Dynamical mean field theory analysis of FORCE training in recurrent neural networks

Pierfrancesco Urbani

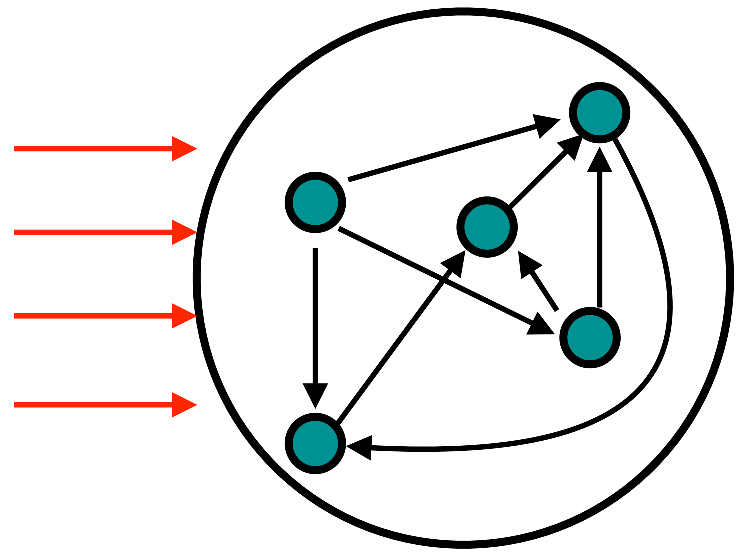
Université Paris-Saclay, CNRS, CEA,
Institut de Physique Théorique

Joint work with Samantha Fournier

J. Stat. Mech. (2023) 113301

Cargèse, 2024

Random RNN models



Sompolinsky, Crisanti, Sommers, PRL (1988)

$$\dot{x}_i(t) = -x_i(t) + \frac{g}{\sqrt{N}} \sum_j J_j^i r_j(t) + H_i(t)$$

$$r_i(t) = \phi(x_i(t)) \qquad \text{Ex: } \phi(x) = \tanh(x)$$

Random (asymmetric) synaptic connections:

$$J_i^j \sim \mathcal{N}(0,1)$$

$$J_i^j, J_j^i \quad \text{i.i.d}$$

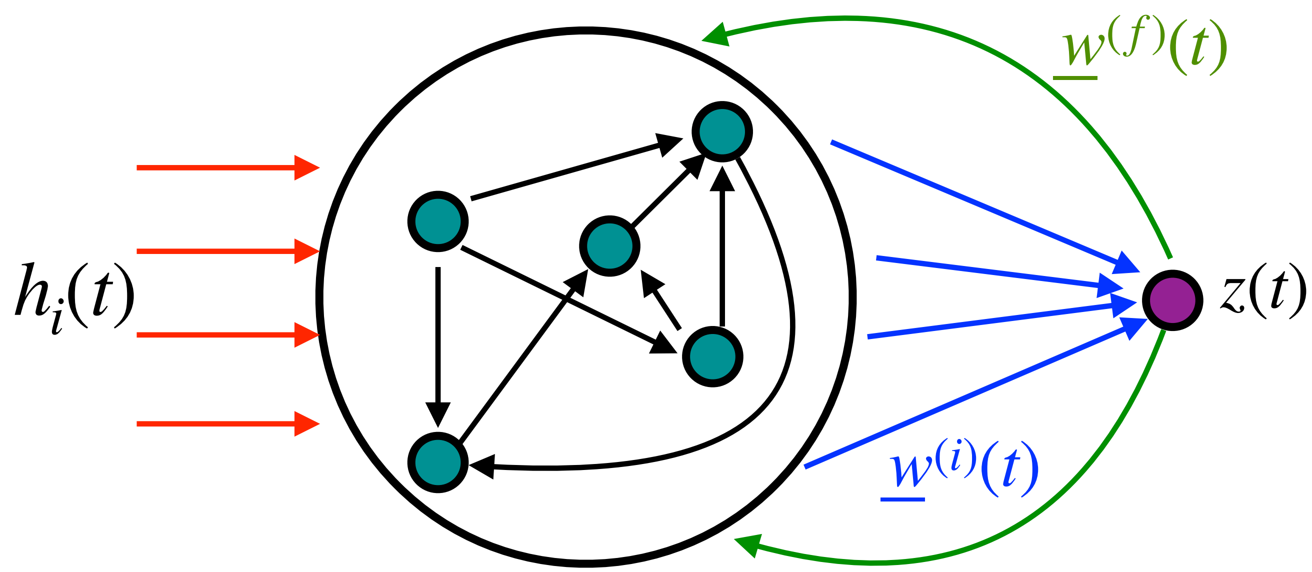
fixed-point to chaos transition driven by the strength of interactions

$$\phi(x): \quad \phi(0) = 0 \quad \phi'(0) = 1$$

$x = 0$ is a fixed point

At $g = 1$ this fixed point is destabilized and the system becomes chaotic

Training RNN models



output: *linear* readout unit

$$\dot{x}_i(t) = -x_i(t) + \frac{g}{\sqrt{N}} \sum_j J^i_j r_j(t) + H_i(t)$$

$$H_i(t) = h_i(t) + z(t) w_i^{(f)}$$

$$r_i(t) = \phi(x_i(t))$$

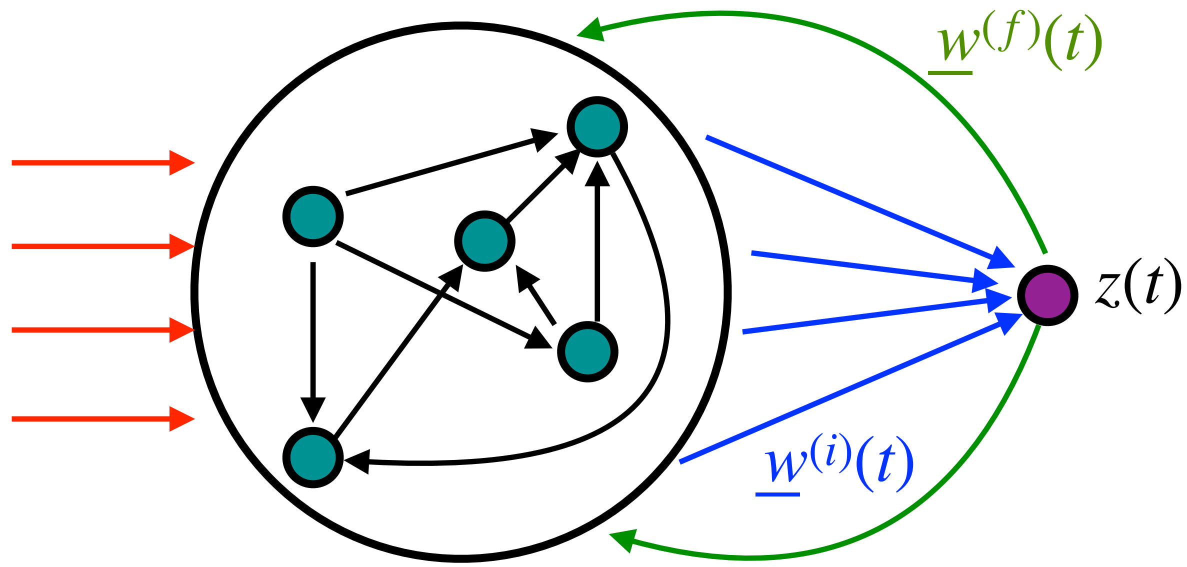
$$z(t) = \underline{w}^{(i)} \cdot \underline{r}(t)$$

TASK : We want to train the weights $\underline{w}^{(i)}$ **such that the output** $z(t) = f(t)$
(Sussillo & Abbott, *Neuron*, 2009)

Two situations:

1. No feedback weights $=>$ ***easy (Gradient descent)***
2. With feedback weights $\underline{w}^{(f)}$ $=>$ ***hard***

Training Algorithms



What we are interested in:

1. Sussillo & Abbott: FORCE

2. Node Perturbation, elegibility traces…

3. Miconi's rule (2017)

1) The current understanding the performance of algorithms is empirical: no control on timescales of learning, *stability with the system size*… this talk: theory of FORCE

2) How the network structure and external inputs shape learning dynamics

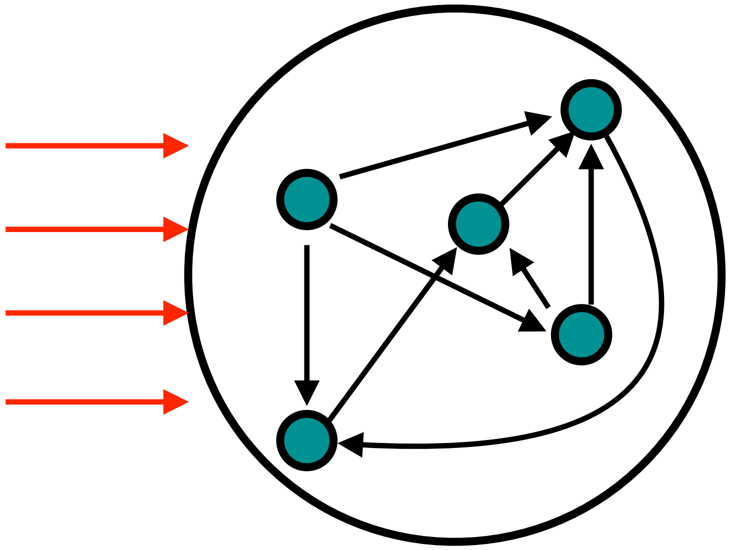
3) Biological plausibility

RNN models *vs* high-d chaotic systems

$$\dot{x}_i(t) = -x_i(t) + \frac{g}{\sqrt{N}} \sum_j J^i_j r_j(t) + H_i(t)$$

$$r_i(t) = \phi(x_i(t))$$

The nonlinearity makes theory complicated and therefore if we want to study training we need to simply the model



Simple Idea: Replace the non-linear dynamical system with

$$\dot{x}_i(t) = -\mu(t)x_i(t) + \frac{g}{N} \sum_{jk} J^i_{jk} x_j(t) x_k(t) + H_i(t)$$

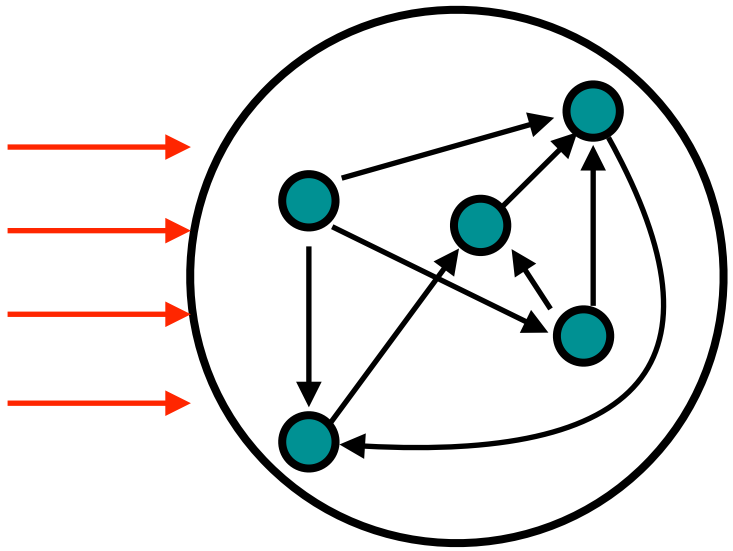
- The main difference is in the non-linear coupling between degrees of freedom.
- In this case the coupling is still non-linear but multi-body.
- Seems not good from biological point of view

We need to show that the new dynamical system has the same properties of the standard RNN.

Hebbian plasticity

1. We know it has a “fixed_point-to-chaotic” transition for large non-linearity.

2. Following Clark & Abbott (2023) we introduce an ***Hebbian plasticity*** term



$$\dot{x}_i(t) = -\mu(t)x_i(t) + \frac{g}{N}\sum_{jk} J^i_{jk} x_j(t)x_k(t) + H_i(t)$$

$$H_i(t) = \sum_j A_{ij}x_j(t)$$

$$p\dot{A}_{ij}(t) = -A_{ij} + \frac{k}{N}x_i(t)x_j(t)$$

Hebbian plasticity: k is a control parameter

Protocol (Clark, Abbott 2023 for standard RNN)

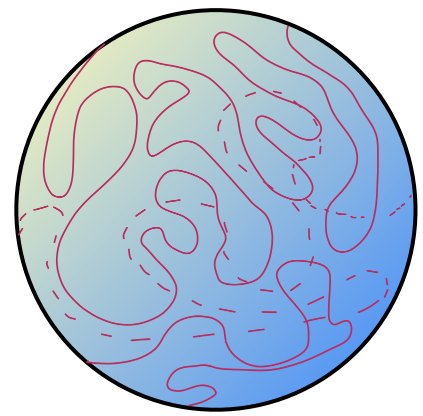
- Run the dynamics with plasticity for a certain amount of time
- At $t = t^*$ stop the plasticity dynamics. The matrix A is fixed to the value it had at t^*
- Characterize the resulting dynamics after synaptic freezing.

Hebbian plasticity

Possible outcomes for the dynamics of the system after the freezing time t^*

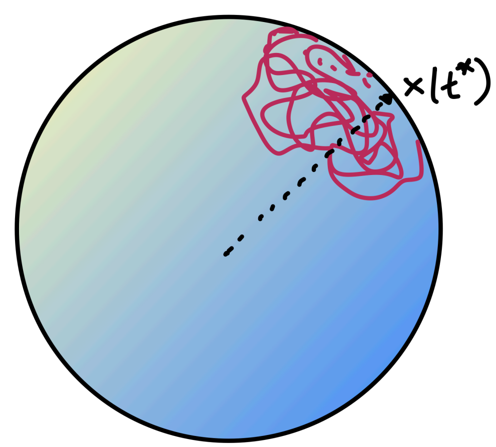
Increasing k (Hebbian forcing)

Chaos



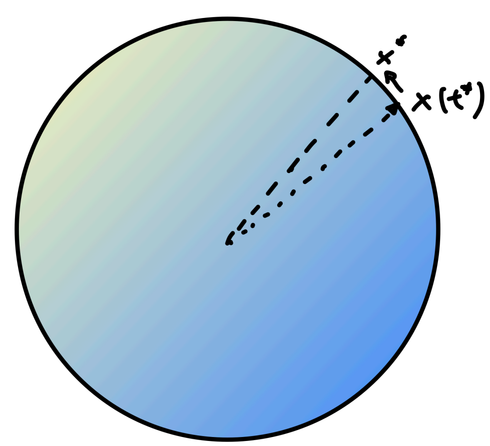
The system keeps exploring in a chaotic way *all the phase space.*

Semi-Freezable Chaos



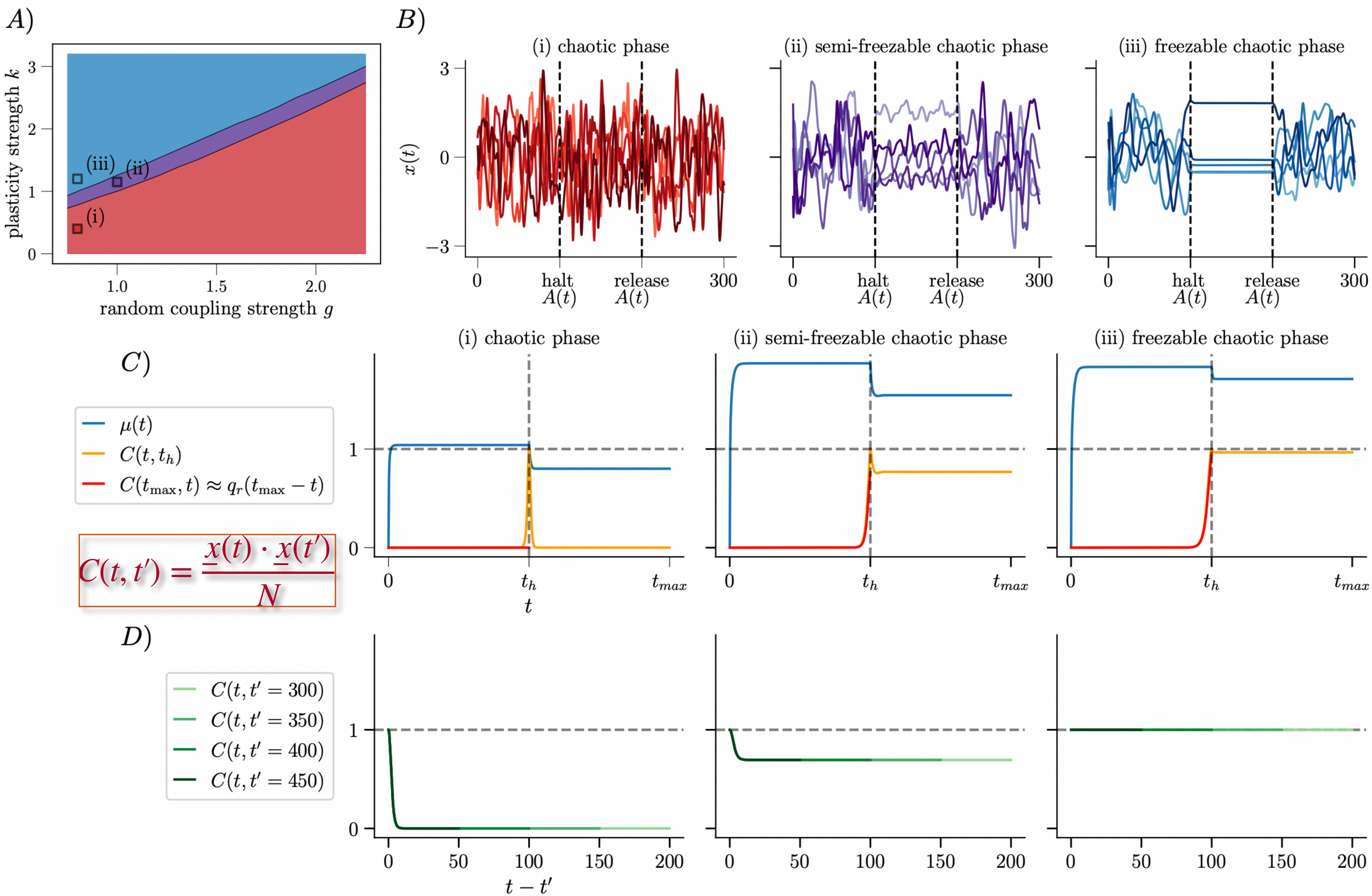
The system explores chaotically a phase space region around the point where the system was at freezing. *Memory & chaos.*

Freezable Chaos



The system goes to a fixed point close to the point where it was at the freezing time. *Memory & Fixed Point.*

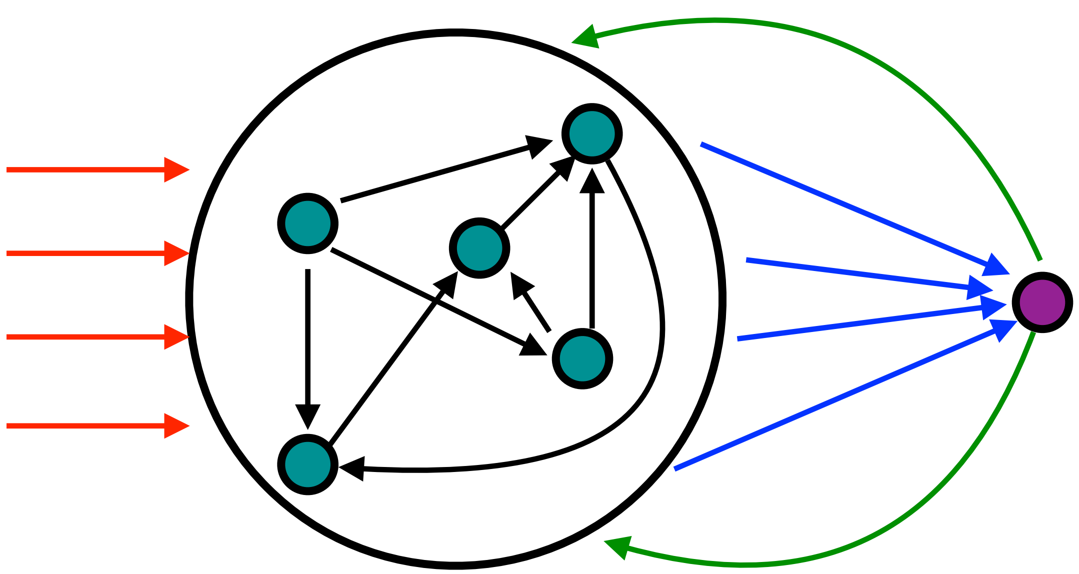
<u>Hebbian plasticity can be used to tune the level of chaos</u>



$$C(t,t') = \frac{\underline{x}(t) \cdot \underline{x}(t')}{N}$$

✔ We get exactly the same results as for standard RNN models

FORCE Training



$$\dot{x}_i(t) = -x_i(t) + \frac{g}{\sqrt{N}} \sum_j J^i_j r_j(t) + H_i(t)$$

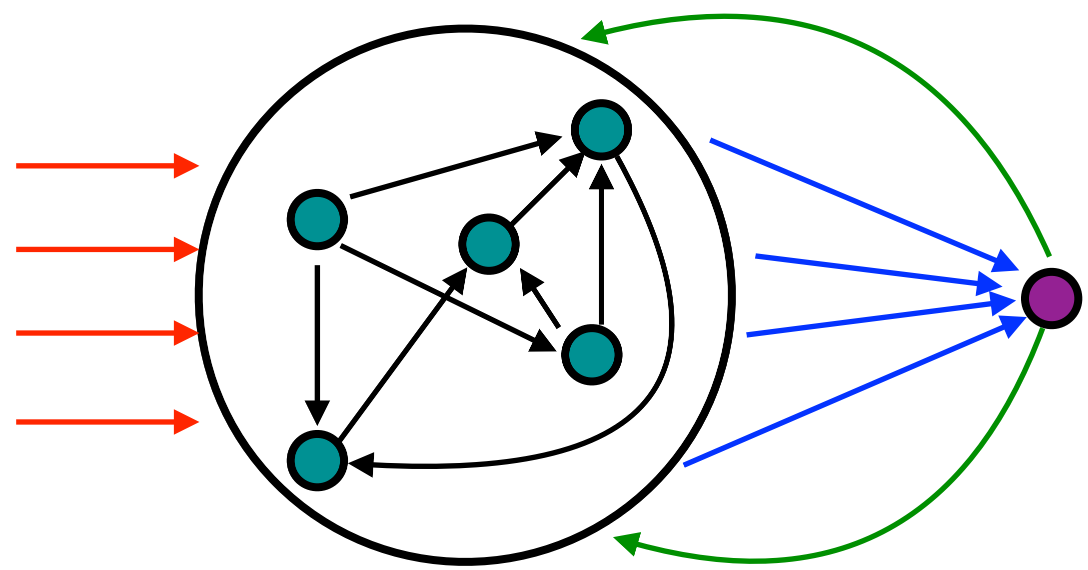
$$r_i(t) = \phi(x_i(t))$$

TASK: find the linear readout weights such that the output of the readout neuron is a desired periodic function

$$z(t) = f(t)$$

Feedback loops make the dynamics unstable (exploding and/or vanishing gradients)!

FORCE Training



First Order Reduced and Controlled Error (**FORCE**) (Sussillo, Abbott, 2009)

Main Idea:
- If the readout weights can be time dependent, one can produce the desired output instantaneously
- Try to damp out the dynamics of the readout weights while keeping small the error

Practically
- The feedback loop on the dynamical system drives it towards an attractor.
- The algorithms looks for configurations of the readout weights that keep close to the attractor while slowing down their dynamics

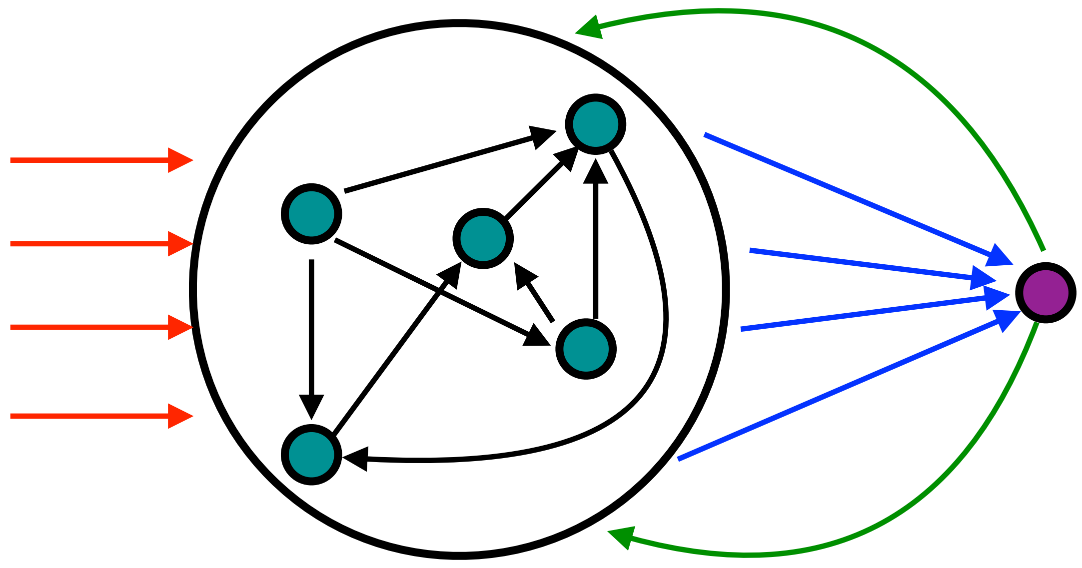
1. *Can be adapted to perform many tasks and can be also implemented in spiking recurrent neural networks (Nicola, Clopath, 2017)*
2. *Studied numerically (simulations): no theory*

FORCE Training

$$\dot{x}_i(t) = -\mu(t)x_i(t) + \frac{g}{N}\sum_{jk} J^i_{jk} x_j(t) x_k(t) + H_i(t)$$

$$H_i(t) = z(t)$$

$$z(t) = \frac{1}{N}\underline{w}(t)\cdot\underline{x}(t)$$



We adapt **FORCE** to this dynamical system

$$z^+(t) = \frac{1}{N}\underline{w}(t)\cdot\underline{x}(t+\mathrm{d}t)$$

$$e_-(t) = z^+(t-\mathrm{d}t) - f(t)$$

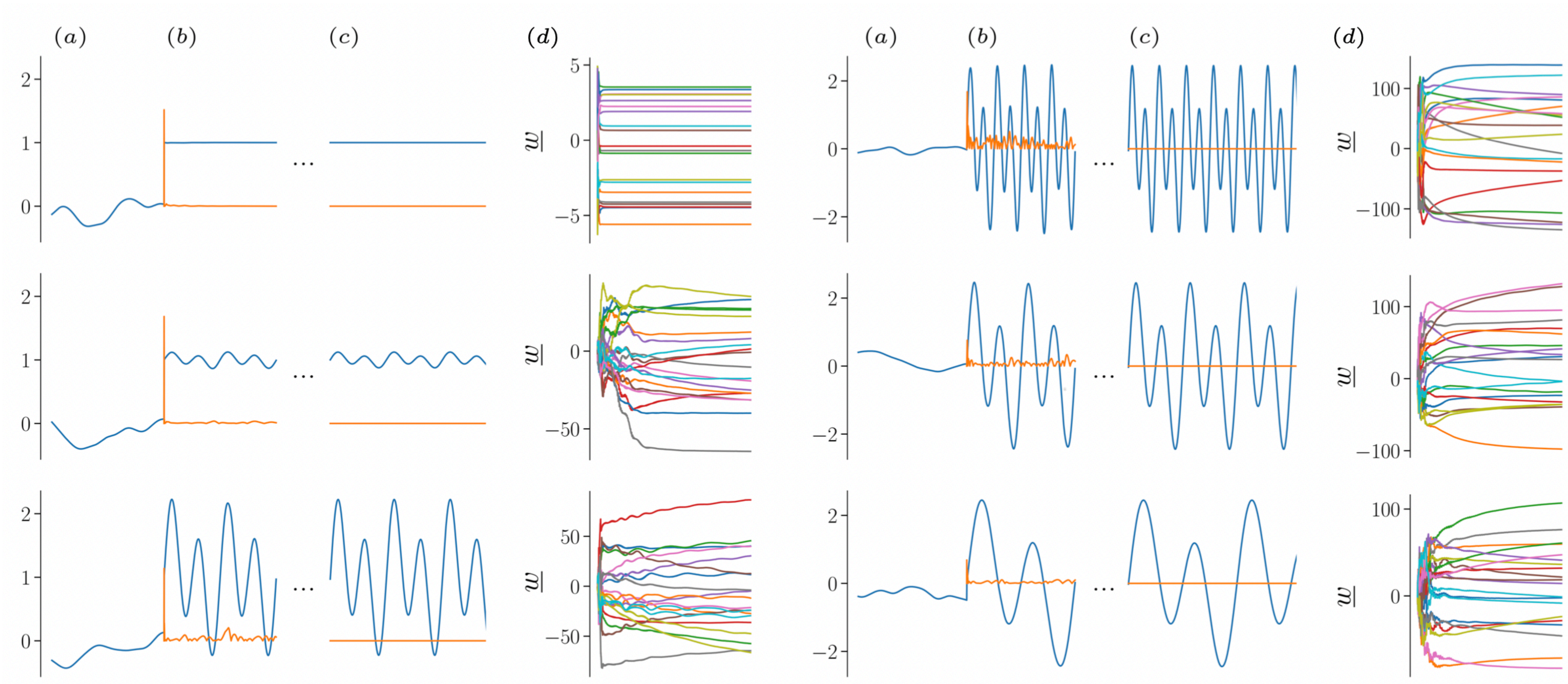
$$\underline{w}(t+\mathrm{d}t) = \underline{w}(t) - e_-(t+\mathrm{d}t)P(t+\mathrm{d}t)\underline{x}(t+\mathrm{d}t)$$

$$P(0) = \frac{1}{\alpha}\mathbf{1}$$

$$P(t+\mathrm{d}t) = P(t) - \frac{1}{N}\frac{P(t)\underline{x}(t+\mathrm{d}t)\underline{x}(t+\mathrm{d}t)^T P(t)}{1+\frac{1}{N}\underline{x}(t+\mathrm{d}t)^T P(t)\underline{x}(t+\mathrm{d}t)}$$

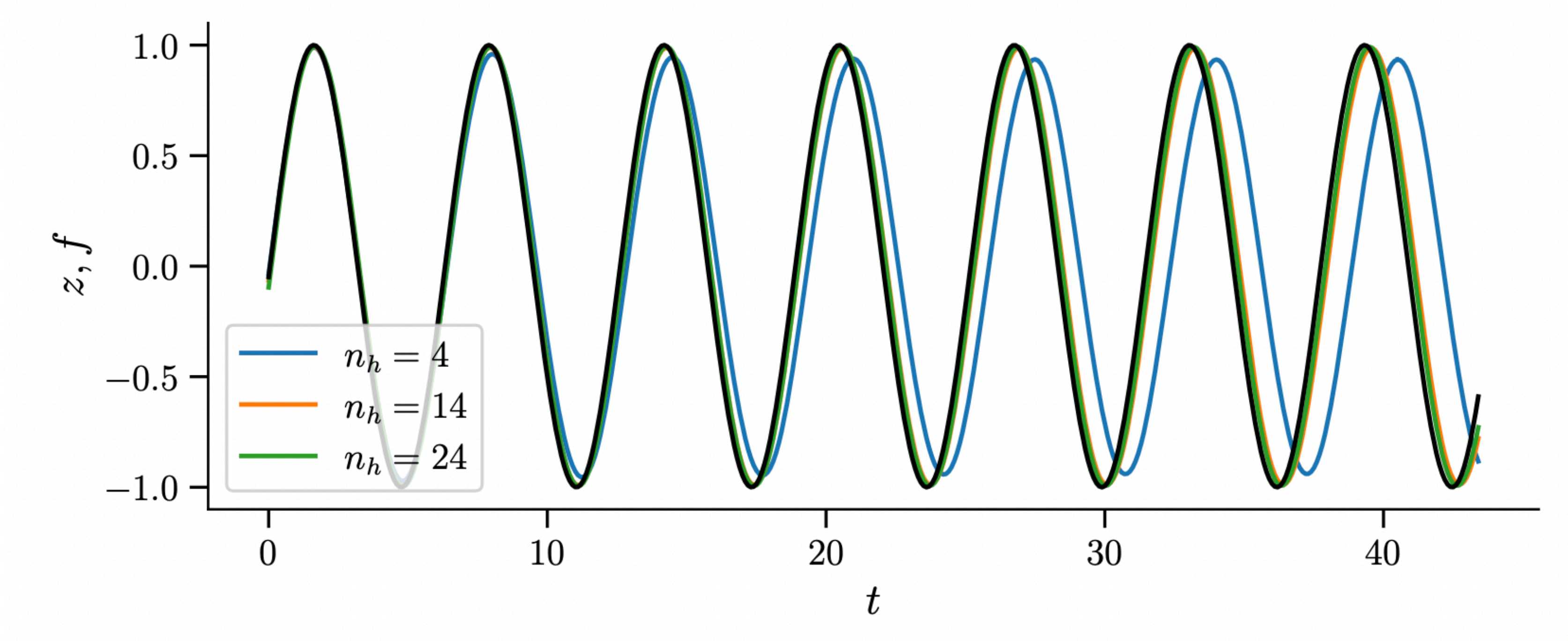
FORCE Training high-d chaotic system

Numerical simulations

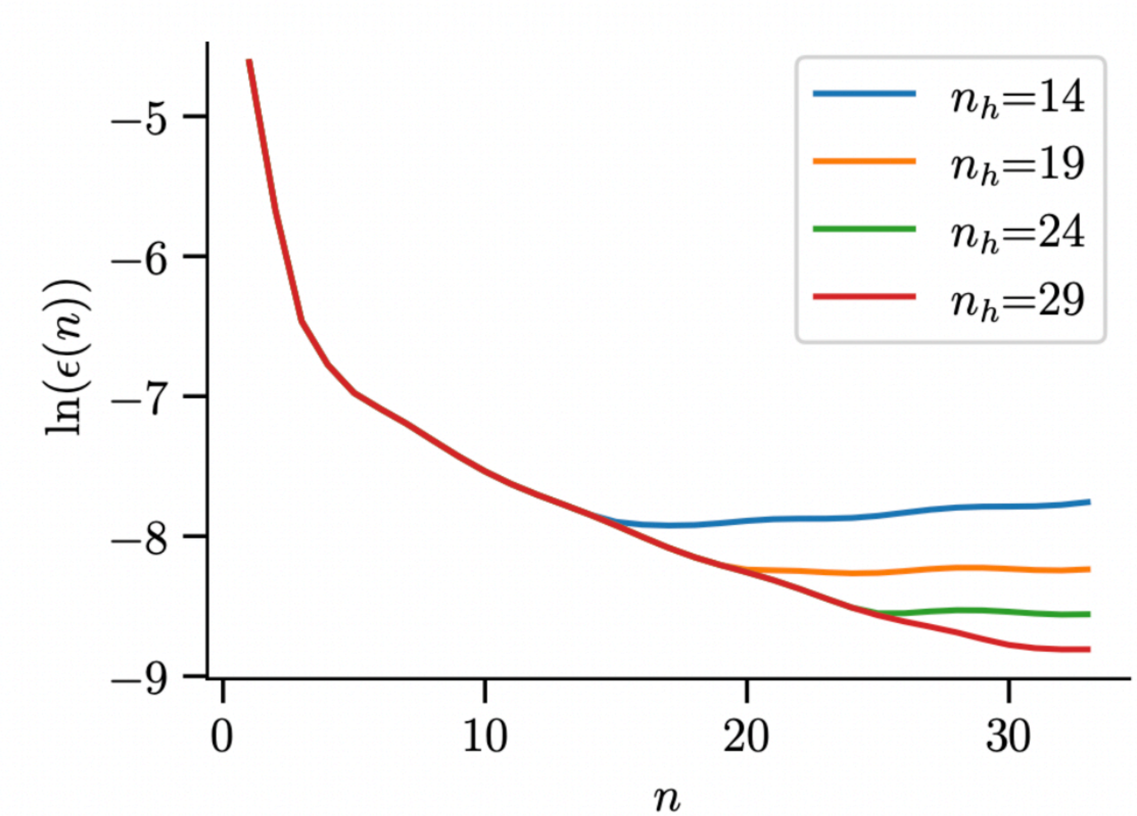


Next: we can study the training dynamics via DMFT => Results for large N

DMFT of FORCE Training



Error during training stopping
at different training epochs



Perspectives

We can track the dynamics, and have access to the statistics of the synaptic weights.

- Stability of the learned dynamical attractor: Lyapunov exponents
- The **space of synaptic weights**? Geometry? Connectivity? Statistics?
- Is it possible to “pack” multiple attractors? Limit of capacity?
- The level of chaos is essential for the performances. Can we mix Hebbian plasticity and FORCE training?
- **More biological plausibility**
- **Everything is general here: “training”/*control* other chaotic systems (ecolo/econo/socio)?**