Learning Temporal Structures in Time Series Data
Non-Autonomous Dynamical Systems View

Peter Tino
School of Computer Science
University of Birmingham
UK

Alan Turing Institute Fellowship

The Alan Turing Institute

Machine Learning and Dynamical Systems Interest Group

Special thanks to ...

- Barbara Hammer
- Lee Giles
- Alessio Micheli
- Juan-Pablo Ortega
- Jochen Steil
- Pietro Verzelli
- Lorenzo Livi
- Simon Fong
- Ali Rodan
- Alessandro Sperdutti
- Luba Benuskova
- Igor Farkas
- Lyudmila Grigoryeva
- Whit Tabor
- ...

Parametrized Autonomous Dynamical Systems

The "core" dynamics: $f : \mathcal{X} \times \Theta_f \rightarrow \mathcal{X}$

$$\boldsymbol{x}(t) = f(\boldsymbol{x}(t-1);\ \theta_f)$$

Reading out from state space $h : \mathcal{X} \times \Theta_h \rightarrow \mathcal{Y}$

$$\boldsymbol{y}(t) = h(\boldsymbol{x}(t);\ \theta_h)$$

Other reformulations

Core dynamics:

$$\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t-1);\ \theta_f)$$

$$p(\boldsymbol{x}(t)|\ \boldsymbol{x}(t-1);\ \theta_f)$$

Reading out from state space:

$$\boldsymbol{y}(t) = h(\boldsymbol{x}(t);\ \theta_h)$$

$$p(\boldsymbol{y}(t)|\boldsymbol{x}(t);\ \theta_h)$$

Non-Autonomous Dynamical Systems

The "core" dynamics: $f : \mathcal{X} \times \mathcal{U} \times \Theta_f \rightarrow \mathcal{X}$

$$\boldsymbol{x}(t) = f(\boldsymbol{x}(t-1), \boldsymbol{u}(t);\ \theta_f)$$

Reading out from the state space $h : \mathcal{X} \times \Theta_h \rightarrow \mathcal{Y}$

$$\boldsymbol{y}(t) = h(\boldsymbol{x}(t);\ \theta_h)$$

Other reformulations

Core dynamics:

$$\dot{\boldsymbol{x}}(t) = f(\boldsymbol{x}(t-1), \boldsymbol{u}(t);\ \theta_f)$$

$$p(\boldsymbol{x}(t)|\ \boldsymbol{x}(t-1), \boldsymbol{u}(t);\ \theta_f)$$

Reading out from state space:

$$\boldsymbol{y}(t) = h(\boldsymbol{x}(t);\ \theta_h)$$

$$p(\boldsymbol{y}(t)|\ \boldsymbol{x}(t);\ \theta_h)$$

ML - Learning from time series data

We observe an "input time series" $\{\boldsymbol{u}(t)\}_t$, $\boldsymbol{u}(t) \in \mathcal{U}$.

At certain times t_i we are asked to produce responses $\{\boldsymbol{y}(t_i)\}_i$, $\boldsymbol{y}(t_i) \in \mathcal{Y}$, based on what we have observed till the response times t_i, $\{\boldsymbol{u}(t) | \ t \leq t_i\}$.

How to organize learning in this case? We need to capture temporal dependencies among the inputs.

How to represent histories of inputs $\{\boldsymbol{u}(t) | \ t \leq t_i\}$ given that they are of different length?

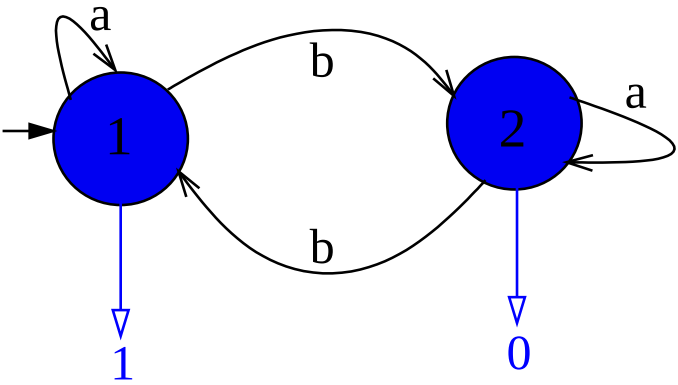
Key Notion - **Information Processing State**

IPS at time t codes information from the entire history of time series items seen up to time t that is needed in order to perform a given task.

IPS - Representation Learning in ML

- States - equivalence classes over temporal sequences - e.g. in FSA, Markov chains, PST etc.
- States - emerge naturally through the information bottleneck method (N. Tishby) on temporal data and an associated task on it.

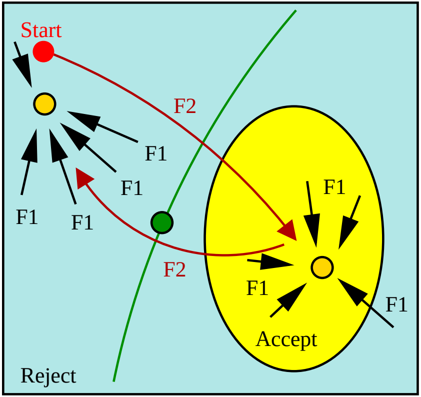
A familiar state space model in Computer Science



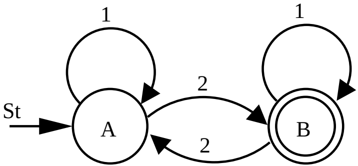
Simple example - IFS as attractive sets

To latch a piece of information for a potentially unbounded number of time steps we need attractive sets.

Information latching problem



FSM extraction



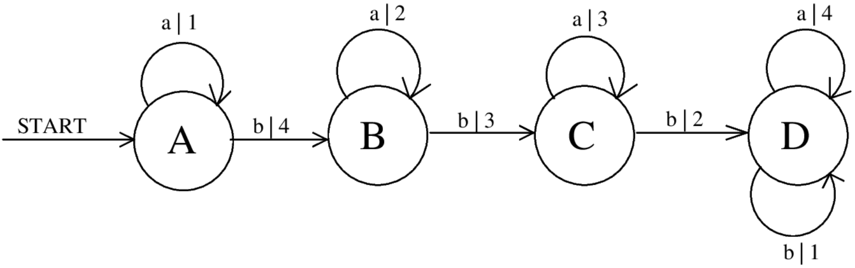
Grammatical:

all strings containg odd number of 2's

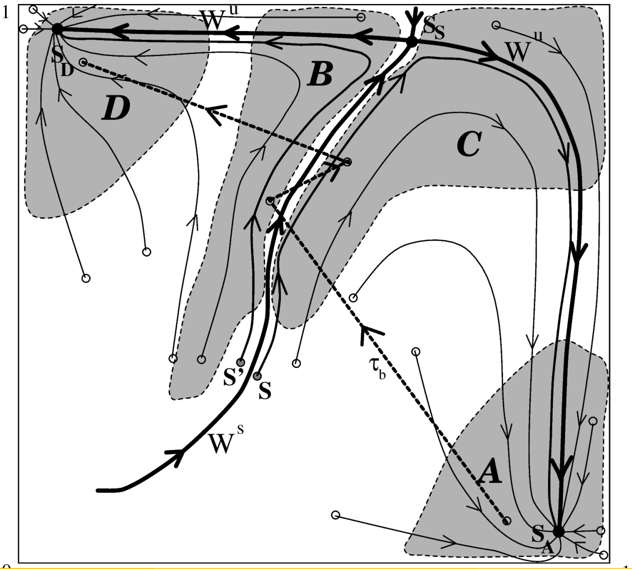
$\boldsymbol{x}(t) = F_1(\boldsymbol{x}(t-1)) = f(\boldsymbol{x}(t-1), \boldsymbol{u}(t) = 1)$

Attractive sets may be difficult to learn

Need to bifurcate to several attractive sets for f_a.



Cheating along stable manifolds of saddle type equilibria



Do we need to learn IPS? **"General purpose" IPS**

Inspiration from information theory ("general purpose" information sources) - Observable Markovian IPS.

Simplest construction of IPS - based on concentrating on the very recent (finite) past,
e.g. definite memory machines, finite memory machines, Markov models

Example:
Only the last 5 input symbols matter when reasoning about what symbol comes next

... 1 2 1 1 2 3 2 1 2 1 2 4 3 2 1 1
... 2 1 1 1 1 3 4 4 4 4 1 4 3 2 1 1
... 3 3 2 2 1 2 1 2 2 1 3 4 3 2 1 1

All three sequences belong to the same IPS "43211"

Probabilistic framework - Markov model (MM)

finite context-conditional next-symbol distributions

... 1 2 1 1 2 3 2 1 2 1 2 $\underline{4\ 3\ 2\ 1\ 1} \rightarrow$?
... 2 1 1 1 1 3 4 4 4 4 1 $\underline{4\ 3\ 2\ 1\ 1} \rightarrow$?
... 3 3 2 2 1 2 1 2 2 1 3 $\underline{4\ 3\ 2\ 1\ 1} \rightarrow$?

$P(s \mid 11111)$
$P(s \mid 11112)$
$P(s \mid 11113)$
...
$P(s \mid 11121)$
...
$P(s \mid 43211)$
...
$P(s \mid 44444)$

Variable memory length MM (VLMM)

Efficient implementation of potentially high-order MM

Save resources - make memory depth context dependent

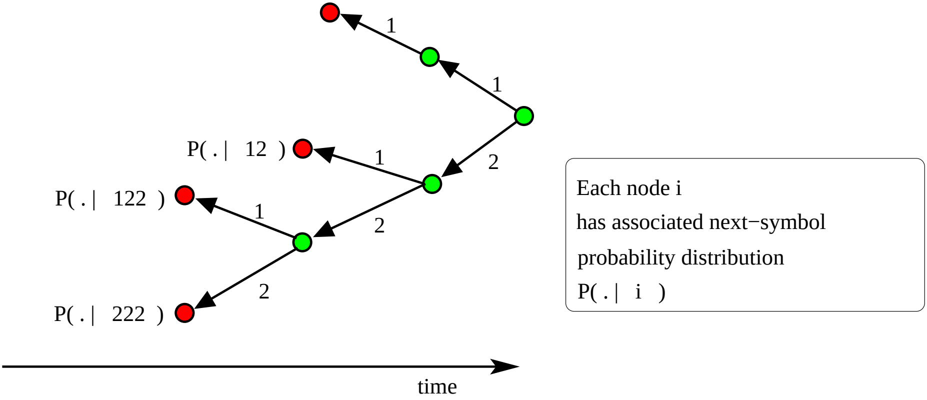
Use deep memory only when it is needed

Natural representation of IPS in form of Prediction Suffix Trees

Closely linked to the idea of universal simulation/approximation of information sources.

Prediction suffix tree (PST)

IFS are organized on nodes of PST
Traverse the tree from root towards leaves in reversed order
Complex and potentially time-consuming training



Contractive maps lead to Markovian IPS (1)

Given an input symbol $s \in \mathcal{U}$ at time t, and activations of recurrent units from the previous step, $\mathbf{x}(t-1)$,

$$\mathbf{x}(t) = F_s(\mathbf{x}(t-1)).$$

This notation can be extended to sequences over $\mathcal{U}$. The state reached after t time steps is

$$\mathbf{x}(t) = F_{s_1 \ldots s_t}(\mathbf{x}(0)).$$

If maps F_s are contractions with Lip. constant $0 < C < 1$,

$$\begin{aligned} \|F_{vq}(\mathbf{x}) - F_{wq}(\mathbf{x})\| &\leq C^{|q|} \cdot \|F_v(\mathbf{x}) - F_w(\mathbf{x})\| \\ &\leq C^{|q|} \cdot diam(\mathcal{X}). \end{aligned}$$

Contractive maps lead to Markovian IPS (2)

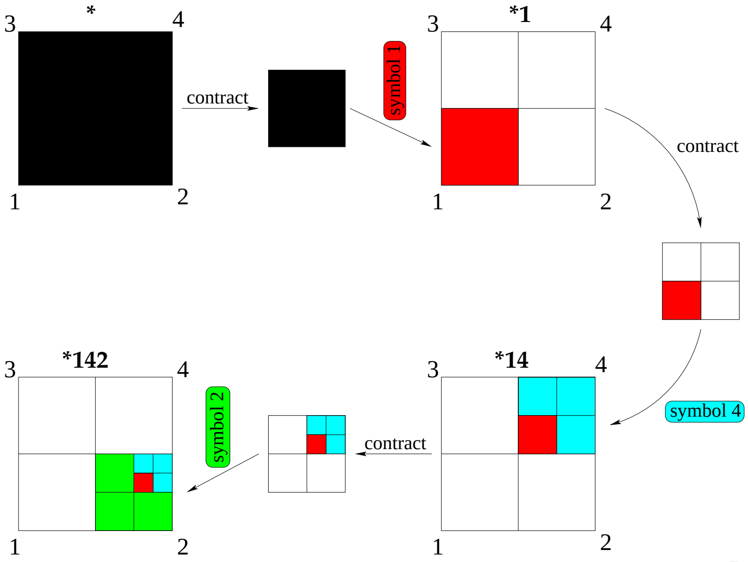
Markov representation learning

Sequences with shared recent histories of observed symbols have close state space representations.

The longer is the common suffix of two sequences, the closer are they mapped to each other.

Now we can add a simple output-generating readout on top of the fixed state dynamics!

Contractive maps lead to Markovian IPS (3)



Theoretical grounding (RNN vs DMM)

Theorem (Hammer & Tino)

- *Every RNN with contractive (fixed-input) transition functions can be approximated arbitrarily well on input sequences of unbounded length by a definite memory machine. Conversely,*
- *Every definite memory machine can be simulated by an RNN with contractive transition function.*

Hence, contractive RNNs (e.g. initialized with small weights) have **Markovian architectural bias**.

Learnability (PAC framework)

The architectural bias emphasizes one possible region of the weight space where generalization ability can be formally proved.

Standard RNN are not distribution independent learnable in the PAC sense if arbitrary precision and inputs are considered.

Theorem (Hammer & Tino)

- *Recurrent networks with contractive transition function with a fixed contraction parameter fulfil the distribution independent UCED property and so*
- *unlike general recurrent networks, are distribution independent PAC-learnable.*

State space complexity $\leftrightarrow$ sequence complexity

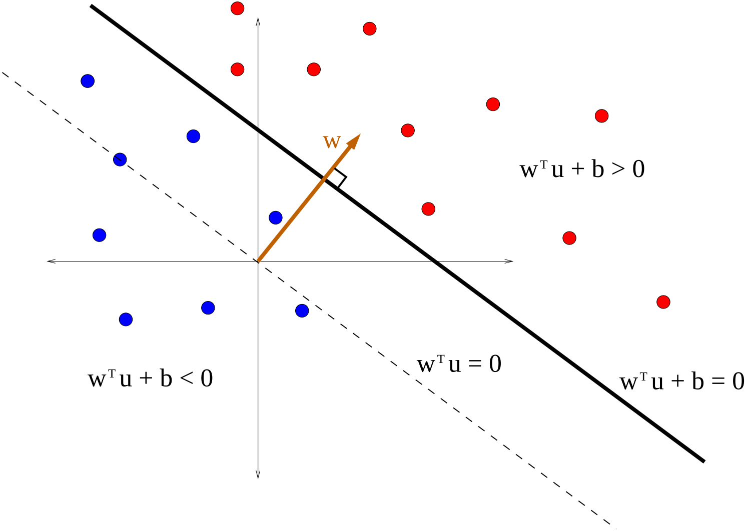
Natural "general purpose" state space organization:
Temporal complexity of the driving input time series (topological entropy) is directly reflected in the spatial complexity of the states (fractal dimension).

Theorem (Tino & Hammer)

Recurrent activations inside contractive RNN form fractals the dimension of which can be bounded by the scaled entropy of the underlying driving source. The scaling factors are fixed and are given by the RNN parameters.

Can be **extended to entropy spectra and multifractal spectra**.

"Learnt" representations - kernel feature spaces



Classification

$$y = sgn(\boldsymbol{w}^T \mathbf{u} + b)$$

often

$$\boldsymbol{w} = \sum_i \alpha_i \cdot \mathbf{u}_i$$

and so

$$\begin{aligned} y &= sgn\left(\sum_i \alpha_i \cdot \mathbf{u}_i^T \mathbf{u} + b \right) \\ &= sgn\left(\sum_i \alpha_i \cdot \langle \mathbf{u}_i, \mathbf{u} \rangle + b \right) \end{aligned}$$

Classification - feature space, kernel trick

Embed the input space in a high-dimensional feature space:

$$\mathbf{u} \mapsto \phi(\mathbf{u}) \in \mathcal{H}$$

$$\begin{aligned} y &= sgn\left(\sum_i \alpha_i \cdot \langle \phi(\mathbf{u}_i), \phi(\mathbf{u}) \rangle_{\mathcal{H}} + b\right) \\ &= sgn\left(\sum_i \alpha_i \cdot K(\mathbf{u}_i, \mathbf{u}) + b\right) \end{aligned}$$

Kernel $K(\cdot,\cdot)$ can allow us to implicitly work in a "rich" Hilbert space $\mathcal{H}$, even though it can be itself parametrized by just few free parameters!

State space as a temporal feature space

View the state as a feature representation of previous inputs:

$$
\begin{aligned}
\phi(\ldots u(t-3), u(t-2), u(t-1)) &= \boldsymbol{x}(t-1) \\
\phi(\ldots u(t-2), u(t-1), u(t)) &= \boldsymbol{x}(t) \\
&= f(\boldsymbol{x}(t-1), u(t))
\end{aligned}
$$

$$
\begin{aligned}
y(t) &= h(\boldsymbol{x}(\boldsymbol{t})) \\
&= \sum_i \alpha_i \cdot \langle \mathbf{x}(t_i), \mathbf{x}(t) \rangle + b \\
&= \sum_i \alpha_i \cdot \langle \phi(\ldots, u(t_i - 1), u(t_i)), \phi(\ldots, u(t-1), u(t)) \rangle + b
\end{aligned}
$$

Fading memory dynamical system

Input stream: $...u(t-2), u(t-1), u(t),\ u(i) \in \mathbb{R}$

$$
\begin{aligned}
\mathbf{x}(t) &= f(\mathbf{x}(t-1), u(t)) \\
&= \mathbf{W}\ \mathbf{x}(t-1) + u(t)\ \mathbf{w}
\end{aligned}
$$

$\mathbf{W} \in \mathbb{R}^{N\times N}$ is a weight matrix providing the dynamical coupling, $\mathbf{w} \in \mathbb{R}^N$ is the input-to-state coupling vector.

Contractive dynamics: $\sigma_{max}(\mathbf{W}) < 1$

Understanding the temporal kernel

$$\begin{aligned} K(\mathbf{u},\mathbf{v}) &= \langle \mathbf{u},\mathbf{v}\rangle_{\mathbf{Q}} \\ &= \mathbf{u}^\top \mathbf{Q}\ \mathbf{v} \\ &= \mathbf{u}^\top \mathbf{M}\Lambda\mathbf{M}^\top \mathbf{v} \\ &= \left\langle \Lambda^{\frac{1}{2}}\mathbf{M}^\top\mathbf{u}, \Lambda^{\frac{1}{2}}\mathbf{M}^\top\mathbf{v} \right\rangle \end{aligned}$$

Eigen-analysis of the positive (semi-)definite "metric tensor" $\mathbf{Q}$

- **feature space "motifs":** eigenvectors of $\mathbf{Q}$
- **their weighting:** (square roots of) eigenvalues of $\mathbf{Q}$

"Rich" feature space - large variety of motifs with significant weights.

The kernel and "metric tensor"

$$K(\mathbf{u},\mathbf{v}) = \mathbf{u}^\top \mathbf{Q}\ \mathbf{v} = \langle \mathbf{u},\mathbf{v}\rangle_{\mathbf{Q}},$$

where $\mathbf{Q}$ is a symmetric, positive semi-definite $\tau \times \tau$ matrix of rank $N_m = \mathrm{rank}(\mathbf{Q}) \leq N$ and elements

$$Q_{i,j} = \mathbf{w}^\top \left(\mathbf{W}^\top\right)^{i-1}\ \mathbf{W}^{j-1}\ \mathbf{w}, \quad i,j = 1,2,...,\tau.$$

With increasing time indices $i,j = 1,2,...,\tau$,

$$|Q_{i,j}| < \nu^{i+j-2}\ \|\mathbf{w}\|_2^2.$$

Note: $\mathbf{Q} = \mathbf{C}^\top\mathbf{C}$, where $\mathbf{C}$ is the controllability matrix $\mathbf{C} = [\mathbf{w}, \mathbf{W}\mathbf{w}, \mathbf{W}^2\mathbf{w}, ..., \mathbf{W}^{\tau-1}\mathbf{w}]$.

Various ESN reservoir settings

- random dynamic coupling $\mathbf{W}$, zero-mean i.i.d. entries
- random symmetric dynamic coupling $\mathbf{W}$
- deterministic minimalist setting - ring topology of $\mathbf{W}$
  SCR, 1 degree of freedom
  - a-periodic input coupling $\mathbf{w}$
  - periodic input coupling $\mathbf{w}$

Illustrative examples – setting

Influence of the dynamic and input coupling, **W** and **w**, respectively, on the strength and richness of motifs of the temporal kernel.

State space dimensionality $N = 100$.

Re-normalize $\mathbf{W} \in \mathbb{R}^{100\times100}$ to $\sigma_{max} = 0.995$.

Input coupling **w** is renormalized to unit length.

Show motifs with motif weights up to 10^{-2} of the highest motif weight.

Random $\mathbf{W}$. $W_{i,j}, w_i \sim N(0,1)$, i.i.d.

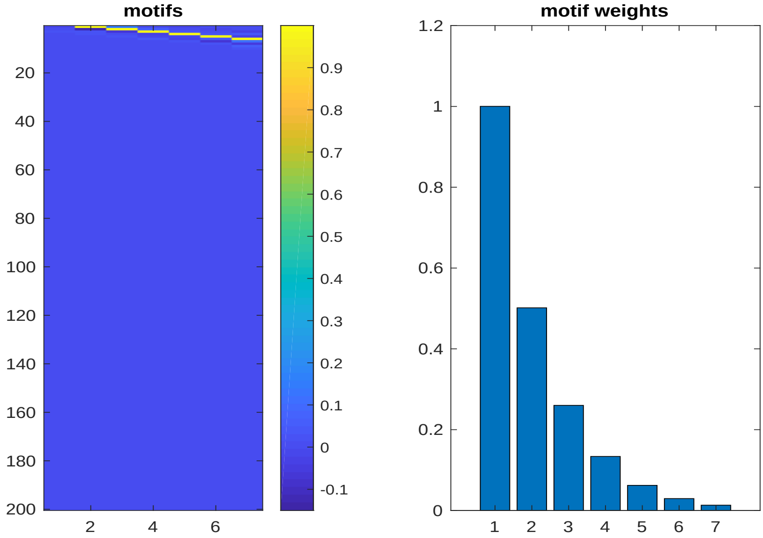
$\mathbf{u} = \ldots\ u_3\ u_2\ u_1, \quad \mathbf{v} = \ldots\ v_3\ v_2\ v_1$

The temporal kernel has a rigid Markovian flavor with shallow memory:

$$K(\mathbf{u},\mathbf{v}) \approx \|\mathbf{w}\|^2 \sum_i \left(\frac{\sigma_{max}}{2}\right)^{2(i-1)} u_i\ v_i$$

Compares the corresponding recent entries of the time series and weights down comparisons of past elements with rapidly decaying weights.

Random $\mathbf{W}$. $W_{i,j}, w_i \sim N(0,1)$, i.i.d.



Almost identical results for other distributions for $\mathbf{W}, \mathbf{w}$ and settings of $\mathbf{w}$.

Random symmetric Wigner $\mathbf{W}$, random $\mathbf{w}$

Theorem

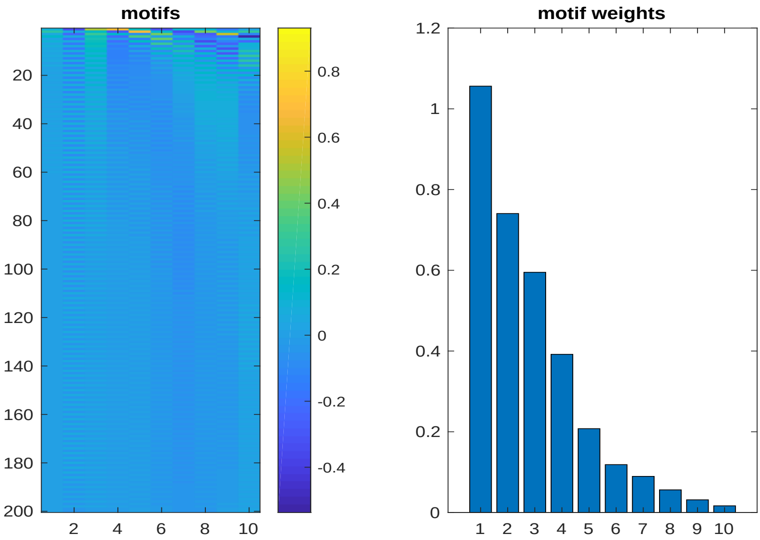
Symmetric coupling $\mathbf{W}$ *of rank* $N_k \leq N$. *Let* $\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_{N_k}$, *be the eigenvectors of* $\mathbf{W}$ *corresponding to non-zero eigenvalues* $\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_{N_k}$. *Denote by* $\widetilde{w}_a = \mathbf{s}_a^\top \mathbf{w}$ *the projection of the input coupling* $\mathbf{w}$ *onto the eigenvector* $\mathbf{s}_a$. *Then,*

$$K(\mathbf{u}, \mathbf{v}) = \sum_{a=1}^{N_k} \widetilde{w}_a^2 \ K^{(a)}(\mathbf{u}, \mathbf{v}),$$

each kernel $K^{(a)}$ *with a single motif*

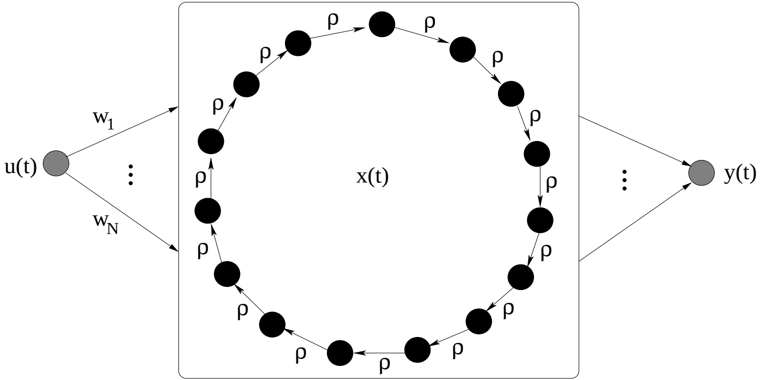
$$\mathbf{m}^{(a)} = (1, \sigma_a, \sigma_a^2, \sigma_a^3, ...)^\top$$

Random symmetric Wigner **W**, random **w**



The number and nature of the motifs stayed unchanged across a variety of generative mechanisms for **W** and **w**.

Ring topology of **W** - SCR



A. Rodan, P. Tino: Minimum Complexity Echo State Network. IEEE TNN, 22(1), pp 131-144, 2011.

Ring topology of $\mathbf{W}$

Theorem

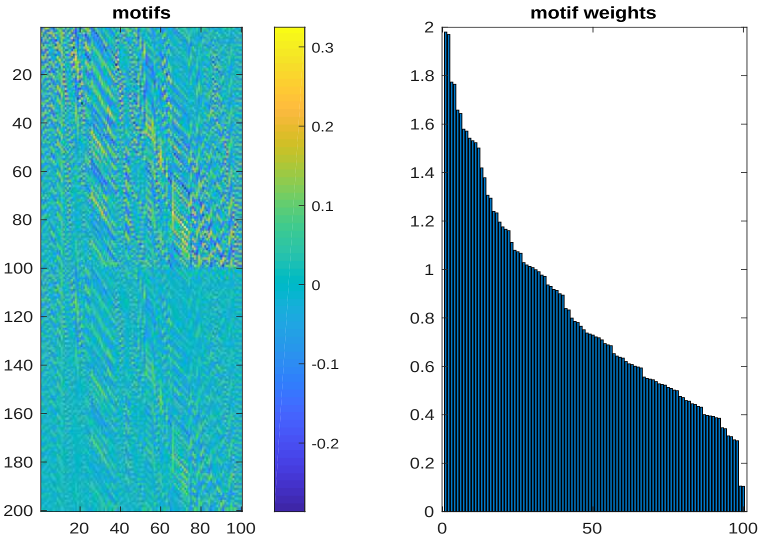
Let $\widetilde{\mathbf{m}}_i \in \mathbb{R}^N$, $i = 1, 2, ..., N$, be motifs of the temporal kernel under past time horizon N, with motif weights $\widetilde{\omega}_i$. Then, the motifs $\mathbf{m}_i$ have the following block form:

$$\mathbf{m}_i = \left(\widetilde{\mathbf{m}}_i^\top,\ \sigma_{max}^N \widetilde{\mathbf{m}}_i^\top,\ \sigma_{max}^{2N} \widetilde{\mathbf{m}}_i^\top, ...\right)^\top, \quad i = 1, 2, ...N.$$

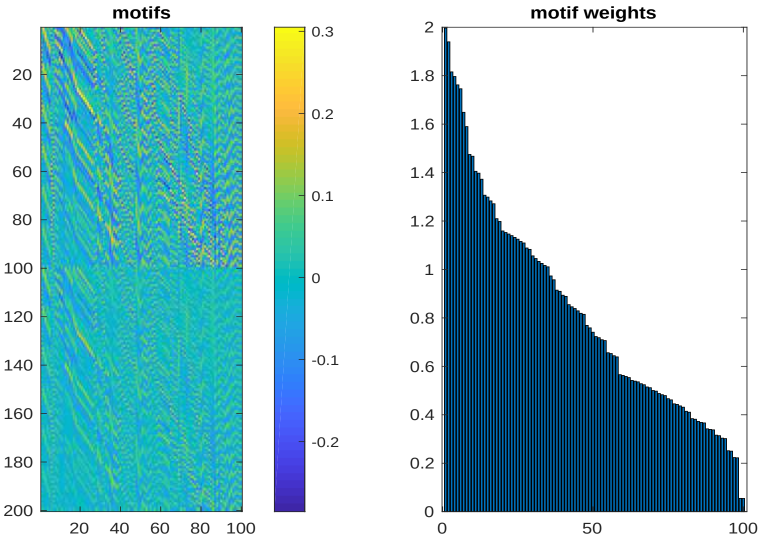
with the corresponding motif weights

$$\omega_i \approx \widetilde{\omega}_i \ \frac{1}{\sqrt{1 - \sigma_{max}^{2N}}}.$$

SCR, $\mathbf{w} \in \{-1, +1\}^N$, signs follow binary expansion of π

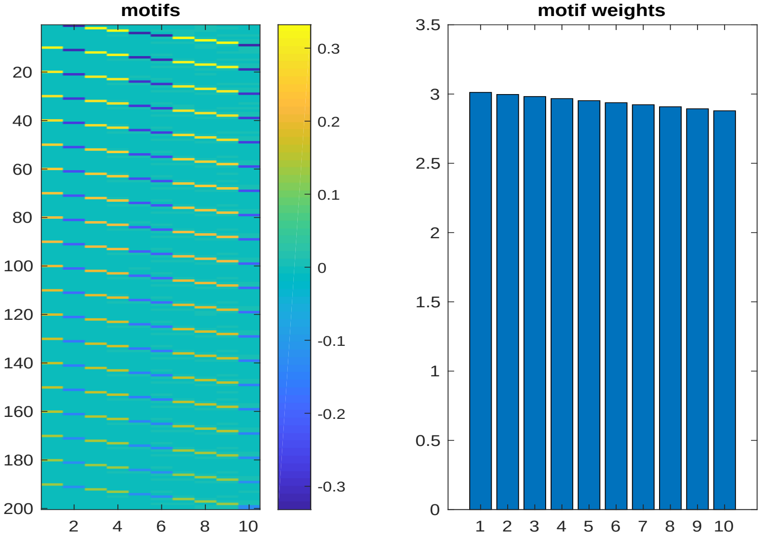


SCR, random **w**



motifs
20
40
60
80
100
120
140
160
180
200
20
40
60
80
100
0.3
0.2
0.1
0
-0.1
-0.2
motif weights
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
0
50
100

SCR, $\mathbf{w} \in \{-1,+1\}^N$, periodic, $p = 10$



Back to Markovianity - Fading Memory Filter

Input-output systems where the influence of inputs from deeper past on the present output is gradually fading out.

Given a time instance t, two input sequences $\mathbf{u}$ and $\mathbf{u}'$ having "similar recent histories" up to time t (but not necessarily the deeper past ones) will yield "similar outputs" at t.

Formalised for example through topological continuity arguments (Grigoryeva et al.):

Back to Markovianity - Fading Memory Filter

Norm $\|\cdot\|$ on the input space.

Left-infinite sequences $\{\mathbf{u}_t\}_{t\in\mathbb{Z}_-}$ (infinite product space) - endowed with a Banach space structure e.g. using the supremum norm

$$\|\mathbf{u}\|_\infty := \sup_{t\in\mathbb{Z}_-} \|\mathbf{u}_t\|.$$

Assign more weight on the recent items - weighted norm

$$\|\mathbf{u}\|_\mathbf{w} := \sup_{t\in\mathbb{Z}_-} \|w_t \cdot \mathbf{u}_t\|,$$

where $\mathbf{w} = \{w_t\}_{t\in\mathbb{Z}_-}$ is a weighting sequence, i.e. a strictly decreasing sequence (in the reverse time order) of positive real numbers with a fixed maximal element (e.g. $w_0 = 1$).

The space $\{\mathbf{u} \in (\mathbb{C}^m)^{\mathbb{Z}_-} \mid \|\mathbf{u}\|_\mathbf{w} < \infty\}$ forms a Banach space

Back to Markovianity - Fading Memory Filter

Analogously, norm on the output left-infinite sequences $\mathbf{y} = \{\mathbf{y}_t\}_{t \in \mathbb{Z}_-}$

a (possibly different) weighting sequence $\mathbf{v} = \{v_t\}_{t \in \mathbb{Z}_-}$ -
$\|\mathbf{y}\|_{\mathbf{v}} := \sup_{t \in \mathbb{Z}_-} \|v_t \cdot \mathbf{y}_t\|$ (the norm $\|\cdot\|$ is this time defined on the output space..

Require the maps realised by the causal input-output systems be continuous with respect to the topologies generated by the weighted norms $\|\cdot\|_{\mathbf{w}}$ and $\|\cdot\|_{\mathbf{v}}$.

Such input-output systems are said to have the *fading memory property*.

Formalizing the general purpose nature of reservoir architectures

J-P Ortega, L. Grigoryeva et al.:
Reservoir models (ESNs in particular) are a rich class of models - any time-invariant fading memory filter can be approximated to arbitrary precision by a linear reservoir system with polynomial readouts.

Given *any* time-invariant fading-memory filter F and $\epsilon > 0$, *there exists* a linear reservoir system R with polynomial readout h (and the corresponding linear reservoir function H_R) such that H_R is ϵ-close to F in the space of real-valued continuous functions over the space of *uniformly bounded inputs*.

Extensions and non-linear counterparts...

Universality of SCR

Show that given any ESN we can construct an (arbitrarily) close SRC in the sense of sup norm on input-output systems.

Then use universality results by J-P Ortega, L. Grigoryeva et al. to conclude that SCR are universal!

Universality of SCR

Key proof ideas:

$\rightarrow$ reservoir with **a given dynamic coupling**

$\rightarrow$ reservoir with **unitary coupling** (eigenvalues are on unit circle in $\mathbb{C}$)

$\rightarrow$ reservoir with **ring coupling** (eigenvalues are regularly placed on unit circle in $\mathbb{C}$ - roots of unity)

$\rightarrow$ reservoir with **ring coupling** and **constrained input coupling (same magnitude)**

B. Li, S.R. Fong, P. Tino: Simple Cycle Reservoirs are Universal. Journal of Machine Learning Research, 25(158), pp. 128, 2024.

Thank you! Interested?

- R.S. Fong, B. Li, P. Tino: Linear Simple Cycle Reservoirs at the edge of Stability perform Fourier decomposition of the input driving signals. **Chaos**, 4 (35), pp. 043109, 2025.
- R.S. Fong, B. Li, P. Tino: Universality of Real Minimal Complexity Reservoir. **AAAI** Conference on Artificial Intelligence, 39(16), 16622-16629, 2025.
- B. Li, S.R. Fong, P. Tino: Simple Cycle Reservoirs are Universal. **JMLR**, 25(158), pp. 128, 2024.
- P. Tino: Dynamical Systems as Temporal Feature Spaces. **JMLR**, 21(44), pp. 142, 2020.
- P. Verzelli, C. Alippi, L. Livi, P. Tino: Input-to-State Representation in linear reservoirs dynamics. **IEEE TNNLS**, 9 (33), pp. 4598-4609, 2022.
- A. Rodan, P. Tino: Minimum Complexity Echo State Network. **IEEE TNN**, 22(1), pp 131-144, 2011.
- P. Tino: Asymptotic Fisher Memory of Randomized Linear Symmetric Echo State Networks. **Neurocomp**, 298, pp. 48, 2018.
- H. Chen, F. Tang, P. Tino, A. G. Cohn, X. Yao: Model Metric Co-learning for Time Series Classification. **IJCAI** 2015, pp. 3387-3394, AAAI Press, 2015.
- H. Chen, F. Tang, P. Tino, X. Yao: Model-based Kernel for Efficient Time Series Analysis. **KDD** 2013, pp. 392-400, ACM New York, NY, USA. 2013.