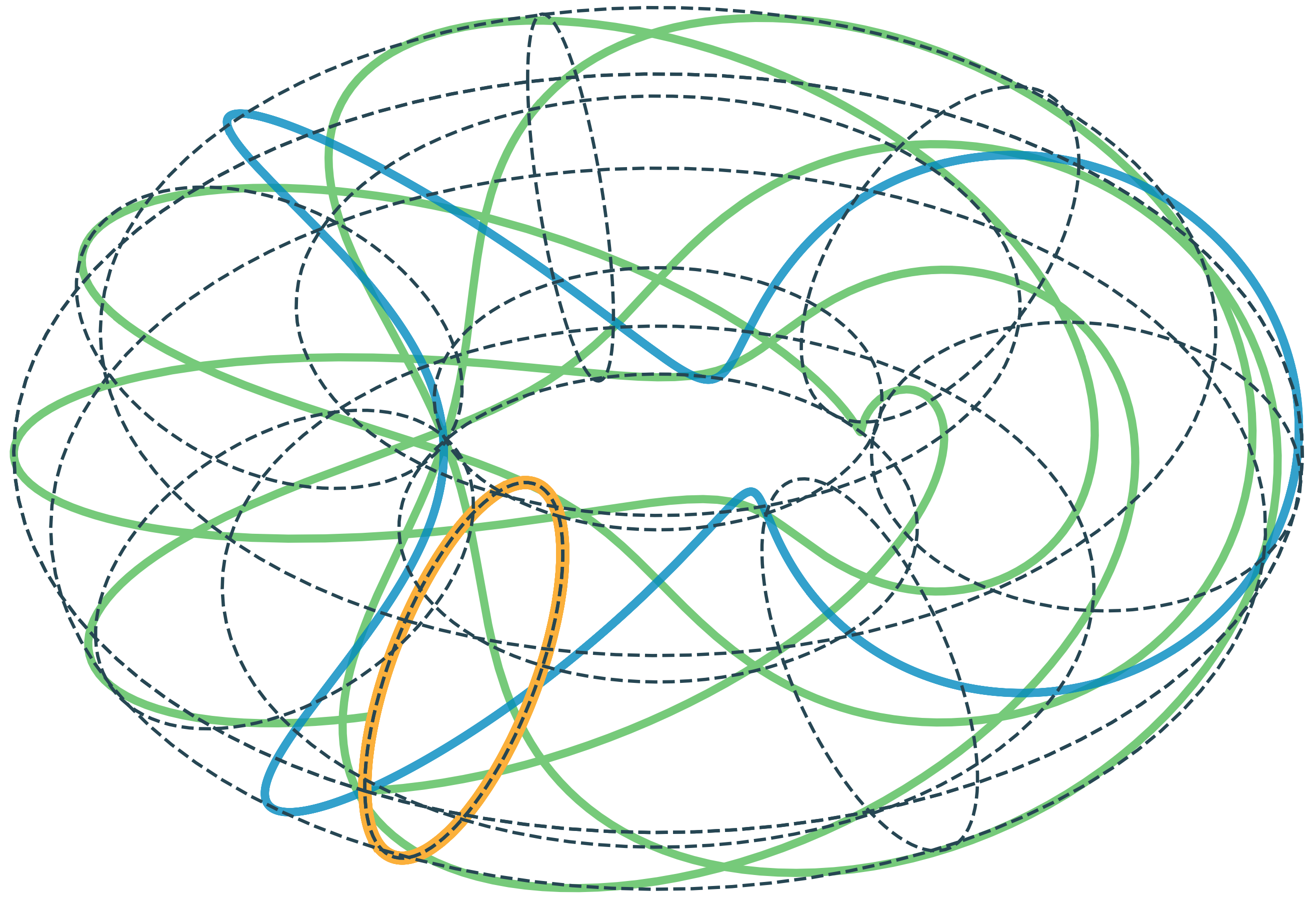
- Multiple independent nonlinear oscillators (with different frequencies)
- Does not suffer from the fine tuning problem (structurally stable)

Only two types dynamical structures

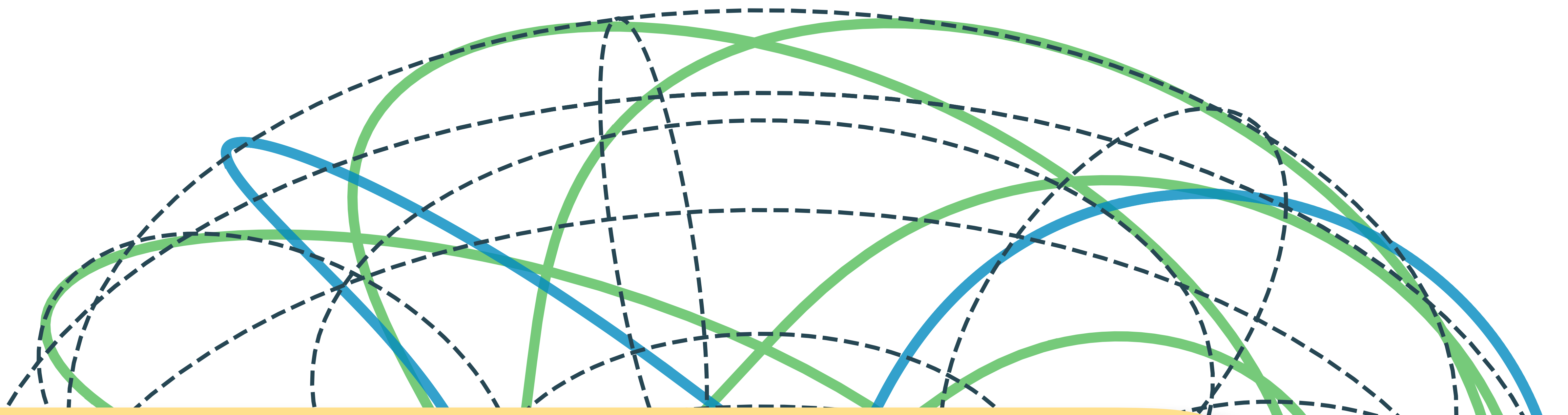
Or their mixture

- Continuous attractors
  - D-dimensional **arbitrary manifold** = D zero-LE
- Periodic / quasi-periodic attractors
  - D-dimensional **torus** = D zero-LE
  - periodic / quasi-periodic learning signals
  - **robust to perturbation of parameters**





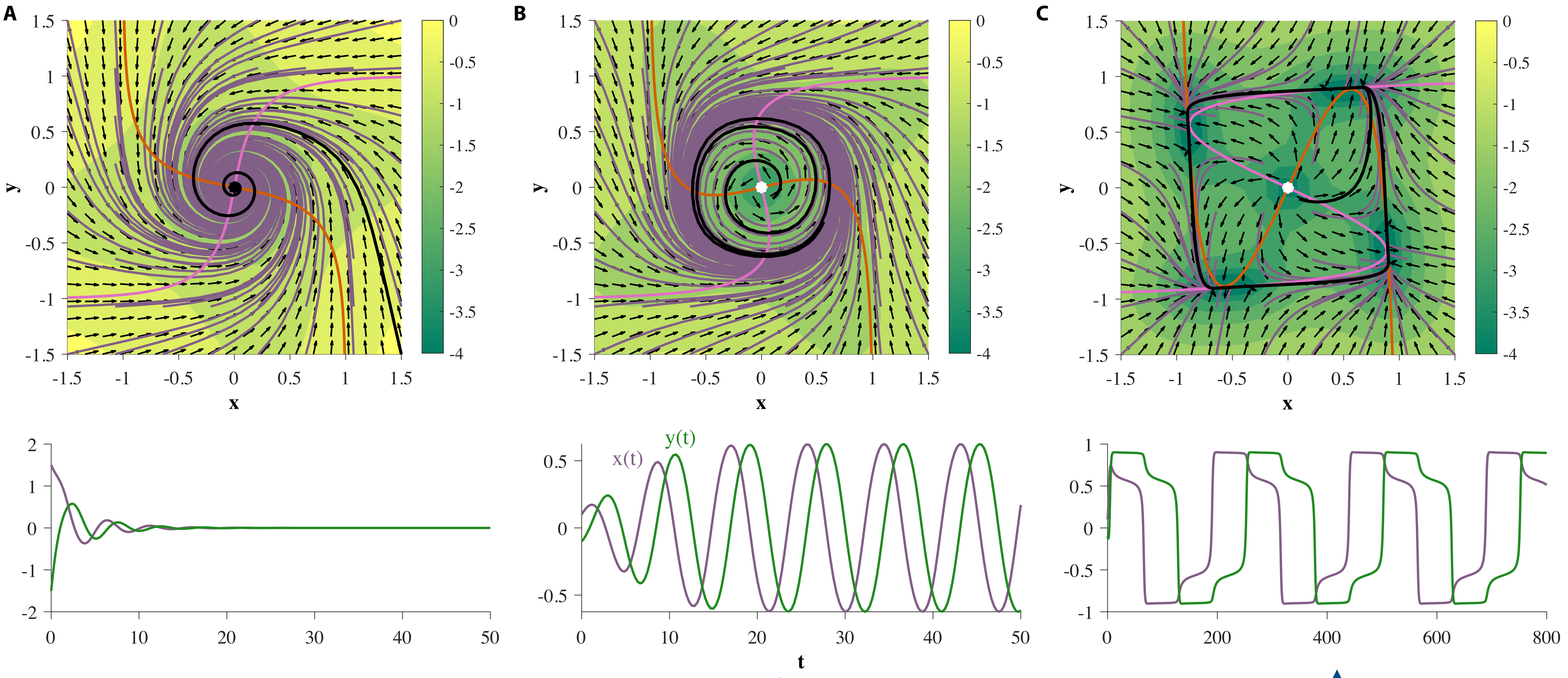
Conjecture



The only “robust” dynamical structure that can carry sensitivities without EVGP for any interval is the **stable limit cycle attractor** or more generally the **quasi-periodic toroidal attractor**.

Application: initialization scheme for RNNs

- Next steps
  - find parameters for RNNs that exhibit stable limit cycle
  - initialize RNNs in this regime and train on difficult tasks
  - ?
  - profit!
- Let's consider the tanh-RNN and GRU (gated recurrent unit) RNNs [Jordan et al., 2018]



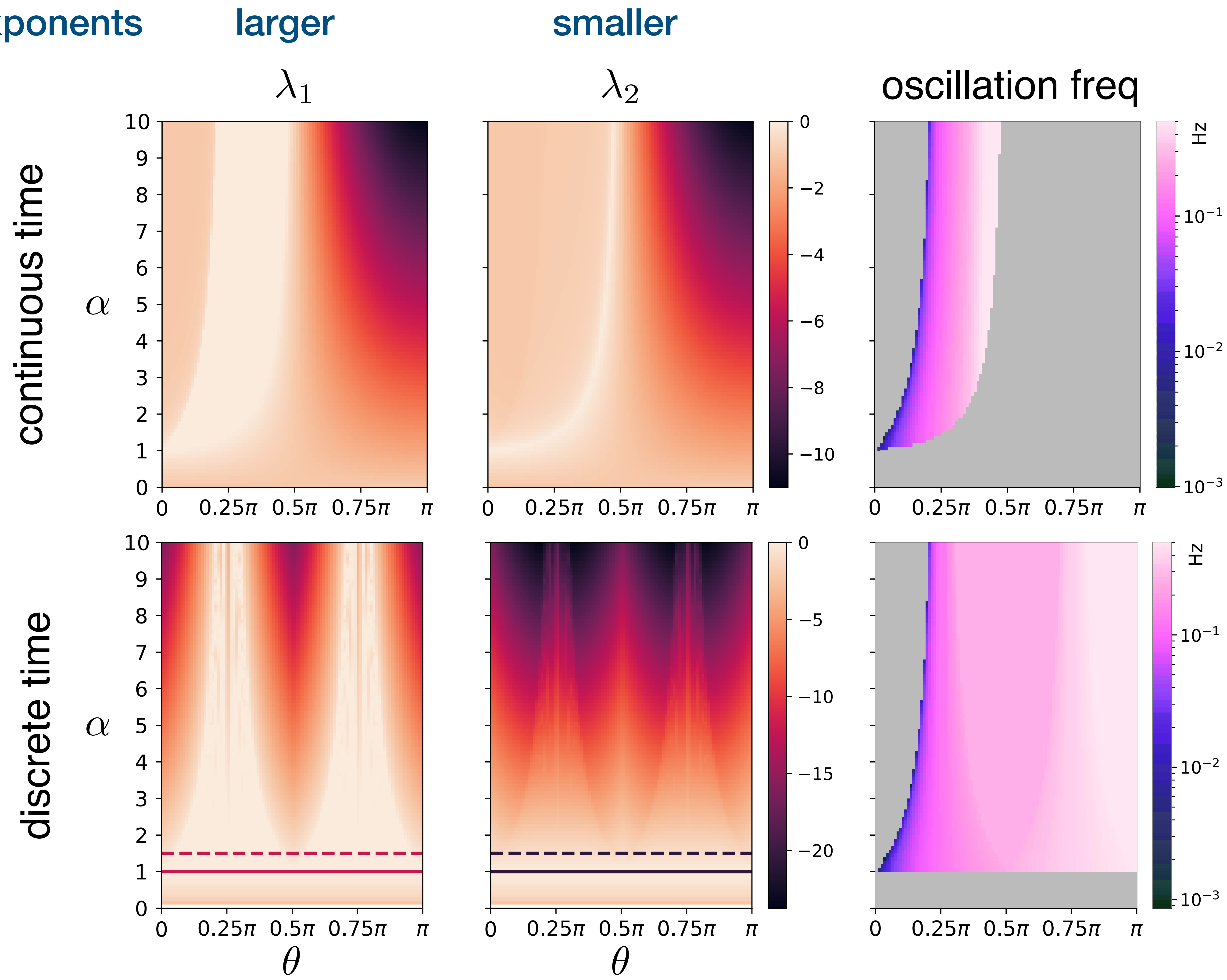
[Jordan et al., 2018]

limit cycle emerges!
(Hopf bifurcation)

nonlinear oscillation

numerically estimated Lyapunov/Floquet exponents

$$\tanh\left(\alpha \begin{bmatrix}\cos\theta & -\sin\theta \\ \sin\theta & \cos\theta\end{bmatrix} \mathbf{x}_t\right) = \begin{cases}\dot{\mathbf{x}} \\ \mathbf{x}(t+1) - \mathbf{x}(t)\end{cases}$$



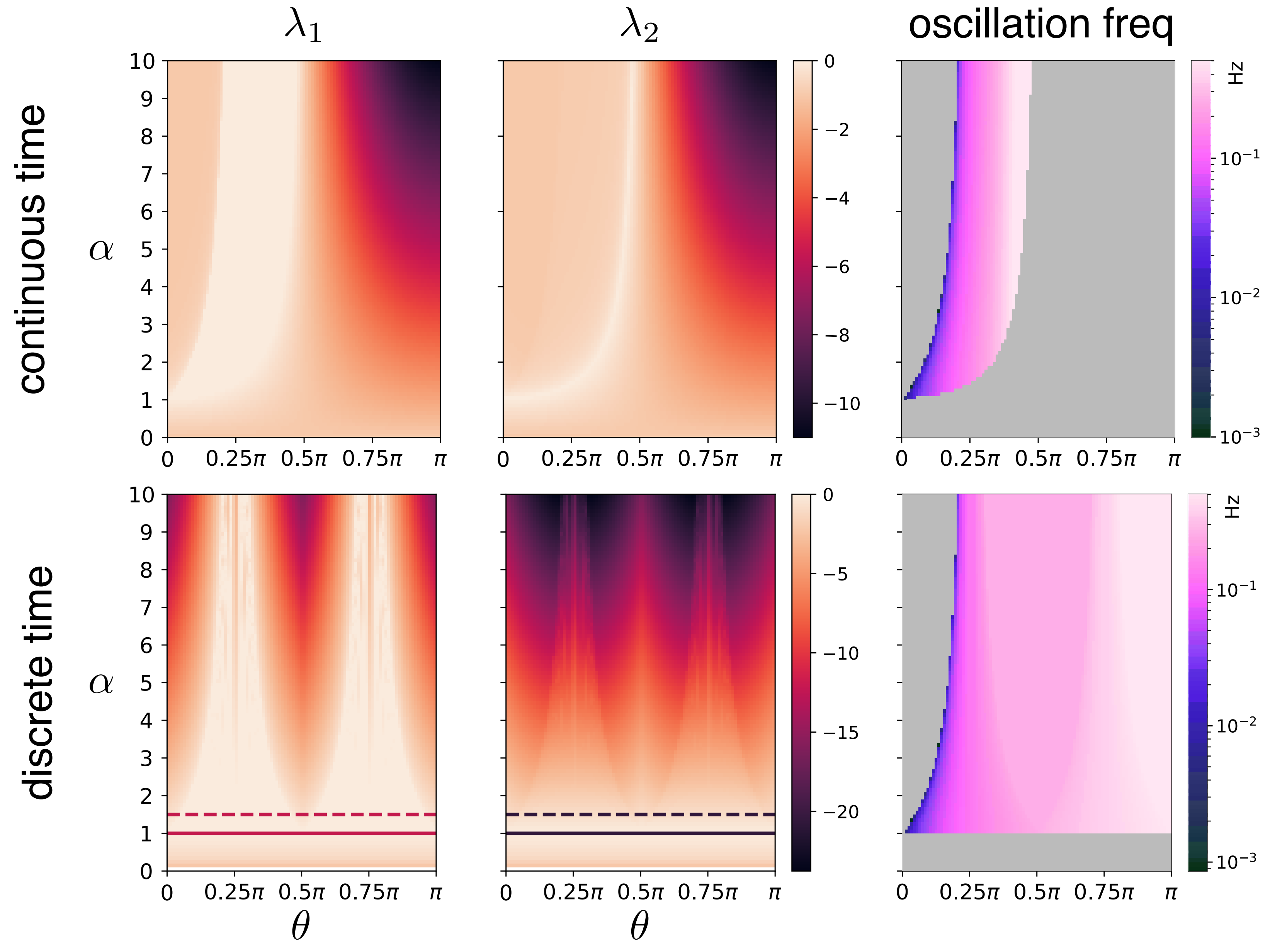
- A region of parameter space corresponds to stable limit cycle.
- Discrete time system has more interesting features emerging from the failure of Euler integration connection…

Block diagonal initialization

A collection of 2D uncoupled oscillators with random parameters

- At initialization we have "good" gradients

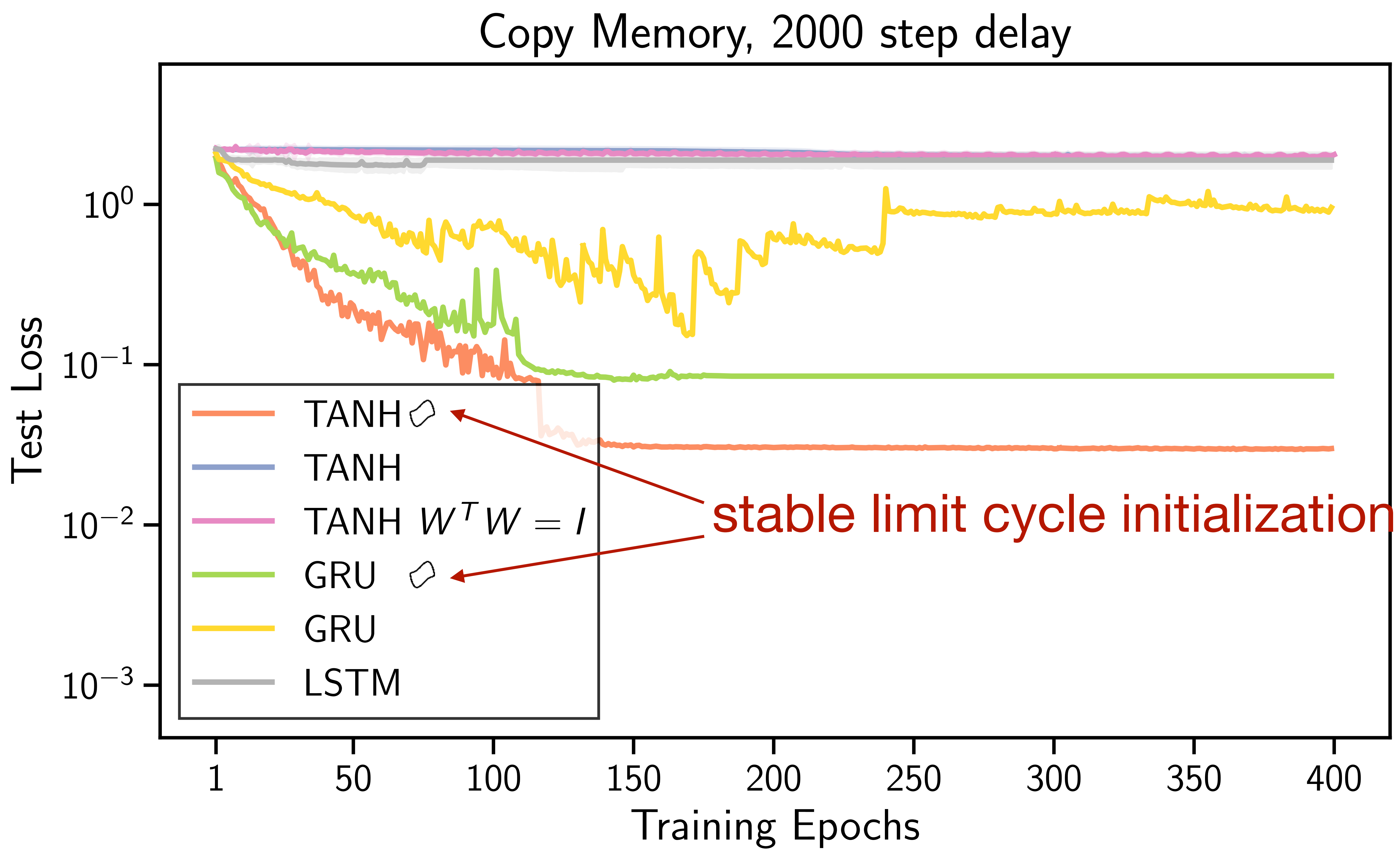
$$\mathbf{W}_{\text{init}} = \begin{bmatrix} \alpha_1 \begin{pmatrix} \cos(\theta_1) & -\sin(\theta_1) \\ \sin(\theta_1) & \cos(\theta_1) \end{pmatrix} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \alpha_m \begin{pmatrix} \cos(\theta_m) & -\sin(\theta_m) \\ \sin(\theta_m) & \cos(\theta_m) \end{pmatrix} \end{bmatrix}$$



Results:

Numerical experiments in discrete time

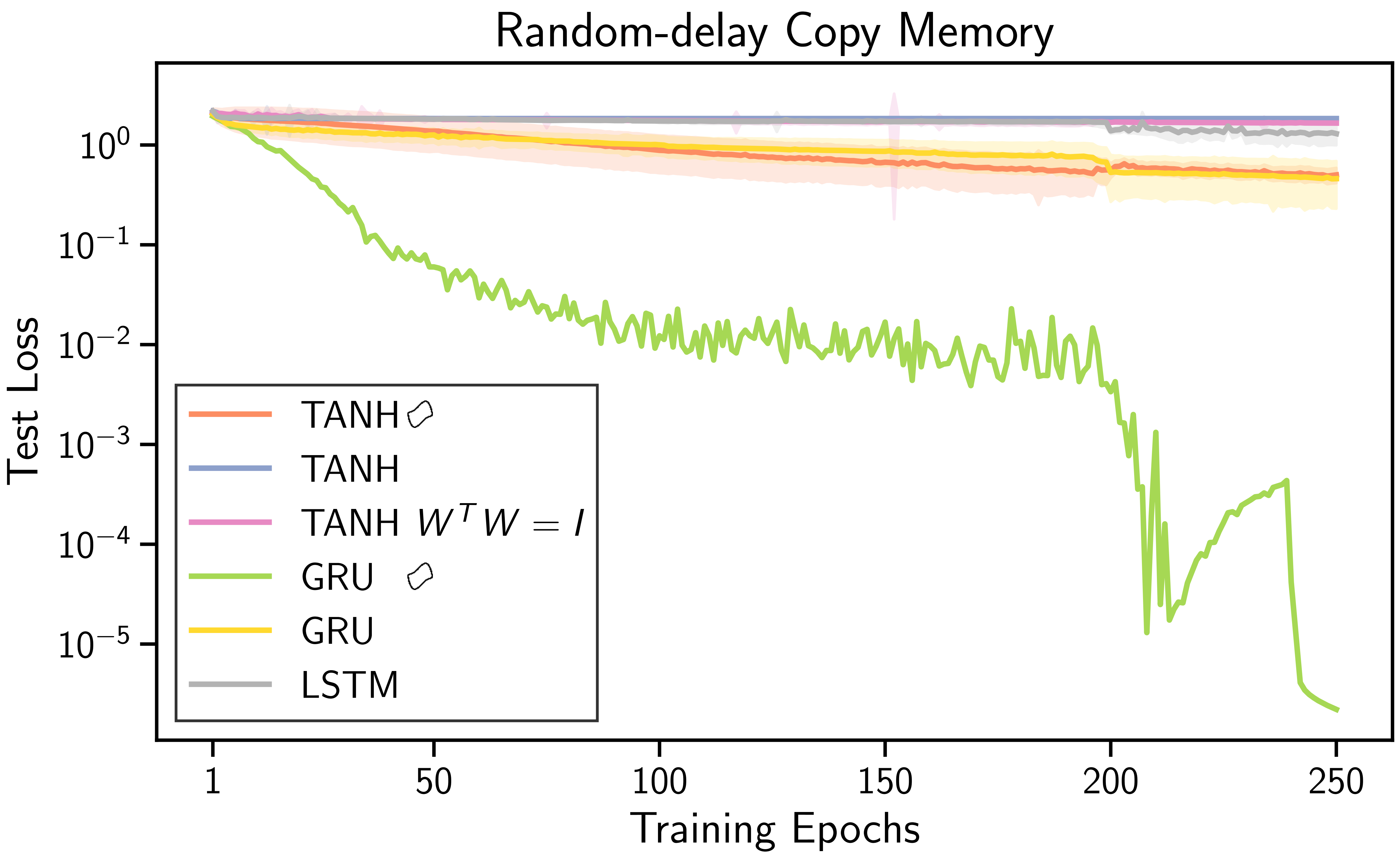
- **Copy Memory task**: remember a sequence of symbols during delay and spit them out later.
- Stable limit cycle initializations converges quickly and solves the task with no tricks.



Results:

Numerical experiments in discrete time

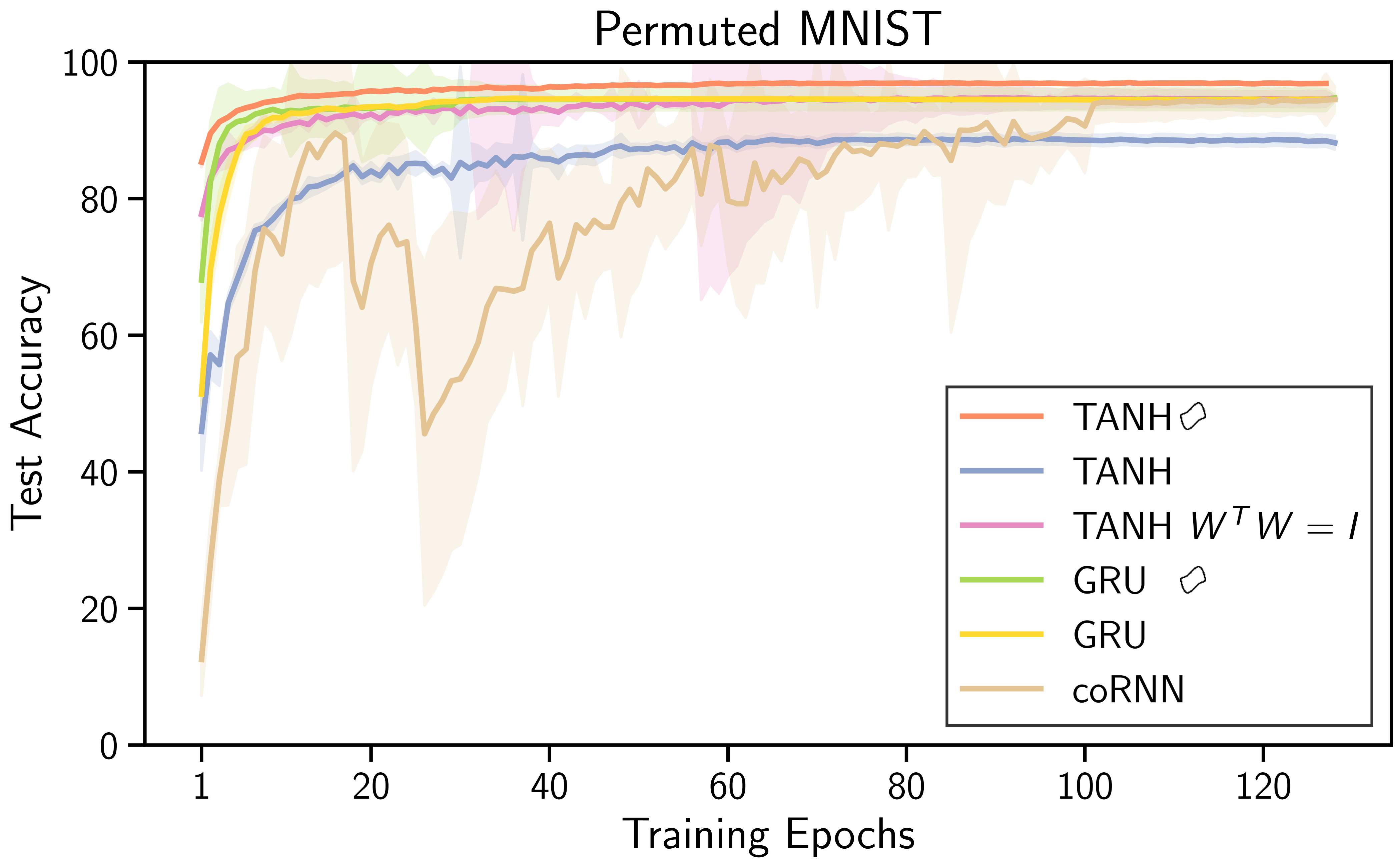
- **Random-delay version**: GRU with stable limit cycle init reliably found a solution.
- Final GRU solution shows no sign of oscillations!

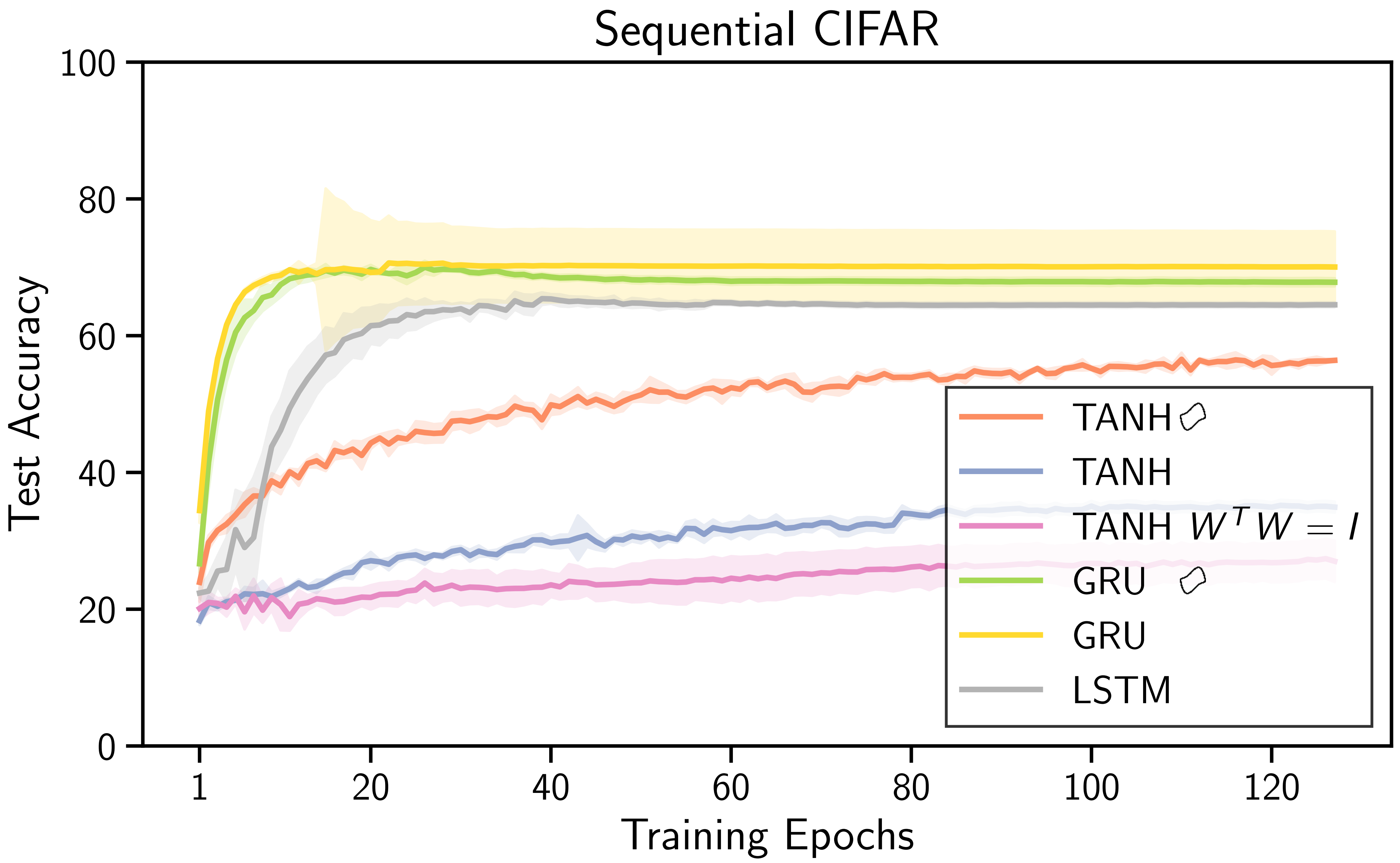


Results:

Numerical experiments in discrete time

TANH top accuracy 96.99 vs coRNN's ("State-of-the-art") 94.68



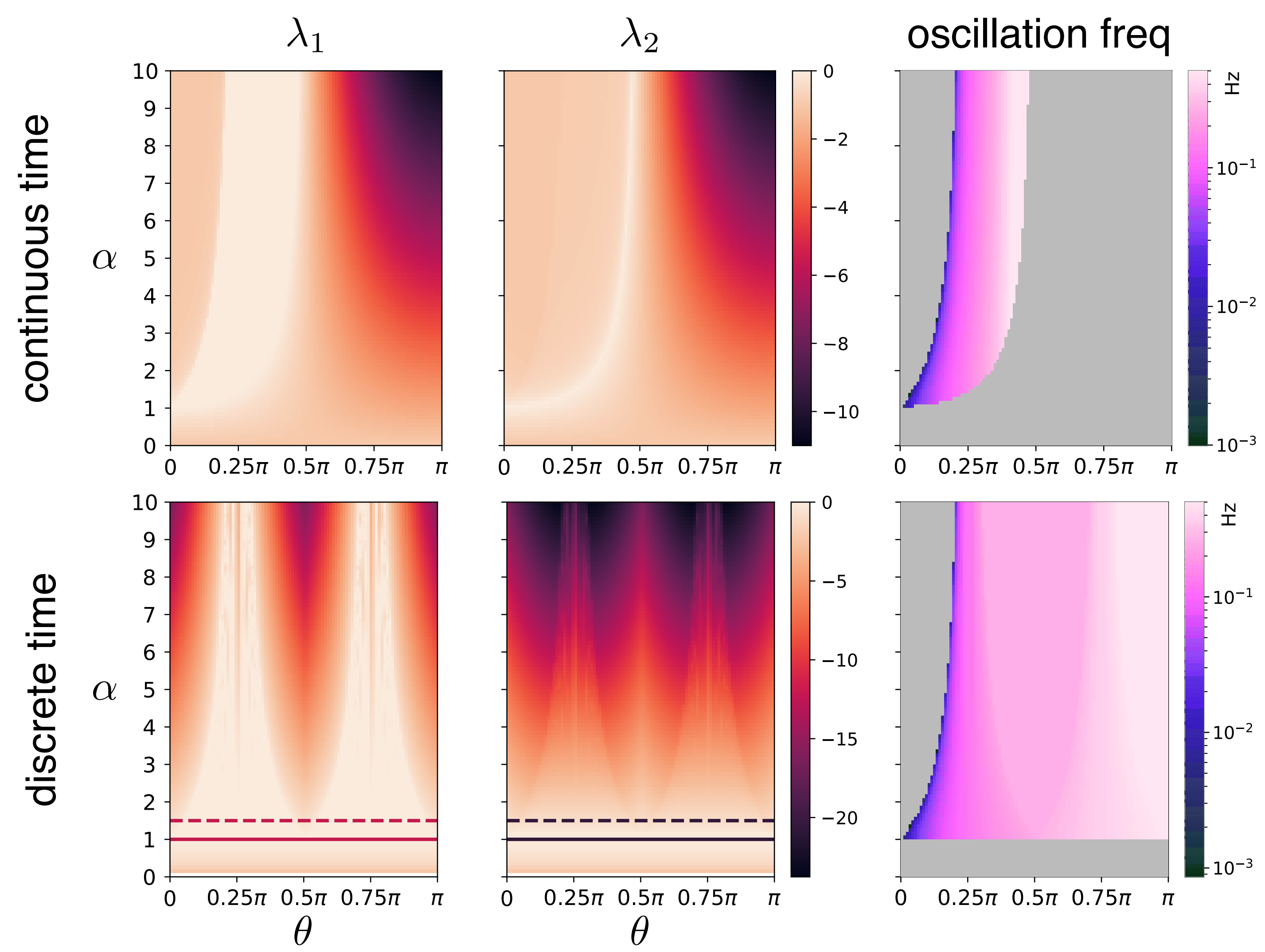


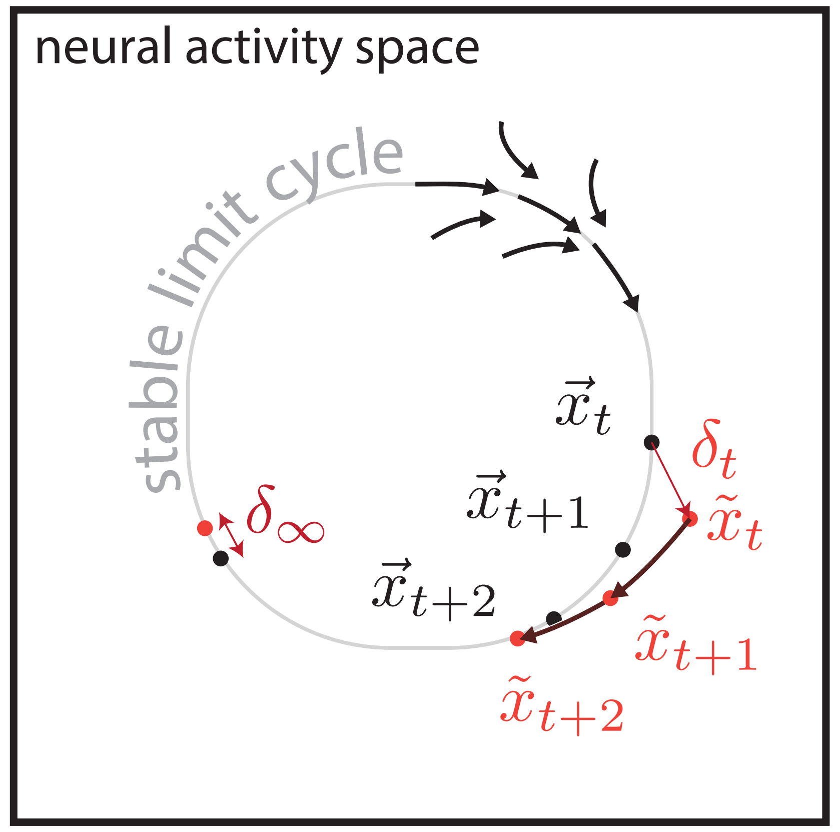
Biological implications
(we don’t know how it could be implemented yet!)

- Oscillations at rest, multiple frequencies, not fully synchronous.
- Persistent form of eligibility trace is quasi-periodic.
  - In the absence of oscillations, long temporal relations should be hard to learn.
  - Resetting oscillations should disrupt learning.
- Input should have lasting desynchronization effect.
- Spiking neurons with baseline quasi-periodic firing pattern may learn temporal dependence better.
- Current issue:
  - Adjoint (back-propagating gradient) is not physically causal.
  - (Biological implementation of forward sensitivity calculation may be implemented with a reference oscillation?)

Summary

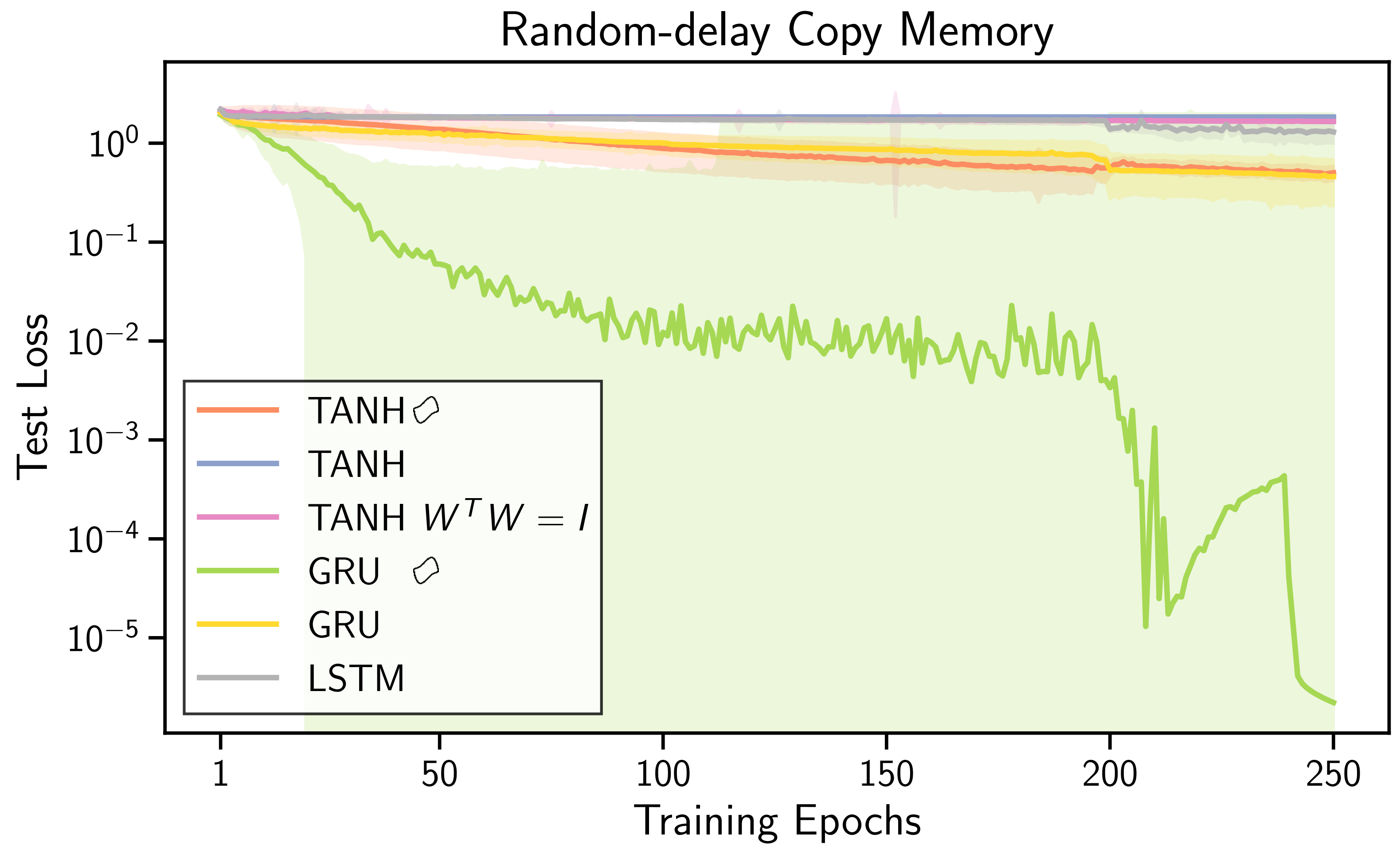
- Only topology of dynamics matters for EVGP.
- Non-vanishing/non-exploding gradients can be achieved for non-trivial systems. Robust solution may only be achieved with stable limit cycles.
- Stable limit cycle initialization is effective in solving long temporal memory tasks with RNNs.





sensitivity remains for infinite time

$$\mathbf{W}_{\text{init}} = \begin{bmatrix} \alpha_1 \begin{pmatrix} \cos(\theta_1) & -\sin(\theta_1) \\ \sin(\theta_1) & \cos(\theta_1) \end{pmatrix} & & \mathbf{0} \\ & \ddots & \\ \mathbf{0} & & \alpha_m \begin{pmatrix} \cos(\theta_m) & -\sin(\theta_m) \\ \sin(\theta_m) & \cos(\theta_m) \end{pmatrix} \end{bmatrix}$$



Computational And Theoretical
Neural Information Processing (CATNIP) Lab

Piotr Sokol	Yuan Zhao (NIH/NIMH)
Ian Jordan	Josue Nassar
Ayesha Vermani	Logan Becker (UTAustin)
Matthew Dowling	David Hocker (NYU)
Tushar Arora	Diego Arribas
Ábel Ságodi	Eben Kadile (IGI TU Graz)
André Mendonça	Kathleen Esfanany (MIT)

https://catniplab.github.io/

Stony Brook University
NSF
NIH
Champalimaud Foundation

Special thanks to Dongsung Ben Huh, Rainer Engelken, Braden Brinkman, Giancarlo LaCamera, Yifan Sun