Fisher Memory of linear Wigner Echo State Networks

Peter Tino
School of Computer Science
The University of Birmingham
UK
Time series data - state space model

Time series ..., \( s(t - 2), s(t - 1), s(t), s(t + 1), ... \)
\( s(t) \in S \)

"State" at time \( t \), \( x(t) \in \mathcal{X} \),
codes the entire history of sequence elements up to \( t \): ..., \( s(t - 1), s(t) \)

\[
\begin{align*}
x(t) & = f(x(t - 1), s(t)), \\
y(t) & = h(x(t))
\end{align*}
\]

- \( f(\cdot) \) - state-transition function
- \( y(t) \) - output at time \( t \) - readout from the state space.
- \( h(\cdot) \) - output readout function
State space model - an example

1 a b
\[ \rightarrow \]

2 a b
\[ \downarrow 1 \]
\[ \downarrow 0 \]
Learning the state space model

- parameterize the state transition and readout functions.
- "learn" their parameters.
Parameterized state space model - RNN (ESN)

$N$-dimensional continuous state space $\mathbf{x}(t) \in [-1, 1]^N$

$$
\begin{align*}
\mathbf{x}(t) &= f(\mathbf{x}(t - 1), \mathbf{s}(t)) \\
&= \sigma(R\mathbf{x}(t - 1) + V\mathbf{s}(t) + \mathbf{b}), \\
\mathbf{y}(t) &= h(\mathbf{x}(t)) \\
&= W\mathbf{x}(t) + \mathbf{a}
\end{align*}
$$
"General purpose” state space model

\[
x(t) = \sigma(Rx(t - 1) + Vs(t) + b),
\]

\[
y(t) = Wx(t) + a
\]

Fix the DS part to a high-dimensional DS - often randomized construction

General purpose state space organization (dynamical filter) that allows for rich and redundant representations of input histories.

Learn the (static) linear readout only. Can be done efficiently.
How useful is the "general purpose" state space?

Memory capacity - [Jaeger] Measures richness of the DS state space representations.
How much information about the past is stored in the states?

Make the memory task difficult: DS is driven by a univariate i.i.d. input signal \( s(t) \).

For a given delay \( k > 0 \), find optimal readout for recalling \( s(t - k) \) after seeing the input stream \( ...s(t - 1)s(t) \) up to time \( t \).

Goodness of recall - squared correlation coefficient between the desired output \( s(t - k) \) and the observed network output \( y(t) \).
Memory capacity of dynamical systems

\[ MC_k = \frac{Cov^2(s(t - k), y(t))}{Var(s(t)) \cdot Var(y(t))}, \]

where \( Cov \) denotes the covariance and \( Var \) the variance operators. Memory Capacity is then given by:

\[ MC = \sum_{k=1}^{\infty} MC_k. \]

Theorem (Jaeger): MC of every smooth \( N \)-dimensional DS is \( \leq N \).

Can be achieved with a linear DS!
Fisher memory of linear dynamical systems

\[ x(t) = s(t)v + Wx(t-1) + z(t), \]

where \( z(t) \) are zero-mean dynamic noise terms.

The state coupling matrix \( W \) is typically scaled to a prescribed spectral radius \(< 1 \) (ESP).

If \( W \) is a random variable, we assume that the parameters of the distribution over \( W \) are set so that asymptotically, almost surely, \( W \) is a contractive linear operator.
Dynamical systems corrupted by a memoryless zero-mean spherical Gaussian i.i.d. dynamic noise $z(t)$ with variance $\sigma^2 I$.

Given an input driving stream $s(\ldots t) = \ldots s(t-2) s(t-1) s(t)$, the input-conditional state distribution

$$p(x(t) \mid \ldots s(t-2) s(t-1) s(t))$$

is a Gaussian with covariance

$$C = \epsilon \sum_{\ell=0}^{\infty} W^\ell (W^T)^\ell.$$
Sensitivities of \( p(x(t) | s(\ldots t)) \) with respect to small perturbations in the input driving stream \( s(\ldots t) \) are collected in the Fisher memory matrix \( F \),

\[
F_{k,l}(s(\ldots t)) = -\mathbb{E}_{p(x(t) | s(\ldots t))} \left[ \frac{\partial^2}{\partial s(t-k) \partial s(t-l)} \log p(x(t) | s(\ldots t)) \right]
\]

Diagonal elements \( J_N(k) = F_{k,k}(s(\ldots t)) \) quantify the information that the state distribution \( p(x(t) | s(\ldots t)) \) retains about a change entering the network \( k > 0 \) time steps in the past.
Fisher memory of dynamical systems

The collection of terms \( \{ J_N(k) \}_{k=1}^{\infty} \) - Fisher memory curve:

\[
J_N(k; W, v) = v^T (W^T)^k C^{-1} W^k v.
\]

Global memory quantification - Fisher memory:

\[
J_N(W, v) = \sum_{k=1}^{\infty} J_N(k; W, v).
\]
Wigner dynamic coupling

\[ X_N - \text{a random symmetric } N \times N \text{ matrix} \]

“Upper triangular” off-diagonal elements \( X_{i,j}, 1 \leq i < j \leq N \) distributed i.i.d. with zero mean and finite moments (variance \( \sigma^2_o > 0 \)).

Diagonal elements \( X_{i,i}, 1 \leq i \leq N \) are distributed i.i.d. with a zero-mean distribution of finite moments and variance \( \sigma^2_d > 0 \).

The elements below the diagonal are copies of their symmetric counterparts: for \( 1 \leq j < i \leq N \), \( X_{i,j} = X_{j,i} \).
Wigner dynamic coupling

Asymptotic properties of such matrices have been widely studied, in particular the convergence of eigenvalues, as the dimension $N \rightarrow \infty$.

In the general case, scaling down of random matrices is necessary to ensure convergence of their spectral properties:

$$W_N = \frac{1}{\sqrt{N}} X_N.$$
Fisher memory of Wigner DS

Study asymptotic properties of the normalized Fisher memory as the state space dimensionality \( N \) grows.

What kind of input coupling \( \mathbf{v} \) is needed to maximize its expectation?

Observation: Increasing state space dimension \( N \) will increase the amount of memory that can be usefully captured by the dynamical system.

Remove this bias - \textit{normalized Fisher memory} measures the amount of memory realizable by the dynamical system \textit{per state space dimension}:

\[
\bar{J}_N(\mathbf{W}_N, \mathbf{v}) = \frac{1}{N} J_N(\mathbf{W}_N, \mathbf{v}).
\]
Observation: As the state space dimensionality $N$ grows, so does the input weight dimensionality.

Keeping the input weight norm constant while increasing the state space dimensionality would result in diminishing individual weights.

Normalize the scales, so that asymptotic statements can be made: require that the input weights live on $(N - 1)$-dimensional hypersphere, $\mathbf{v} \in S_{N-1}(\sqrt{N})$, where for $r > 0$,

$$S_{N-1}(r) = \{ \mathbf{x} \in \mathbb{R}^N | \|\mathbf{x}\|_2 = r \}.$$
Theorem 1: Consider a sequence of Wigner dynamical systems with couplings \( \{W_N\}_{N>1} \). The maximum normalized Fisher memory is attained when for every realization of Wigner coupling \( W_N \), the input weights \( v \) are collinear with the dominant eigenvector. In that case, as \( N \to \infty \), almost surely,

\[
\tilde{J}_N(W_N, v) \to 4 \frac{\sigma^2_o}{\sigma^2}.
\]
Lessons learnt

1. As the system size grows, the contribution of self-coupling of the states (self-loops on reservoir units in ESN) to the normalized Fisher memory is negligible (no $\sigma_d^2$ term);

2. Stronger dynamic noise (larger $\sigma^2$) degrades memory that can be usefully stored in the dynamical system;

3. Stronger dynamical coupling between different state space units (larger $\sigma_o^2$) enhances memory (but see the need for ESP!).

4. Theoretical support for the observation that randomized DS need to be tuned to the edge of stability.
4σ̄_o^2/σ^2 - asymptotic upper bound on normalized Fisher memory of Wigner ESNs.

\( \bar{J}_N(W_N, v) \) can be made vanishingly small by making the input weights v collinear with the least significant eigenvector of \( W_N \) (semicircular law of eigenvalue distribution for Wigner matrices).

To what degree will the Fisher memory degrade if instead of the dominant eigenvector the input weights are made collinear with the sum of eigenvectors of \( W_N \)?
Theorem 2: Consider a sequence of Wigner dynamical systems with couplings \( \{W_N\}_{N>1} \). For every realization of Wigner coupling \( W_N \), let the input weights \( v \) be collinear with the sum of eigenvectors of \( W_N \). Then, as \( N \to \infty \), for the expected normalized Fisher memory we have,

\[
E_{W_N}[\tilde{J}_N(W_N, v)] \to \frac{\sigma_o^2}{\sigma^2}.
\]