The universality problem in dynamic machine learning

Lukas Gonon[1], Lyudmila Grigoryeva[2], and Juan-Pablo Ortega[1,3]

[1]Universität St. Gallen, Switzerland [2]Universität Konstanz, Germany [3]Centre 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:

1. Linear reservoir systems with either polynomial or neural network readout maps are universal;
2. 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^∞ and L^p-type criteria.

Mathematical model for reservoir computing

A **reservoir computer (RC)** is a particular case of recurrent neural network (RNN):

$$\begin{cases} \mathbf{x}_t = F(\mathbf{x}_{t-1}, \mathbf{z}_t), \\ y_t = h(\mathbf{x}_t), \end{cases}$$

where a **reservoir** map $F : \mathbb{R}^N \times \mathbb{R}^n \longrightarrow \mathbb{R}^N$ and a **readout** map $h : \mathbb{R}^N \to \mathbb{R}^d$ transform (or filter) an infinite discrete-time input $\mathbf{z} = (\ldots, \mathbf{z}_{-1}, \mathbf{z}_0, \mathbf{z}_1, \ldots) \in (\mathbb{R}^n)^{\mathbb{Z}}$ into an output signal $\mathbf{y} \in (\mathbb{R}^d)^{\mathbb{Z}}$. Additionally,

- $\mathbf{z}_t \in \mathbb{R}^n$ is the input, $\mathbf{x}_t \in \mathbb{R}^N$ is the **reservoir state**.
- The static readout $h : \mathbb{R}^N \to \mathbb{R}^d$ is trained in order to obtain the desired output $\mathbf{y}_t$ out of the input $\mathbf{z}_t$.
- 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:

$$\begin{cases} \mathbf{x}_t = A\mathbf{x}_{t-1} + \mathbf{c}\mathbf{z}_t, & A \in \mathbb{M}_N, \mathbf{c} \in \mathbb{M}_{N,n}, \quad (1) \\ y_t = h(\mathbf{x}_t), & h \in \mathbb{R}[\mathbf{x}]. \quad (2) \end{cases}$$

Non-homogeneous state-affine systems (SAS): Let $p(z) \in \mathbb{M}_{N,N}[z]$ and $q(z) \in \mathbb{M}_{N,1}[z]$ be two polynomials on the variable z with matrix coefficients, that is

$$p(z) := A_0 + zA_1 + z^2A_2 + \cdots + z^{n_1}A_{n_1},$$
$$q(z) := B_0 + zB_1 + z^2B_2 + \cdots + z^{n_2}B_{n_2},$$

the SAS associated to p, q and $\mathbf{W}$ is:

$$\begin{cases} \mathbf{x}_t = p(z_t)\mathbf{x}_{t-1} + q(z_t), & (3) \\ y_t = \mathbf{W}^\top \mathbf{x}_t. & (4) \end{cases}$$

Echo state networks (ESN):

$$\begin{cases} \mathbf{x}_t = \sigma\left(A\mathbf{x}_{t-1} + C\mathbf{z}_t + \boldsymbol{\zeta}\right), & (5) \\ \mathbf{y}_t = W\mathbf{x}_t. & (6) \end{cases}$$

Setups and tools

Deterministic setup: [3, 2]

- The **Stone-Weierstrass theorem** for polynomial subalgebras of real-valued functions defined on compact metric spaces.
- **Internal approximation theorem**: universality in the space of reservoir maps translates into universality into the space of reservoir filters.

Stochastic setup: [3, 1]

- L^∞ 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.
- 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 $\mathcal{R}_w := \{H_h^F : K_M \longrightarrow \mathbb{R} \mid h \in C^\infty(D_N), F : D_N \times \overline{B_n(\mathbf{0}, M)} \longrightarrow D_N\}$ with inputs in the set K_M of uniformly bounded sequences by a constant M and that have the fading memory property (FMP) w.r.t. a given weighted norm $\|\cdot\|_w$ is universal, that is, it is dense in the set $(C^0(K_M), \|\cdot\|_w)$ of real-valued continuous functions on $(K_M, \|\cdot\|_w)$. In other words, let $\mathcal{A}(\mathcal{R}_w)$ be the polynomial algebra generated by $\mathcal{R}_w$, then any causal, time-invariant FMP filter $H : K_M \longrightarrow \mathbb{R}$ can be uniformly approximated by elements in $\mathcal{A}(\mathcal{R}_w)$, that is, for any $\epsilon > 0$

$$\|H - H_h^F\|_\infty := \sup_{\mathbf{z} \in K_M} |H(\mathbf{z}) - H_h^F(\mathbf{z})| < \epsilon.$$

Corolary (Universality of linear reservoirs)

The set $\mathcal{L}_\epsilon$ formed by all the linear reservoir systems as in (1)-(2) with matrices $A \in \mathbb{M}_N$ such that $\sigma_{\max}(A) < 1 - \epsilon$ is made of λ_ρ-exponential fading memory reservoir functionals, with $\lambda_\rho := (1-\epsilon)^\rho$, for any $\rho \in (0, 1)$. This family is dense in $(C^0(K_M), \|\cdot\|_{w^\rho})$. The same universality result can be stated for two smaller subfamilies of $\mathcal{L}_\epsilon$ generated by diagonal and nilpotent matrices.

Theorem (Universality of SAS)

Let $I^{\mathbb{Z}_-} := \{\mathbf{z} \in \mathbb{R}^{\mathbb{Z}_-} \mid z_t \in [-1, 1], \text{ for all } t \leq 0\}$, and let $\mathcal{S}_\epsilon$ be the family of functionals $H_{\mathbf{W}}^{p,q} : I^{\mathbb{Z}_-} \longrightarrow \mathbb{R}$ induced by the state-affine systems in (3)-(4) that satisfy that $M_p := \max_{z \in I} \|p(z)\|_2 < 1 - \epsilon$ and $M_q := \max_{z \in I} \|q(z)\|_2 < 1 - \epsilon$. The subfamily $\mathcal{S}_\epsilon$ is dense in $(C^0(I^{\mathbb{Z}_-}), \|\cdot\|_{w^\rho})$.
Equivalently, for any fading memory filter H and any $\epsilon > 0$, there $\exists$ $N \in \mathbb{N}$, polynomials $p(z) \in \mathbb{M}_N[z], q(z) \in \mathbb{M}_{N,1}[z]$ with $M_p, M_q < 1 - \epsilon$, and a vector $\mathbf{W} \in \mathbb{R}^N$ s.t.

$$\|H - H_{\mathbf{W}}^{p,q}\|_\infty := \sup_{z \in I^{\mathbb{Z}_-}} |H(z) - H_{\mathbf{W}}^{p,q}(z)| < \epsilon.$$

The same universality result can be stated for the smaller SAS subfamily determined by nilpotent polynomials.

Universality: the stochastic setup

Theorem (Deterministic-stochastic transfer principle)

Let $M > 0$ and let K_M and $K_M^{L^\infty}$ be the sets of deterministic and stochastic uniformly bounded inputs.

- *Let $H : (K_M, \|\cdot\|_w) \longrightarrow \mathbb{R}$ be a causal and time-invariant filter. Then H has the fading memory property if and only if the corresponding filter with almost surely uniformly bounded inputs has almost surely bounded outputs, that is, $H : (K_M^{L^\infty}, \|\cdot\|_{L_w^\infty}) \longrightarrow L^\infty(\Omega, \mathbb{R})$, and it has the fading memory property.*
- *Let $\mathcal{T} := \{H_i : (K_M, \|\cdot\|_w) \longrightarrow \mathbb{R} \mid i \in I\}$ be a family of causal and time-invariant fading memory filters. Then, $\mathcal{T}$ is dense in the set $(C^0(K_M), \|\cdot\|_w)$ if and only if the corresponding family with inputs in $K_M^{L^\infty}$ is universal in the set of continuous maps of the type $H : (K_M^{L^\infty}, \|\cdot\|_{L_w^\infty}) \longrightarrow L^\infty(\Omega, \mathbb{R})$.*

Theorem (Universality of SAS reservoir computers with stochastic inputs)

Let $K_I^{L^\infty} \subset L^\infty(\Omega, \mathbb{R}^{\mathbb{Z}_-})$ be the set of a.s. uniformly bounded processes in the interval $I = [-1, 1]$. Let $\mathcal{S}_\epsilon$ be the family of functionals $H_{\mathbf{W}}^{p,q} : K_I^{L^\infty} \longrightarrow L^\infty(\Omega, \mathbb{R})$ induced by the SAS that satisfy $M_p := \max_{z \in I} \|p(z)\| < 1-\epsilon$ and $M_q := \max_{z \in I} \|q(z)\| < 1 - \epsilon$. The family $\mathcal{S}_\epsilon$ forms a polynomial subalgebra of $\mathcal{R}_{w^\rho}$ with $w_t^\rho := (1-\epsilon)^{\rho t}$, made of FM reservoir filters that map into $L^\infty(\Omega, \mathbb{R})$.
For any time-invariant and causal FM filter $H : (K_I^{L^\infty}, \|\cdot\|_{L_{w^\rho}^\infty}) \longrightarrow L^\infty(\Omega, \mathbb{R})$ and any $\epsilon > 0$, there exists $N \in \mathbb{N}$, polynomials $p(z) \in \mathbb{M}_{N,N}[z], q(z) \in \mathbb{M}_{N,1}[z]$ with $M_p, M_q < 1 - \epsilon$, and a vector $\mathbf{W} \in \mathbb{R}^N$ such that

$$\|H - H_{\mathbf{W}}^{p,q}\|_\infty := \sup_{z \in K_I^{L^\infty}} \|H(z) - H_{\mathbf{W}}^{p,q}(z)\|_{L^\infty} < \epsilon.$$

The same universality result can be stated for SAS reservoir systems with nilpotent polynomials $p(z) \in \mathrm{Nil}[z]$.

Theorem (Internal approximation)

let $F_1, F_2 : \overline{B_{\|\cdot\|}(\mathbf{0}, L)} \times \overline{B_{\|\cdot\|}(\mathbf{0}, M)} \longrightarrow \overline{B_{\|\cdot\|}(\mathbf{0}, L)}$ be two continuous reservoir maps such that F_1 is a contraction with constant $0 < r < 1$ and F_2 has the existence of solutions property. Let $U_{F_1}, U_{F_2} : K_M \longrightarrow K_L$ be the corresponding filters (if F_2 does not have the ESP, then U_{F_2} is just a generalized filter). Then, for any $\epsilon > 0$, we have that

$$\|F_1 - F_2\|_\infty < \delta(\epsilon) := (1 - r)\epsilon$$

implies that

$$|||U_{F_1} - U_{F_2}|||_\infty < \epsilon.$$

Internal approximation in connection with the classical universality theorems for one-hidded-layer feedforward neural networks yields the universality of ESNs.

Echo state networks are universal

Let $U : I_n^{\mathbb{Z}_-} \longrightarrow \left(\mathbb{R}^d\right)^{\mathbb{Z}_-}$ be a causal and time-invariant filter that has the fading memory property. Then, for any $\epsilon > 0$ and any weighting sequence w, there is an echo state network

$$\begin{cases} \mathbf{x}_t = \sigma\left(A\mathbf{x}_{t-1} + C\mathbf{z}_t + \boldsymbol{\zeta}\right), & (7) \\ \mathbf{y}_t = W\mathbf{x}_t. & (8) \end{cases}$$

whose associated generalized filters U_{ESN} satisfy that

$$|||U - U_{\mathrm{ESN}}|||_\infty < \epsilon. \quad (9)$$

In these expressions $C \in \mathbb{M}_{N,n}$ for some $N \in \mathbb{N}$, $\boldsymbol{\zeta} \in \mathbb{R}^N$, $A \in \mathbb{M}_{N,N}$, and $W \in \mathbb{M}_{d,N}$. The function $\sigma : \mathbb{R}^N \longrightarrow [-1, 1]^N$ in (7) is constructed by componentwise application of a continuous squashing function $\sigma : \mathbb{R} \longrightarrow [-1, 1]$.
When the approximating echo state network (7)-(8) satisfies the echo state property, then it has a unique filter U_{ESN} associated which is necessarily time-invariant. The corresponding reservoir functional $H_{\mathrm{ESN}} : I_n^{\mathbb{Z}_-} \longrightarrow \mathbb{R}^d$ satisfies that

$$|||H_U - H_{\mathrm{ESN}}|||_\infty < \epsilon. \quad (10)$$

Perspectives

1. What about unbounded inputs?
2. We know a lot about continuity. What about differentiability?
3. Performance bounds. Maurey-Barron-Jones Theorems and the curse of dimensionality.
4. Capacity estimates.
5. We solved the approximation error problem. What about the estimation error problem?
6. Relation to time series analysis.

References

[1] Lukas Gonon and Juan-Pablo Ortega.
Reservoir computing universality with stochastic inputs.
Conditionally accepted in IEEE Transactions on Neural Networks and Learning Systems, 2018.

[2] Lyudmila Grigoryeva and Juan-Pablo Ortega.
Echo state networks are universal.
Neural Networks, 108:495–508, 2018.

[3] Lyudmila Grigoryeva and Juan-Pablo Ortega.
Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems.
Journal of Machine Learning Research, 19(24):1–40, 2018.

Acknowledgements

The authors acknowledge partial financial support of the French ANR "BIPHOPROC" project (ANR-14-OHRI-0002-02) as well as the hospitality of the Centre Interfacultaire Bernoulli of the Ecole Polytechnique Fédérale de Lausanne during the program "Stochastic Dynamical Models in Mathematical Finance, Econometrics, and Actuarial Sciences" 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 acknowledges partial financial support coming from the Research Commission of the Universität Sankt Gallen and the Swiss National Science Foundation (grant number 200021_175801/1).

University of St.Gallen Universität Konstanz