The universality problem in dynamic machine learning

Lukas Gonon1, Lyudmila Grigoryeva2, and Juan-Pablo Ortega3,3
1Universität St. Gallen, Switzerland
2Universität Konstanz, Germany
3Centre National de Recherche Scientifique, France

Contributions
The universal approximation properties of three important families of reservoir computers (RC) are shown. We prove that both in deterministic and stochastic setups and for discrete-time semi-infinite inputs. We show that:
- Linear reservoir systems with either polynomial or neural network readout maps are universal:
- Two RC families with linear readouts, namely, state-affine systems (SAS) and echo state networks (ESN) (the most widely used RC systems in applications) are universal.
- The linearity in the readouts is a key feature in supervised machine learning. It guarantees that these systems can be used in high-dimensional/large-volume dataset situations. In the stochastic case proofs of two different types are constructed, in order to establish the universality of the RC systems with respect to $L^\infty$ and $L^2$-type criteria.

Mathematical model for reservoir computing
A reservoir computer (RC) is a particular case of recent network neural (RNN):
\[
\{ x_t = F(x_t-1, x_t), \} \\
\{ y_t = h(x_t), \} 
\]
where a reservoir map $F : \mathbb{R}^N \times \mathbb{R}^D \rightarrow \mathbb{R}^N$ and a readout map $h : \mathbb{R}^N \rightarrow \mathbb{R}^D$ form (or filter) an infinite discrete-time data $\mathbb{x} = \{ x_1, \ldots, x_T \} \in \mathbb{R}^{D \times T}$ into an output signal $\mathbb{y} \in \mathbb{R}^N$.

Additionally, $\mathbb{x} = 0$ in the input, $\mathbb{x} = \mathbb{r}$ in the reservoir state:
- The static readout $h : \mathbb{R}^N \rightarrow \mathbb{R}^D$ is trained in order to obtain the desired output $y$, output of the input $\mathbb{x}$.
- Different readouts can be trained on the same reservoir output for different tasks (multitasking).

Goal: identify families of reservoir filters that are able to uniformly approximate any time-invariant, causal, and fading memory filter with deterministic or stochastic inputs with any desired degree of accuracy. Such families of reservoir computers are said to be universal.

Reservoir systems
Linear reservoirs with a polynomial readout:
\[
\{ x_{t+1} = A x_t + c x_t, \} \quad A \in \mathbb{M}_N, \quad c \in \mathbb{M}_N, \quad x_0 \in \mathbb{R}^N. 
\]
Non-homogeneous state-affine systems (SAS):
\[
\{ p(z) = x_t = x_0 + z + x_t A + \cdots + x_t A^{n-1}, \} \\
\{ y_t = h(x_t), \} 
\]
where the associated to $p$ and $w$ is:
\[
\{ x_{t+1} = \sigma(A x_t + c x_t + z), \} \\
\{ y_t = w x_t. \}
\]

Problem statement (ESN):
\[
\{ x_{t+1} = \sigma(A x_t + C x_t + \zeta), \} \\
\{ y_t = w x_t. \}
\]

Deterministic setup \[3, 2]\n
1. The Stone-Weierstrass theorem for polynomial subalgebras of real-valued functions defined on compact metric spaces.
2. Internal approximation theorem: universality in the space of reservoir maps translates into universality into the space of reservoir filters.

Stochastic setup \[3, 1]\n
1. $L^p$ criterion using a transfer theorem: fading memory universal filters with deterministic uniformly bounded inputs have the same properties when presented with stochastic almost surely uniformly bounded inputs.
2. $L^p$ criterion: allows to cover a more general class of input signals. Allows us to formulate universality results for filters that do not necessarily have the fading memory property. Only measurability is required.

Universality: the deterministic setup

**Theorem (Reservoir family is universal)**

The set of all reservoir filters $R_N := \{ H : \mathbb{F} = \mathbb{K} \rightarrow \mathbb{R} \mid h \in \mathcal{C}_0(D), F : D : \mathbb{R}^D \times \mathbb{R}^M \rightarrow \mathbb{R}^D \}$ with inputs in the set $\mathbb{K}$ of uniformly bounded sequences by a constant $M$ and that have the fading memory property (FMP) w.r.t. a given width norm $\| \cdot \|_w$ is universal, that is, it is dense in the set $\mathcal{C}_0(D)$ of real-valued continuous functions on $\mathbb{K}$ equipped with $\| \cdot \|_w$. In other words, let $A(\mathbb{R})$ be the polynomial algebra generated by $\mathbb{R}$, then any causal, time-invariant FMP filter $H : \mathbb{K} \rightarrow \mathbb{R}$ can be uniformly approximated by elements in $A(\mathbb{R})$, that is, for any $\varepsilon > 0$

\[
\| H - H_{(\mathbb{R})} \|_w < \sup \| (H(z) - H_{(\mathbb{R})}(z)) \|_w < \varepsilon. 
\]

**Corollary (Universality of linear reservoirs)**
The set $\mathcal{L}_F$ formed by the all linear reservoirs as in (1)-(2) with matrices $A \in \mathbb{M}_N$, such that $s_{\text{max}}(A) < 1$, is made of $\lambda_0$-expansive fading memory reservoir functionals, with $\lambda_0 := (1 - c)^{-1}$, for any $c \in (0, 1)$. This family is dense in $\mathcal{C}_0(D)$.

The universal result can be stated for two smaller subfamilies of $\mathbb{L}$ generated by diagonal and nilpotent matrices.

**Theorem (Universality of SAS)**

Let $F^2 := \{ z \in \mathbb{R}^N \mid z \in [-1, 1], f \}$ for all $t$, and let $S_N$ be the family of functional $H_{f} : \mathbb{F} \rightarrow \mathbb{R}$ induced by the state-affine systems in (3)-(4) that satisfy $M_N := \max_{l \in \mathbb{N}} \| p^{(l)}(z) \|_w < 1 - c$ and $M_N := \max_{l \in \mathbb{N}} \| s^{(l)}(z) \|_w < 1 - c$. The subfamily $S_N$ is dense in $(C_0(D), \| \cdot \|_w)$.

Equivalently, for any fading memory filter $H$ and any $\varepsilon > 0$, there exist $N \in \mathbb{N}$, polynomials $p(z) \in \mathbb{M}_N[z]$, $c(z) \in \mathbb{M}_N[z]$, with $M_N < 1 - c$, and a vector $w \in \mathbb{R}^N$ s.t.

\[
\| H - H_{(\mathbb{R})} \|_w < \sup \| (H(z) - H_{(\mathbb{R})}(z)) \|_w < \varepsilon. 
\]

Perspectives
- What about unbounded inputs?
- What we know about continuity. What about differentiality?
- Performance bounds. Maurey-Brezis-Jones Theorems and the curse of dimensionality.
- Capacity estimates.
- We solved the approximation error problem. What about the estimation error problem?
- Relation to time series analysis.

References

Acknowledgements
The authors acknowledge partial financial support of the French ANR “RIBRPROM” project (ANR-14-IBSP10002-02) as well as the hospitality of the Centre Interfacultaire Bernoulli of the Ecole Polytechnique Fédérale de Lausanne during the program “StochasticDynamical Model in Mathematical Finance. Econometric and Numerical Statistical Science” that made possible the collaboration that led to some of the results included in this paper. Lyudmila Grigoryeva acknowledges partial financial support of the Graduate School of Decision Sciences and the Young Scholar Fund AFF of the Universität Konstanz. Juan-Pablo Ortega and Lukas Gonon acknowledge partial financial support coming from the Research Commission of the University Saint Gallen and the Swiss National Science Foundation (grant 200021_175891/1).