Recurrent Neural Networks as Parametrized Non-autonomous Dynamical Systems
A Machine Learning Perspective

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
- Huanhuan Chen
- Jochen Steil
- Pietro Verzelli
- Lorenzo Livi
- Simon Fong
- Ali Rodan
- Alessandro Sperdutti
- Luba Benuskova
- Igor Farkas
- Michal Cernansky
- Whit Tabor
- ...

...
Learning from time series data

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

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

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

How to represent histories of inputs $\{u(t) | t \leq t_i\}$ given that they are of different length?
State space model
Input-driven parametrized state space models (e.g. RNNs).

Parametrized non-autonomous DS providing IPSs
\[ f : \mathcal{X} \times \mathcal{U} \times \Theta_f \to \mathcal{X} \]

\[ x(t) = f(x(t - 1), u(t); \theta_f) \]

Reading out the model response from IPSs
\[ h : \mathcal{X} \times \Theta_h \to \mathcal{Y} \]

\[ y(t) = h(x(t); \theta_h) \]
Discrete time state space model

States of the DS form IPS

These states code the entire history of sequence elements we have seen so far.
A possible solution to our example

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

Information latching problem  FSM extraction

Saddle node bifurcation
Attractive sets may be difficult to learn

Need to bifurcate to several attractive sets for $f_a$. 
Attractive sets
“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

\[\ldots 1 2 1 1 2 3 2 1 2 1 2 4 3 2 1 1 \]
\[\ldots 2 1 1 1 1 3 4 4 4 4 1 4 3 2 1 1 \]
\[\ldots 3 3 2 2 1 2 1 2 2 1 3 4 3 2 1 1 \]

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 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 2 1 2 1 3 4 3 2 1 1 → ?

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

\( s \in \{1, 2, 3, 4\} \)
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.
IFS are organized on nodes of PST
Traverse the tree from root towards leaves in reversed order
Complex and potentially time-consuming training

Each node $i$
has associated next-symbol probability distribution $P(.|i)$
Contractive maps lead to Markovian IPS - ESN et al.

[Diagram showing the process of contractive maps leading to Markovian IPS with symbols and contractions]
Contractive maps lead to Markovian IPS (1)

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

$$x(t) = f_s(x(t - 1)).$$

This notation can be extended to sequences over $\mathcal{A}$ the recurrent activations after $t$ time steps are

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

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

$$\|f_{vq}(x) - f_{wq}(x)\| \leq C^{|q|} \cdot \|f_v(x) - f_w(x)\| \leq C^{|q|} \cdot diam(\mathcal{X}).$$
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.

Add a simple output-generating readout on top of the fixed state dynamics!
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 fulfill the distribution independent UCED property and so

- unlike general recurrent networks, are distribution independent PAC-learnable.
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.
Feature space view - classification

\[ w^T u + b > 0 \]

\[ w^T u + b = 0 \]

\[ w^T u + b < 0 \]
Classification

\[ y = sgn(w^T u + b) \]

often

\[ w = \sum_i \alpha_i \cdot u_i \]

and so

\[ y = sgn \left( \sum_i \alpha_i \cdot u_i^T u + b \right) \]

\[ = sgn \left( \sum_i \alpha_i \cdot \langle u_i, u \rangle + b \right) \]
Embed the input space in a feature space:

\[ u \mapsto \phi(u) \in \mathcal{H} \]

\[ y = \text{sgn} \left( \sum_i \alpha_i \cdot \langle \phi(u_i), \phi(u) \rangle_{\mathcal{H}} + b \right) \]

\[ = \text{sgn} \left( \sum_i \alpha_i \cdot K(u_i, u) + b \right) \]

Kernel \( K(\cdot, \cdot) \) can define a "rich" Hilbert space \( \mathcal{H} \), but can itself be governed only by few free parameters!
Beyond discrete inputs: “temporal kernel”

View the state space model as a “dynamical kernel machine”:

\[
\begin{align*}
\phi(...u(t-3), u(t-2), u(t-1)) & = x(t-1) \\
\phi(...u(t-2), u(t-1), u(t)) & = x(t) \\
& = f(x(t-1), u(t))
\end{align*}
\]

\[
\begin{align*}
y(t) & = h(x(t)) \\
& = \sum_i \alpha_i \cdot \langle x(t_i), x(t) \rangle + b \\
& = \sum_i \alpha_i \cdot \langle \phi(..., u(t_i - 1), u(t_i)), \phi(..., u(t - 1), u(t)) \rangle + b
\end{align*}
\]
\[ x(t) = f(x(t - 1), u(t)) \]
\[ = \mathbf{W} \ x(t - 1) + u(t) \ w \]

\( \mathbf{W} \in \mathbb{R}^{N \times N} \) 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

\[ K(u, v) = \langle u, v \rangle_Q \]
\[ = u^\top Q v \]
\[ = u^\top M \Lambda M^\top v \]
\[ = \left\langle \Lambda^{\frac{1}{2}} M^\top u, \Lambda^{\frac{1}{2}} M^\top v \right\rangle \]

Eigen-analysis of the positive (semi-)definite “metric tensor” \( Q \)

- **basis - feature space “motifs”:** eigenvectors of \( Q \)
- **their weighting:** (square roots of) eigenvalues of \( Q \)
Random \( \mathbf{W} \): \( W_{i,j}, w_i \sim N(0, 1) \), i.i.d.

**Basis** - standard basis \( \mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \ldots \)
\( e_i \) picks the \( i \)-th last seen element

**Weighting** - fast exponential decay \( \left( \frac{\sigma_{\text{max}}}{2} \right)^{i-1} \)

\[ u = \ldots \ u_3 \ u_2 \ u_1, \quad v = \ldots \ v_3 \ v_2 \ v_1 \]

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

\[ K(u, v) \approx \| \mathbf{w} \|^2 \sum_i \left( \frac{\sigma_{\text{max}}}{2} \right)^{2(i-1)} u_i \ v_i \]
Random $\mathbf{W}$. $W_{i,j}$, $w_i \sim \mathcal{N}(0, 1)$, i.i.d.

Almost identical results for other distributions for $\mathbf{W}$, $\mathbf{w}$ and settings of $\mathbf{w}$. 
Ring topology of $W$ - SCR
Theorem

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

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

with the corresponding motif weights

$$\omega_i \approx \tilde{\omega}_i \frac{1}{\sqrt{1 - \sigma_{\text{max}}^{2N}}}.$$
SCR, $\mathbf{w} \in \{-1, +1\}^N$, signs follow binary expansion of $\pi$
Interested?


