Random Neural Network Approximation of Dynamic Barron Functions

Lukas Gonon

based on joint work with Lyudmila Grigoryeva and Juan-Pablo Ortega

University of Munich, Germany

September 26th, 2022
3rd Symposium on MLDS
Fields Institute, Toronto

Overview

- Reservoir computing & echo state networks: highly efficient at learning dynamical / chaotic systems (see, e.g., Jaeger & Haas [JH04], Pathak, Hunt, Girvan, Lu & Ott [PHG+18], ...).
- Learning theoretical foundations?
  - Universality:
    - Grigoryeva & Ortega, *Neural Netw. (2018)* [GO18]
    - G. & Ortega, *IEEE TNNLS (2020)* [GO20]
    - G. & Ortega, *Neural Netw. (2021)* [GO21]
  - Generalization error: G., Grigoryeva & Ortega *JMLR (2020)* [GGO20]
  - Approximation error: G., Grigoryeva & Ortega *Ann. Appl. Probab. (2022+)* [GGO22]
  - RC systems via random projection: Cuchiero et al. *IEEE TNNLS (2021)* [CGG+21]
  - ...
- Quantitative bounds: for sufficiently smooth functionals.
- Goal: full learning error bounds for inherently infinite-dimensional, not necessarily smooth functionals.

Overview

- Reservoir computing & echo state networks: highly efficient at learning dynamical / chaotic systems (see, e.g., Jaeger & Haas [JH04], Pathak, Hunt, Girvan, Lu & Ott [PHG+18], ...).
- Learning theoretical foundations?
  - Universality:
    - Grigoryeva & Ortega, *Neural Netw. (2018)* [GO18]
    - G. & Ortega, *IEEE TNNLS (2020)* [GO20]
    - G. & Ortega, *Neural Netw. (2021)* [GO21]
  - Generalization error: G., Grigoryeva & Ortega *JMLR (2020)* [GGO20]
  - Approximation error: G., Grigoryeva & Ortega *Ann. Appl. Probab. (2022+)* [GGO22]
  - RC systems via random projection: Cuchiero et al. *IEEE TNNLS (2021)* [CGG+21]
  - ...
- Quantitative bounds: for sufficiently smooth functionals.
- Goal: full learning error bounds for inherently infinite-dimensional, not necessarily smooth functionals.

Overview

- Reservoir computing & echo state networks: highly efficient at learning dynamical / chaotic systems (see, e.g., Jaeger & Haas [JH04], Pathak, Hunt, Girvan, Lu & Ott [PHG+18], ...).
- Learning theoretical foundations?
    - Universality:
        - Grigoryeva & Ortega, *Neural Netw. (2018)* [GO18]
        - G. & Ortega, *IEEE TNNLS (2020)* [GO20]
        - G. & Ortega, *Neural Netw. (2021)* [GO21]
    - Generalization error: G., Grigoryeva & Ortega *JMLR (2020)* [GGO20]
    - Approximation error: G., Grigoryeva & Ortega *Ann. Appl. Probab. (2022+)* [GGO22]
    - RC systems via random projection: Cuchiero et al. *IEEE TNNLS (2021)* [CGG+21]
    - ...
- Quantitative bounds: for sufficiently smooth functionals.
- Goal: full learning error bounds for inherently infinite-dimensional, not necessarily smooth functionals.

Neural network approximation

Let us look at neural network approximation in the much further developed static case:
Neural networks are able to overcome the curse of dimensionality for classes of

- compositional functions (built from lower-dimensional functions),
- solutions to certain PDEs,
- Barron functions / functions with dimension-dependent regularity.
  - Originally proposed by Barron [Bar92], [Bar93].
  - Extended to the larger class of "generalized Barron functions" in E et al. [EW20], [EMWW20], [EMW19].
  - Goal: dynamic analogue
    - Rich class of functionals
    - Approximation and learning bounds.

Neural network approximation

Let us look at neural network approximation in the much further developed static case:
Neural networks are able to overcome the curse of dimensionality for classes of

- compositional functions (built from lower-dimensional functions),
- solutions to certain PDEs,
- Barron functions / functions with dimension-dependent regularity.
  - Originally proposed by Barron [Bar92], [Bar93].
  - Extended to the larger class of "generalized Barron functions" in E et al. [EW20], [EMWW20], [EMW19].
  - Goal: dynamic analogue
    - Rich class of functionals
    - Approximation and learning bounds.

Neural network approximation

Let us look at neural network approximation in the much further developed static case:
Neural networks are able to overcome the curse of dimensionality for classes of

- compositional functions (built from lower-dimensional functions),
- solutions to certain PDEs,
- Barron functions / functions with dimension-dependent regularity.
  - Originally proposed by Barron [Bar92], [Bar93].
  - Extended to the larger class of "generalized Barron functions" in E et al. [EW20], [EMWW20], [EMW19].
  - Goal: dynamic analogue
    - Rich class of functionals
    - Approximation and learning bounds.

Recurrent generalized Barron functionals

- Let $D_d \subset \mathbb{R}^d$ bounded, $\mathcal{I}_d \subset (D_d)^{\mathbb{Z}_-}$,
- $\sigma_1, \sigma_2 \colon \mathbb{R} \to \mathbb{R}$ activations (applied componentwise),
- $p \in [1, \infty]$, q such that $\frac{1}{p} + \frac{1}{q} = 1$.

Definition

$H \colon \mathcal{I}_d \to \mathbb{R}$ is called recurrent generalized Barron functional, if there exist

- a probability measure μ on $\mathbb{R} \times \ell^p \times \mathbb{R}^d \times \mathbb{R}$ with finite expectation,
- $B \in \ell^q$ and linear maps $A \colon \ell^q \to \ell^q$, $C \colon \mathbb{R}^d \to \ell^q$

such that for each $\mathbf{z} \in \mathcal{I}_d$ the system

$$\mathbf{x}_t = \sigma_1(A\mathbf{x}_{t-1} + C\mathbf{z}_t + B), \quad t \in \mathbb{Z}_-,$$

admits a unique solution $(\mathbf{x}_t)_{t \in \mathbb{Z}_-}$ with $\mathbf{x}_t = \mathbf{x}_t(\mathbf{z}) \in \ell^q$ and

$$H(\mathbf{z}) = \int_{\mathbb{R} \times \ell^p \times \mathbb{R}^d \times \mathbb{R}} w\sigma_2(\mathbf{a} \cdot \mathbf{x}_{-1}(\mathbf{z}) + \mathbf{c} \cdot \mathbf{z}_0 + b)\mu(dw, d\mathbf{a}, d\mathbf{c}, db), \quad \mathbf{z} \in \mathcal{I}_d.$$

Properties

Denote by $\mathcal{H}$ the class of all recurrent generalized Barron functionals. For natural choices of activation functions ($\sigma_1(x) = x$ and either $\sigma_2(x) = \max(x, 0)$ or σ_2 is bounded, continuous and non-constant) we obtain:

- $\mathcal{H}$ is a vector space
- $\mathcal{H}$ contains
  - sufficiently smooth functionals
  - functionals associated to convolutional filters
- $\mathcal{H} \cap L^{\bar{p}}(\mathcal{I}_d, \gamma)$ is dense in $L^{\bar{p}}(\mathcal{I}_d, \gamma)$ for any $\bar{p} \in [1, \infty)$ and any probability measure γ on $\mathcal{I}_d \subset (D_d)^{\mathbb{Z}_-}$.

Key example

Proposition

Suppose $p = 2$. Let $\mathcal{Y}$ be a separable Hilbert space, let $\bar{A}\colon \mathcal{Y} \to \mathcal{Y}$, $\bar{C}\colon \mathbb{R}^d \to \mathcal{Y}$ be linear and $\bar{B} \in \mathcal{Y}$ and assume that for each $\mathbf{z} \in \mathcal{I}_d$ the system

$$\bar{\mathbf{x}}_t = \bar{A}\bar{\mathbf{x}}_{t-1} + \bar{C}\mathbf{z}_t + \bar{B}, \quad t \in \mathbb{Z}_-, \tag{1}$$

admits a unique solution $(\bar{\mathbf{x}}_t)_{t\in\mathbb{Z}_-} \in \mathcal{Y}^{\mathbb{Z}_-}$. Let $\bar{\mu}$ be a (Borel) probability measure on $\mathbb{R} \times \mathcal{Y} \times \mathbb{R}^d \times \mathbb{R}$ with $\int_{\mathbb{R}\times\mathcal{Y}\times\mathbb{R}^d\times\mathbb{R}} |w|(\|\mathbf{a}\|_{\mathcal{Y}} + \|\mathbf{c}\| + |b|)\bar{\mu}(dw, d\mathbf{a}, d\mathbf{c}, db) < \infty$ and consider

$$H(\mathbf{z}) = \int_{\mathbb{R}\times\mathcal{Y}\times\mathbb{R}^d\times\mathbb{R}} w\sigma_2(\langle \mathbf{a}, \bar{\mathbf{x}}_{-1}(\mathbf{z})\rangle_{\mathcal{Y}} + \mathbf{c}\cdot\mathbf{z}_0 + b)\bar{\mu}(dw, d\mathbf{a}, d\mathbf{c}, db), \ \mathbf{z} \in \mathcal{I}_d. \tag{2}$$

Then $H \in \mathcal{H}$.

Such systems arise, e.g., in quantum reservoir computing.

Learning system

Goal: approximate (unknown) $H \in \mathcal{H}$ using random neural networks:

- Dynamics: captured by a (possibly linear) echo state network mapping an input $\mathbf{z}$ to

$$\mathbf{x}_t^{\mathrm{ESN}} = \sigma_1(\mathbf{A}^{\mathrm{ESN}}\mathbf{x}_{t-1}^{\mathrm{ESN}} + \mathbf{C}^{\mathrm{ESN}}\mathbf{z}_t + \mathbf{B}^{\mathrm{ESN}}), \quad t \in \mathbb{Z}_-, \tag{3}$$

  with given (randomly generated) matrices $\mathbf{B}^{\mathrm{ESN}} \in \mathbb{R}^N$, $\mathbf{A}^{\mathrm{ESN}} \in \mathbb{R}^{N\times N}$, $\mathbf{C}^{\mathrm{ESN}} \in \mathbb{R}^{N\times d}$.
- Random feedforward neural network readout: H is approximated by

$$\hat{H}(\mathbf{z}) = \hat{H}_{\mathbf{W}}(\mathbf{z}) = \sum_{i=1}^{N} W_i\sigma_2(\mathbf{a}^{(i)} \cdot \mathbf{x}_{-1}^{\mathrm{ESN}}(\mathbf{z}) + \mathbf{c}^{(i)} \cdot \mathbf{z}_0 + b_i) \tag{4}$$

  with randomly generated coefficients $\mathbf{a}^{(i)}$, $\mathbf{c}^{(i)}$, b_i valued in $\mathbb{R}^N$, $\mathbb{R}^d$ and $\mathbb{R}$, respectively.
- Only $\mathbf{W} \in \mathbb{R}^N$ is trainable.

Approximation result

- Let $T = \lceil \frac{N}{d} \rceil$ and suppose $\mathbf{A}^{\mathrm{ESN}}$, $\mathbf{C}^{\mathrm{ESN}}$ are generated so that $\|\mathbf{A}^{\mathrm{ESN}}\| < 1$ and K is invertible, with
  $$K = \pi_{N\times N}(\mathbf{C}^{\mathrm{ESN}}|\mathbf{A}^{\mathrm{ESN}}\mathbf{C}^{\mathrm{ESN}}|\cdots|(\mathbf{A}^{\mathrm{ESN}})^{T-2}\mathbf{C}^{\mathrm{ESN}}|(\mathbf{A}^{\mathrm{ESN}})^{T-1}\mathbf{C}^{\mathrm{ESN}}).$$
- Suppose $H \in \mathcal{H}$ with $\|A\| < 1$, μ has finite second moments.
- Assume that hidden readout weights are sampled from a generic measure ν satisfying an absolute continuity condition w.r.t. μ.

Theorem (Approximation error bound)

Consider the setting above and let $\sigma_1(x) = x$, $p \in (1, \infty)$, $\lambda \in (\|A\|, 1)$. Then there exists f such that $\hat{H}$ with readout $\mathbf{W} = f((w^{(i)}, \mathbf{a}^{(i)}, \mathbf{c}^{(i)}, b_i)_{i=1,\ldots,N})$ satisfies for any $\mathbf{z} \in \mathcal{I}_d$

$$\mathbb{E}[|H(\mathbf{z}) - \hat{H}(\mathbf{z})|^2]^{1/2} \leq C_{H,\mathrm{ESN}}[\lambda^{\frac{N}{d}} + \|\mathbf{A}^{\mathrm{ESN}}\|^T + \frac{1}{N^{\frac{1}{2}}}].$$

The constant $C_{H,\mathrm{ESN}}$ is available explicitly.

Approximation result

- Let $T = \lceil \frac{N}{d} \rceil$ and suppose $\mathbf{A}^{\mathrm{ESN}}$, $\mathbf{C}^{\mathrm{ESN}}$ are generated so that $\|\mathbf{A}^{\mathrm{ESN}}\| < 1$ and K is invertible, with
$$K = \pi_{N \times N}(\mathbf{C}^{\mathrm{ESN}} | \mathbf{A}^{\mathrm{ESN}}\mathbf{C}^{\mathrm{ESN}} | \cdots | (\mathbf{A}^{\mathrm{ESN}})^{T-2}\mathbf{C}^{\mathrm{ESN}} | (\mathbf{A}^{\mathrm{ESN}})^{T-1}\mathbf{C}^{\mathrm{ESN}}).$$
- Suppose $H \in \mathcal{H}$ with $\|A\| < 1$, μ has finite second moments.
- Assume that hidden readout weights are sampled from a generic measure ν satisfying an absolute continuity condition w.r.t. μ.

Theorem (Approximation error bound)

Consider the setting above and let $\sigma_1(x) = x$, $p \in (1, \infty)$, $\lambda \in (\|A\|, 1)$. *Then there exists* f *such that* $\hat{H}$ *with readout* $\mathbf{W} = f((w^{(i)}, \mathbf{a}^{(i)}, \mathbf{c}^{(i)}, b_i)_{i=1,\ldots,N})$ *satisfies for any* $\mathbf{z} \in \mathcal{I}_d$

$$\mathbb{E}[|H(\mathbf{z}) - \hat{H}(\mathbf{z})|^2]^{1/2} \leq C_{H,\mathrm{ESN}}[\lambda^{\frac{N}{d}} + \|\mathbf{A}^{\mathrm{ESN}}\|^T + \frac{1}{N^{\frac{1}{2}}}].$$

The constant $C_{H,\mathrm{ESN}}$ is available explicitly.

- Suppose the dynamic learning part (in particular $\mathbf{A}^{\mathrm{ESN}}$, $\mathbf{B}^{\mathrm{ESN}}$, $\mathbf{C}^{\mathrm{ESN}}$) remains "bounded" in N, i.e.,
  - there exists $\bar{c} > 0$ and $\underline{l}, \bar{l} \in (0,1)$, $\underline{l} < \bar{l}$ such that for any choice of N the ESN parameters satisfy that
    - K is invertible
    - $\underline{l} < \|\mathbf{A}^{\mathrm{ESN}}\| < \bar{l}$
    - $\|\mathbf{B}^{\mathrm{ESN}}\| \leq \bar{c}$, $\|\mathbf{C}^{\mathrm{ESN}}\| \leq \bar{c}$
    - $\|K^{-1}\mathrm{diag}(\mathbb{1}_d, \mathbb{1}_d \underline{l}^{k-1}, \ldots, \underline{l}^{T-1})\| \leq \bar{c}$

$\Rightarrow$ the constant $C_{H,\mathrm{ESN}}$ does not depend on N.

Universality

Let σ_1, σ_2 as before.

Corollary

Let $H\colon \mathcal{I}_d \to \mathbb{R}$ be an arbitrary functional and let ν_0 be a given hidden weight distribution with finite first moment.
Then for any $\varepsilon > 0$ and any probability measure γ on $\mathcal{I}_d \subset (D_d)^{\mathbb{Z}_-}$ with $H \in L^2(\mathcal{I}_d, \gamma)$ there exists a probability measure ν with $\mathcal{W}_1(\nu_0, \nu) < \varepsilon$ and a readout $\mathbf{W}$ such that $\hat{H}$ with readout $\mathbf{W}$ and distribution ν for the hidden layer weights satisfies

$$\left(\int_{\mathcal{I}_d} \mathbb{E}[|H(\mathbf{z}) - \hat{H}(\mathbf{z})|^2] \gamma(d\mathbf{z}) \right)^{1/2} < \varepsilon.$$

Special case: static situation

- Let μ_0 be a probability measure on $\mathbb{R} \times \mathbb{R}^d \times \mathbb{R}$ and

$$H(\mathbf{u}) = \int_{\mathbb{R}\times\mathbb{R}^d\times\mathbb{R}} w\sigma_2(\mathbf{c}\cdot\mathbf{u} + b)\mu_0(dw, d\mathbf{c}, db), \quad \mathbf{u} \in D_d \subset \mathbb{R}^d.$$

- $$\hat{H}(\mathbf{u}) = \sum_{i=1}^{N} W_i \sigma_2(\mathbf{c}^{(i)} \cdot \mathbf{u} + b_i)$$

  is used as learning system with randomly generated $\mathbf{c}^{(i)}$, b_i (distribution ν_0) and $\mathbf{W} \in \mathbb{R}^N$ trainable.

Corollary

Let H as above with $\mu_0 \ll \nu_0$. Then there exists $\mathbf{W}$ s. t. for any $\mathbf{u} \in D_d$

$$\mathbb{E}[|H(\mathbf{u}) - \hat{H}(\mathbf{u})|^2]^{\frac{1}{2}} \leq \frac{c}{N^{\frac{1}{2}}} \|\tfrac{d\mu_0}{d\nu_0}\|_\infty^{\frac{1}{2}} \left(\int w^2[\|\mathbf{c}\|^2 + |b|^2 + 1]\mu_0(dw, d\mathbf{c}, db)\right)^{\frac{1}{2}},$$

where $c = (2\max(2L_{\sigma_2}, |\sigma_2(0)|^2)\max(1, \sup_{v\in D_d}\|v\|^2))^{\frac{1}{2}}$.

Learning error bounds

How about learning H from a single trajectory of input/output pairs?

- Observations $(\mathbf{Z}_t, \mathbf{Y}_t)_{t=0,-1,\ldots,-n+1}$ are available.
- Let $H \in \mathcal{H}$ be the unknown functional and assume that the input/output relation between the data is given as $H(\mathbf{Z}) = \mathbb{E}[\mathbf{Y}_0|\mathbf{Z}]$
  - Example: $\mathbf{Y}_t = H(\mathbf{Z}_{t+\cdot}) + \varepsilon_t$ for a stationary process $(\varepsilon_t)_{t\in\mathbb{Z}_-}$ independent of $\mathbf{Z}$ and with $\mathbb{E}[\varepsilon_0] = 0$.
- To learn H from the data we solve

$$\hat{\mathbf{W}} = \arg\min_{\mathbf{W}} \frac{1}{n} \sum_{i=0}^{n-1} \|\hat{H}_{\mathbf{W}}(\mathbf{Z}_{-i}^{-n+1}) - \mathbf{Y}_{-i}\|^2 \tag{5}$$

  where we denote $\mathbf{Z}_{-i}^{-n+1} = (\ldots, 0, 0, \mathbf{Z}_{-n+1}, \ldots, \mathbf{Z}_{-i-1}, \mathbf{Z}_{-i})$.
- Data points are not i.i.d., but only weakly dependent.

Theorem

Consider the setting as above, let $R \geq \frac{c}{\sqrt{N}}$*, assume* $\mathbf{Y}$ *is bounded and* $(\mathbf{Z}, \mathbf{Y})$ *has a causal Bernoulli shift structure with geometric decay of rate* λ_{dep} *and* $\log(n) < n \log(\lambda_{max}^{-1})$*, where* $\lambda_{max} = \max(\|\mathbf{A}^{\mathrm{ESN}}\|, \lambda_{\mathrm{dep}})$*. Then the trained system* $\hat{H}_{\hat{\mathbf{W}}}$ *satisfies the learning error bound*

$$\mathbb{E}[|H(\bar{\mathbf{Z}}) - \hat{H}_{\hat{\mathbf{W}}}(\bar{\mathbf{Z}})|^2]^{1/2} \leq C_{\mathrm{approx}} \left(\lambda^{\frac{N}{d}} + \|\mathbf{A}^{\mathrm{ESN}}\|^T + \frac{1}{N^{\frac{1}{2}}} \right) + C_{\mathrm{est}} \left(R N^{\frac{1}{2}} \frac{\sqrt{\log(n)}}{\sqrt{n}} \right)^{\frac{1}{2}} \tag{6}$$

where $\bar{\mathbf{Z}}$ *is an independent copy of* $\mathbf{Z}$ *and* C_{approx}*,* C_{est} *are explicitly given and not depending on* N*,* n*.*

Thank you!

References I

Andrew R Barron.
Neural net approximation.
In *Yale Workshop on Adaptive and Learning Systems*, volume 1, pages 69–72, 1992.

Andrew R. Barron.
Universal approximation bounds for superpositions of a sigmoidal function.
IEEE Trans. Inform. Theory, 39(3):930–945, 1993.

Christa Cuchiero, Lukas Gonon, Lyudmila Grigoryeva, Juan-Pablo Ortega, and Josef Teichmann.
Discrete-time signatures and randomness in reservoir computing.
IEEE Transactions on Neural Networks and Learning Systems, 10.1109/TNNLS.2021.3076777:1–10, 2021.

Weinan E, Chao Ma, and Lei Wu.
A priori estimates of the population risk for two-layer neural networks.
Commun. Math. Sci., 17(5):1407–1425, 2019.

Weinan E, Chao Ma, Stephan Wojtowytsch, and Lei Wu.
Towards a mathematical understanding of neural network-based machine learning: what we know and what we don't.
Preprint, arXiv 2009.10713, 2020.

Weinan E and Stephan Wojtowytsch.
On the banach spaces associated with multi-layer relu networks: Function representation, approximation theory and gradient descent dynamics.
Preprint, arXiv 2007.15623, 2020.

Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega.
Risk bounds for reservoir computing.
J. Mach. Learn. Res., 21:Paper No. 240, 61, 2020.

References II

Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega.
Approximation bounds for random neural networks and reservoir systems.
Preprint, arXiv 2002.05933; to appear in The Annals of Applied Probability, 2022+.

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

Lukas Gonon and Juan-Pablo Ortega.
Reservoir computing universality with stochastic inputs.
IEEE Trans. Neural Netw. Learn. Syst., 31(1):100–112, 2020.

Lukas Gonon and Juan-Pablo Ortega.
Fading memory echo state networks are universal.
Neural Networks, 138:10–13, 2021.

Herbert Jaeger and Harald Haas.
Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication.
Science, 304(5667):78–80, 2004.

Jaideep Pathak, Brian Hunt, Michelle Girvan, Zhixin Lu, and Edward Ott.
Model-Free Prediction of Large Spatiotemporally Chaotic Systems from Data: A Reservoir Computing Approach.
Physical Review Letters, 120(2):24102, 2018.

Appendix Setup I: Sufficiently regular functionals

Consider an unknown functional $H^*: (\mathbb{R}^d)^{\mathbb{Z}_-} \to \mathbb{R}^m$ and a random input signal $\mathbf{Z}$ (valued in $B_M(0)^{\mathbb{Z}_-}$). The goal is to approximate $H^*(\mathbf{Z})$.

- Key examples:
  - $H^*(\mathbf{Z}) = \mathbf{X}_0^*$, where (for a suitable $F^*: \mathbb{R}^{N^*} \times \mathbb{R}^d \to \mathbb{R}^{N^*}$)

$$\mathbf{X}_t^* = F^*(\mathbf{X}_{t-1}^*, \mathbf{Z}_t), \quad t \in \mathbb{Z}_-.$$

  - $H^*(\mathbf{Z}) = \mathbb{E}[\mathbf{Z}_1 | \mathbf{Z}_0, \mathbf{Z}_{-1}, \ldots]$.
- Assumptions:
  - H^* is d_w-Lipschitz-continuous for some summable weighting sequence w, that is, there exists $L > 0$ such that for all $\mathbf{v}, \mathbf{u} \in B_M(0)^{\mathbb{Z}_-}$

$$\|H^*(\mathbf{v}) - H^*(\mathbf{u})\| \leq L \left[\sum_{t \in \mathbb{Z}_-} w_t \|\mathbf{v}_t - \mathbf{u}_t\|^2 \right]^{1/2}.$$

  - For all $T \in \mathbb{N}$ the truncated function(al) $H_T^*: (\mathbb{R}^d)^{T+1} \to \mathbb{R}^m$, $H_T^*(\mathbf{z}_0, \ldots, \mathbf{z}_{-T}) = H^*(\mathbf{z}_0, \ldots, \mathbf{z}_{-T}, 0, \ldots)$ is sufficiently smooth and integrable (e.g. Sobolev-regularity $W^{2,2}(\mathbb{R}^{d(T+1)})$).

Appendix Setup II: Linear reservoir, random neural network readout

Consider learning based on a recurrent neural network with ReLU activation function and randomly generated $\mathbf{A}, \mathbf{S}, \mathbf{c}, \zeta$ (independent of $\mathbf{Z}$).

- The input signal $\mathbf{Z} \in (\mathbb{R}^d)^{\mathbb{Z}_-}$ is mapped to the output signal $\mathbf{Y} \in (\mathbb{R}^m)^{\mathbb{Z}_-}$ via

$$\begin{aligned} \mathbf{X}_t &= \sigma_1(\mathbf{S}\mathbf{X}_{t-1} + \mathbf{c}\mathbf{Z}_t), \\ \mathbf{Y}_t &= \mathbf{W}\sigma_2(\mathbf{A}\mathbf{X}_t + \zeta), \quad t \in \mathbb{Z}_-. \end{aligned} \tag{7}$$

- $\sigma_1(x) = x$, $\sigma_2(x) = \max(x, 0)$.
- Each of the N rows of $\mathbf{A}$ is generated (i.i.d.) randomly from uniform distribution in $B_R(0) \subset \mathbb{R}^{d(T+1)}$.
- The entries of ζ are generated i.i.d. uniformly on $[-MR, MR]$.
- $\mathbf{S}, \mathbf{c}$ are (random) matrices with $\lim_{k\to\infty} \|\mathbf{S}^k\|_2 = 0$ and such that $\mathbf{K} = [\mathbf{c}\ \mathbf{S}\mathbf{c}\ \cdots\ \mathbf{S}^T\mathbf{c}]$ has full rank.
- R, T, N can be chosen (to make the bound as small as possible).
- $\mathbf{W} \in \mathbb{R}^{m\times N}$ is trained (linear regression!).

Appendix: Approximation error bound

Theorem (G., Grigoryeva & Ortega [GGO22])

For any sufficiently regular functional there exists a readout $\mathbf{W}$ *(a* $\mathbb{R}^{1\times N}$*-valued random variable) such that for any* $\delta \in (0,1)$*, with probability* $1-\delta$ *the approximation error satisfies*

$$\mathbb{E}[|\mathbf{Y}_0 - H^*(\mathbf{Z})|^2 | \mathbf{A}, \mathbf{S}, \mathbf{c}, \zeta]^{\frac{1}{2}} \leq \frac{1}{\delta}\left[\frac{I_1(T,R)^{\frac{1}{2}}}{\sqrt{N}} + I_2(T,R) + I_3(T)\right], \text{ with}$$

$$I_1(T,R) = C_1(\mathbf{S},\mathbf{c}) R \mathrm{Vol}_{d(T+1)}(B_R(0)) \int_{B_R} \max(1, \|u\|^3) |\widehat{H_T^*}(\mathbf{K}^* u)|^2 \mathrm{d}u,$$

$$I_2(T,R) = |\det(\mathbf{K})| \int_{B_R(0)^c} |\widehat{H_T^*}(\mathbf{K}^* u)| \mathrm{d}u,$$

$$I_3(T) = LM \left(\sum_{t=T+1}^{\infty} w_{-t}\right)^{1/2} + C_2(\mathbf{S},\mathbf{c}) \|\mathbf{S}^T\|.$$

Appendix: Weak dependence

Definition

An $\mathbb{R}^k$-valued random process $\mathbf{U}$ is said to have a causal Bernoulli shift structure, if there exist $q \in \mathbb{N}$, $G \colon (\mathbb{R}^q)^{\mathbb{Z}_-} \to \mathbb{R}^k$ measurable and an i.i.d. collection $(\boldsymbol{\xi}_t)_{t \in \mathbb{Z}_-}$ of $\mathbb{R}^q$-valued random variables such that

$$\mathbf{U}_t = G(\ldots, \boldsymbol{\xi}_{t-1}, \boldsymbol{\xi}_t), \quad t \in \mathbb{Z}_-.$$

It is said to have geometric decay, if there exist $C_{\mathrm{dep}} > 0$, $\lambda_{\mathrm{dep}} \in (0,1)$ such that the weak dependence coefficient $\theta(\tau) := \mathbb{E}[\|\mathbf{U}_0 - \tilde{\mathbf{U}}_0^{\tau}\|]$ satisfies $\theta(\tau) \leq C_{\mathrm{dep}} \lambda_{\mathrm{dep}}^{\tau}$ for all $\tau \in \mathbb{N}$, where
$\tilde{\mathbf{U}}_0^{\tau} = G(\ldots, \tilde{\boldsymbol{\xi}}_{-\tau-1}, \tilde{\boldsymbol{\xi}}_{-\tau}, \boldsymbol{\xi}_{-\tau+1}, \ldots, \boldsymbol{\xi}_0)$ for $\tilde{\boldsymbol{\xi}}$ an independent copy of $\boldsymbol{\xi}$.