Memory of recurrent networks: Do we compute it right?

Workshop – Mathematics of Data Stream
Greifswald

Giovanni Ballarin[1], Lyudmila Grigoryeva[2,3], Juan-Pablo Ortega[4,5]

[1]University of Mannheim; [2]University of St.Gallen; [3]University of Warwick;
[5]Nanyang Technical University: [5]CNRS;

April 12, 2024

Introduction

- Linear SSM:

$$\boldsymbol{x}_t = A\boldsymbol{x}_{t-1} + \boldsymbol{C}z_t + \zeta, \tag{1}$$
$$y_t = W^\top \boldsymbol{x}_t, \tag{2}$$

  where z_t and y_t are scalars, $\boldsymbol{x}_t \in \mathbb{R}^N$.

- Assume A, $\boldsymbol{C}$ and ζ are **randomly drawn** from a regular distribution.
  - Random-init. (linear) RNNs
  - Random-weights NNs
  - Reservoir Models / (linear) **Echo State Networks** (LESNs)

- **Memory capacity:**

$$\mathrm{MC}_\tau := 1 - \frac{1}{\mathrm{Var}(z_t)} \min_{W_\tau \in \mathbb{R}^N} \mathbb{E}\left[\| z_{t-\tau} - W_\tau^\top \boldsymbol{x}_t \|^2 \right], \tag{3}$$

- Mapping $\tau \mapsto \mathrm{MC}_\tau$ is called a **memory curve**.
- The sum $\sum_{\tau=0}^{\infty} \mathrm{MC}_\tau$ is the **total memory capacity** of model (1)-(2).

Some Key References

Memory Capacity:
Jaeger (2002), Matthews (1992), Matthews and Moschytz (1994), Jaeger and Haas (2004), Dambre et al. (2012)

Fisher Memory:
Ganguli et al. (2008), Tino and Rodan (2013), Livi et al. (2016), Tino (2018)

Memory Properties:

Empirical:
Whiteaker and Gerstoft (2022a), Whiteaker and Gerstoft (2022b), Verzelli et al. (2021)

Formal:
White et al. (2004), Hermans and Schrauwen (2010), Grigoryeva et al. (2015, 2016a), Marzen (2017), Gonon et al. (2020), Grigoryeva et al. (2021)

Memory Maximization:

Architecture:
Farkas et al. (2016), Strauss et al. (2012), Rodan and Tino (2011, 2012), Tino and Rodan (2013)

Hyperparameters:
Gallicchio (2020), Aceituno et al. (2020)

This Paper

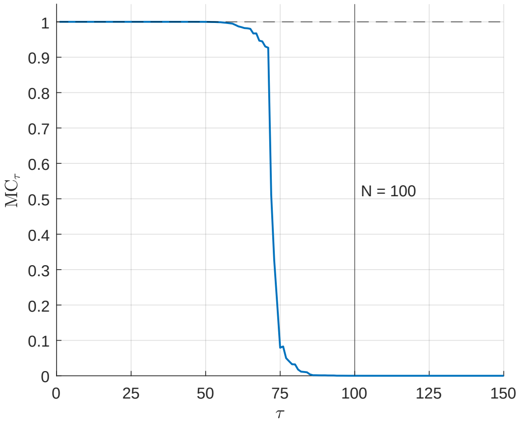
Address two key issues with *theoretical* vs. *practical* memory in LESNs (RNNs):

1. Show how and *why* simulation and "naive" memory capacity estimation methods yield MCs which do not agree with known theoretical results.
2. Develop *robust* memory capacity estimation methods.

Outline

Table of Contents

Memory Curve: Example



Two Fundamental Memory Results

1. (Jaeger, 2002) It holds
$$1 \leq \mathrm{MC} \leq N.$$
2. (Gonon et al., 2020) Linear ESNs have (almost always) *maximal* memory:

Proposition: Perfect Memory

Consider a linear ESN model in (1)-(2) and let $\boldsymbol{\zeta} = \mathbf{0}$. Let A be diagonalizable and such that $\rho(A) < 1$, with $\rho(A)$ the spectral radius of the matrix A. Suppose that all the eigenvalues of A are distinct. Let any of the following equivalent conditions hold

(i) The vectors $\{A\boldsymbol{C}, A^2\boldsymbol{C}, \ldots, A^N\boldsymbol{C}\}$ form a basis of $\mathbb{R}^N$.

(ii) The Kalman controllability condition holds.

(iii) A has full rank and $\boldsymbol{C}$ is neither the zero vector nor an eigenvector of A.

If $(z_t)_{t\in\mathbb{Z}_-}$ is a weakly stationary white noise process, then $\mathrm{MC} = N$.

Naive MC Estimation

Memory estimation of a LESN when z_t are IID:

- One can show that

$$\mathrm{MC}_\tau = \frac{\mathrm{Cov}(z_{t-\tau}, \boldsymbol{x}_t)\Gamma_{\boldsymbol{x}}^{-1}\mathrm{Cov}(\boldsymbol{x}_t, z_{t-\tau})}{\mathrm{Var}(z_t)} \qquad \text{with} \qquad \Gamma_{\boldsymbol{x}} := \mathrm{Var}(\boldsymbol{x}_t). \tag{4}$$

- Run a Monte Carlo experiment:
  1. Randomly generate A, $\boldsymbol{C}$ and ζ
  2. To make the system stable, rescale A such that $\rho(A) < 1$
  3. Fix warmup length T_0 and sample length T
  4. Draw sample $\{z_{-T_0+1}, , z_0, z_1, \ldots, z_T\}$
  5. Set $\boldsymbol{x}_{-T_0} = 0$, collect states $\{\boldsymbol{x}_{T_0-1}, \ldots, \boldsymbol{x}_T\}$ and discard first T_0 states
  6. Compute MC_τ using sample plug-in estimators in (5)

Memory: Uniform A, Gaussian C

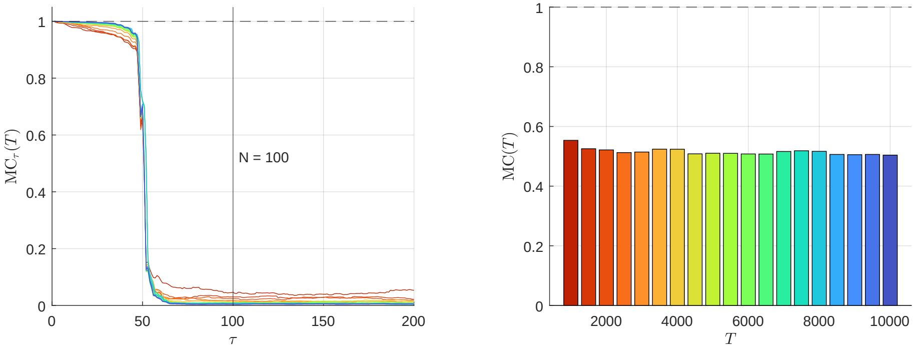


Figure: Memory curves $\widehat{\mathrm{MC}}_\tau(T)$ (left) and bar chart of normalized total memory capacity $\widehat{\mathrm{MC}}(T)/N$ (right) for the same ESN model estimated using progressively larger simulation samples. All memory curves $\widehat{\mathrm{MC}}_\tau(T)$ are computed for $\tau \in \{0, 1, \ldots, \mathbf{2N}\}$ but plot is shortened for clarity. Estimators are computed from simulated $\{z_t\}$ of sample size T, where $z_t \sim$ i.i.d. $\mathcal{N}(0,1)$ and T ranges from 1000 to 10 000 in increments of 500. Input mask $\boldsymbol{C} = [c_i] \in \mathbb{R}^N$ where $c_{i,j} = \overline{\boldsymbol{C}}/\|\overline{\boldsymbol{C}}\|$ for $\overline{\boldsymbol{C}} = [\overline{c}_i] \sim$ i.i.d. $\mathcal{N}(0,1)$.

Uniform distribution connectivity matrix $A = [A_{i,j}] \in \mathbb{R}^{N\times N}$ of size $N = 100$ and spectral radius $\rho(A) = 0.9$.

Memory: Orthogonal A, Gaussian C

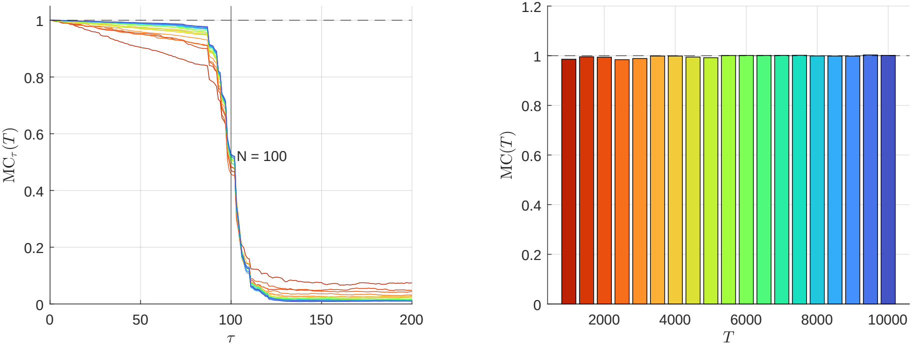


Figure: All memory curves $\widehat{\mathrm{MC}}_\tau(T)$ are computed for $\tau \in \{0, 1, \ldots, \mathbf{2N}\}$ but plot is shortened for clarity.

Orthogonal connectivity matrix $A = [A_{i,j}] \in \mathbb{R}^{N \times N}$ of size $N = 100$ and spectral radius $\rho(A) = 0.9$.

The Trouble with Linear Memory

Q: What if we instead compute MC algebraically?

- It is straightforward (when z_t are iid) to simplify MC_τ to

$$\mathrm{MC}_\tau = \boldsymbol{C}^\top (A^\top)^\tau \underbrace{\left[\sum_{j=0}^{\infty} A^j \boldsymbol{C}\boldsymbol{C}^\top (A^\top)^j \right]}_{G_{\boldsymbol{x}}}^{-1} A^\tau \boldsymbol{C}, \tag{5}$$

- Since we must invert $G_{\boldsymbol{x}}$, we should ask what its properties are:
  - ↳ The infinite sum is not a problem, as Gonon et al. (2020) also provide a closed-form solution:

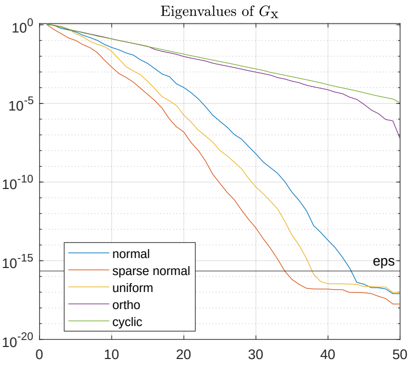
$$G_{\boldsymbol{x}} = \sum_{i,j=1}^{N} \frac{c_i c_j^{\mathrm{H}}}{1 - \lambda_i \lambda_j^{\mathrm{H}}} \boldsymbol{v}_i \boldsymbol{v}_j^{\mathrm{H}}.$$

    where $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\}$ is an eigenbasis of A and $\boldsymbol{C} = \sum_{i=1}^{N} c_i \boldsymbol{v}_i$.

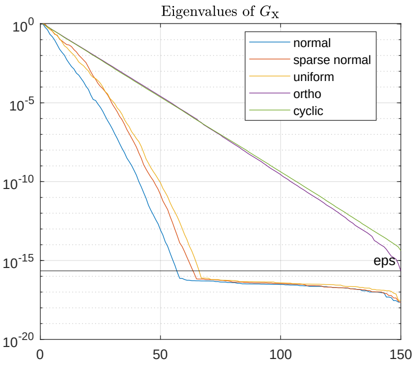
- But the **eigenvalue decay** of $G_{\boldsymbol{x}}$ can be very fast!

▸ Example

The Trouble with Linear Memory



(a) $N = 50$



(b) $N = 150$

Figure: Eigenvalue plot (absolute values) for G_x computed using $\max(1000, 5N)$ series terms, $A = [A_{i,j}] \in \mathbb{R}^{N \times N}$ with $\rho(A) = 0.9$ for all designs. Input mask $\boldsymbol{C} = [c_i] \in \mathbb{R}^N$ where $c_{i,j} = \overline{\boldsymbol{C}}/\|\overline{\boldsymbol{C}}\|$ for $\overline{\boldsymbol{C}} = [\overline{c}_i] \sim$ i.i.d. $\mathcal{N}(0, 1)$. Run in MATLAB with $eps \approx 2.2204 \times 10^{-16}$.

Krylov Conditioning

- For $N \in \mathbb{N}$, $A \in \mathbb{M}_N$, and $\boldsymbol{C} \in \mathbb{R}^N$ define the Krylov matrix

$$K := \left(\boldsymbol{C} \,|\, A\boldsymbol{C} \,|\, A^2\boldsymbol{C} \,|\, \ldots \right).$$

- Under hypothesis $\rho(A) < 1$, Gelfand's formula (see Lax (2002)) implies that for any $\epsilon > 0$ there exists $k \in \mathbb{N}$ such that $\|A^k\|_\infty < \epsilon$.
- We can use the truncation

$$K_m := \left(\boldsymbol{C} \,|\, A\boldsymbol{C} \,|\, A^2\boldsymbol{C} \,|\, \ldots \,|\, A^{m-1}\boldsymbol{C} \right)$$

to derive approximation

$$G_{\boldsymbol{x}} \approx K_m K_m^\top.$$

- K_m is a **Krylov matrix**, for which (Meurant and Duintjer Tebbens, 2020) the inner product $K_m K_m^\top$ may be very ill-conditioned. Tyrtyshnikov (1994) proved Krylov matrices have exponential lower bounds in m for their condition number.

Krylov Conditioning

- Gonon et al. (2020) also showed that

$$\boxed{\mathrm{MC} = \mathrm{rank}(K_N).}$$

- For N large, let us consider the Krylov matrix QR decomposition

$$K_N = (\mathbf{q}_1|\mathbf{q}_2|\dots|\mathbf{q}_N)\begin{pmatrix} r_{1,1} & r_{1,2} & \dots & r_{1,N} \\ 0 & r_{2,2} & \dots & r_{2,N} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & r_{N,N} \end{pmatrix} = QR.$$

- $r_{j,j}$ represent the norm of the orthogonal component in vector $A^j \boldsymbol{C}$ with respect to the subspace spanned by the columns of K_{j-1}.
- In practice, we observe that $r_{j,j}$ decay **superexponentially** compared to $\rho(A)^j$
- We call this phenomenon **Krylov subspace squeezing**.

Krylov Subspace Squeezing

Definition

The ***jth-order Krylov subspace*** generated by a matrix $A \in \mathbb{M}_N$ and a vector $\boldsymbol{C} \in \mathbb{R}^N$ is the linear subspace of $\mathbb{R}^N$ given by

$$\mathcal{K}_j(A, \boldsymbol{C}) = \operatorname{span}\left\{\boldsymbol{C}, A\boldsymbol{C}, A^2\boldsymbol{C}, \ldots, A^{j-1}\boldsymbol{C}\right\}.$$

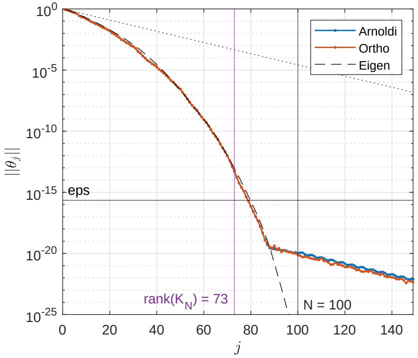
- Let $j \in \mathbb{N}$ and denote as $\boldsymbol{\theta}_j = \operatorname{perp}_{\mathcal{K}_j(A,\boldsymbol{C})}(A^j\boldsymbol{C}) \in \mathcal{K}_j(A, \boldsymbol{C})^\perp$ the orthogonal component of $A^j\boldsymbol{C}$ with respect to $\mathcal{K}_j(A, \boldsymbol{C})$.
- Take the singular value decomposition decomposition of K_j,

$$K_j = U_j \, \Sigma_j \, V_j^\top.$$

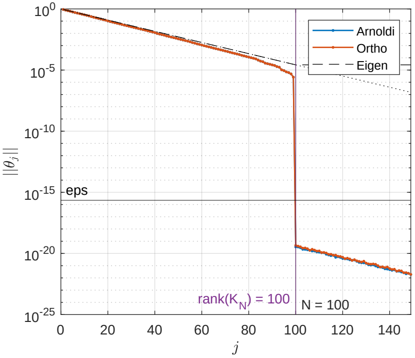
- Orthogonal components $\boldsymbol{\theta}_j$ for every $j \in \{1, \ldots, N\}$ have norm

$$\|\boldsymbol{\theta}_j\| = \|A^j\boldsymbol{C}\,(\mathbb{I}_N - U_{j-1}U_{j-1}^\top)\|.$$

Krylov Subspace Squeezing



(a) $A_{i,j} \sim$ i.i.d. $\mathcal{U}(-1, 1)$, $\rho(A) = 0.9$



(b) $A \sim \mathcal{O}$, $\rho(A) = 0.9$

Figure: Krylov subspace squeezing effects as measured using the norm of the orthogonal component. For Krylov matrix $K_m \in \mathbb{R}^{N \times m}$, $N = 100$ and $m = 5N$. Input mask is $C = [1, \ldots, 1]^\top \in \mathbb{R}^N$. The black dotted line shows the exponential decay of leading eigenvalue $\rho(A)$.

Approximate Subspace Decay

- Empirical conjecture regarding the decay of $\|\boldsymbol{\theta}_j\|$:

  The value of $\|\boldsymbol{\theta}_j\|$ as a function of j is well approximated by the **ordered cumulative product** of the absolute values of eigenvalues of A.

- When A is drawn randomly from standard matrix ensembles, approximately

$$|\lambda_i|^2 \sim \mathcal{U}(0, \rho(A)), \quad i \in \{1, \ldots, N\}.$$

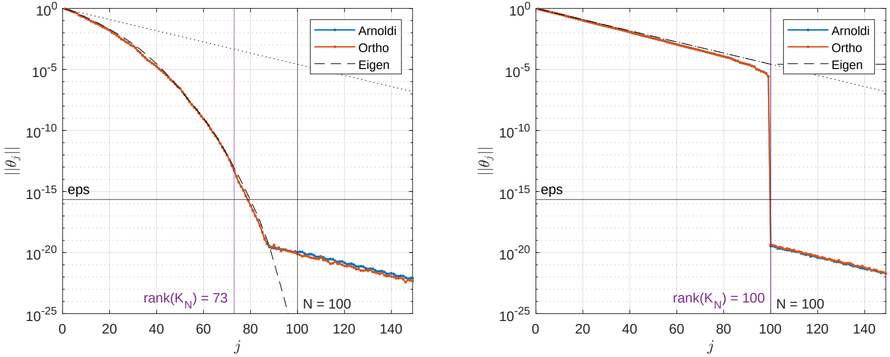
  - Follows from RMT (Tao, 2012, Wood, 2012, Basak and Rudelson, 2019).
  - Then $|\lambda_i|$ are approximately distributed as $\sqrt{\rho(A) Z_i}$ where $Z_i \sim \mathcal{U}(0,1)$.
  - When N is large, $(Z_i)_{i=1}^N$ are approximately uniformly spaced over $(0,1)$.

- Thus a closed-form approximation κ_j for the value of $\|\boldsymbol{\theta}_j\|$ is

$$\boxed{\kappa_j = \sqrt{\rho(A)\,\frac{N!}{N^j (N-j)!}}\,.} \tag{6}$$

Eigenvalue Plots

Krylov Subspace Squeezing



(a) $A_{i,j} \sim$ i.i.d. $\mathcal{U}(-1,1)$, $\rho(A) = 0.9$

(b) $A \sim \mathcal{O}$, $\rho(A) = 0.9$

Figure: Krylov subspace squeezing effects as measured using the norm of the orthogonal component. For Krylov matrix $K_m \in \mathbb{R}^{N\times m}$, $N = 100$ and $m = 5N$. Input mask is $C = [1, \ldots, 1]^\top \in \mathbb{R}^N$. The black dotted line shows the exponential decay of leading eigenvalue $\rho(A)$.

Impact on Naïve Methods

- For A with $\rho(A) < 1$ and large enough N there exists a positive integer $\ell < N$ such that **numerically**

$$R \approx \begin{pmatrix} R_1 & R_2 \\ \mathbb{O}_{N-\ell,\ell} & \mathbb{O}_{N-\ell,N-\ell} \end{pmatrix}.$$

- Naïve methods, which do not control for the ill-conditioning of $G_{\boldsymbol{x}}$, estimate

$$\boxed{\mathrm{MC} = \mathrm{rank}(R) \approx \ell.}$$

Table of Contents

A Memory Neutrality Result

- We prove the following:

Proposition: Input Mask Memory Neutrality

For any *linear* ESN, under the assumptions of the "Perfect Memory" Proposition, the memory capacity is **input mask neutral**, that is, MC_τ is invariant with respect to the choice of $\boldsymbol{C}$, for all $\tau \in \mathbb{N}$.

▸ Proof

- A complementary result in continuous time was derived by Hermans and Schrauwen (2010).

- We exploit input mask neutrality to estimate MC in linear ESNs.

Orthogonalized Subspace Method

- Let $K_m = U \Sigma V^\top$ again be the SVD decomposition of the Krylov matrix.
- Define the projection operator $P : \mathbb{R}^m \to \mathcal{K}_N(A, \mathbf{C})$ with associated matrix

$$P = K_m^\top \left(K_m K_m^\top \right)^{-1} K_m.$$

- By construction we have that $P_{\tau,\tau} \approx \mathrm{MC}_\tau$:

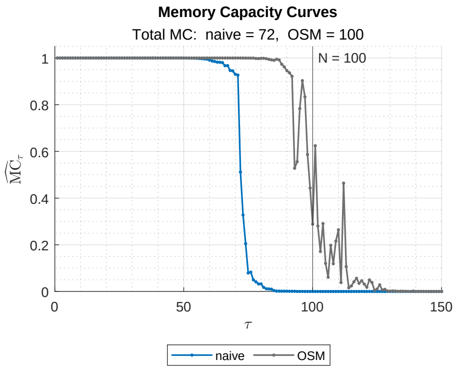
$$\begin{aligned} K_m^\top \left(K_m K_m^\top \right)^{-1} K_m &= V \Sigma U^\top \left(U \Sigma V^\top V \Sigma U^\top \right)^{-1} U \Sigma V^\top \\ &= V(\Sigma U^\top U \Sigma^{-1})(V^\top V)^{-1}(\Sigma^{-1} U^\top U \Sigma) V^\top \\ &= V V^\top. \end{aligned}$$

- The τ-lag memory capacity is well approximated by the diagonal entries of P,

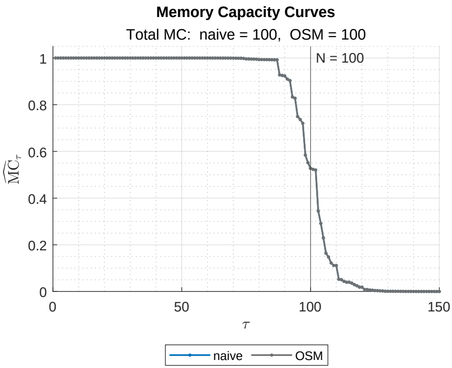
$$\mathrm{MC}_\tau = \left[V V^\top \right]_{\tau,\tau}. \tag{7}$$

- We term this approach the **orthogonalized subspace method** (OSM).

Naive vs OSM



(a) $A \sim$ i.i.d. $\mathcal{U}(-1, 1)$



(b) $A \sim \mathcal{O}$

Figure: Memory capacity curves of LESNs with $A = [A_{i,j}] \in \mathbb{R}^{N\times N}$, $\rho_A = 0.9$, and $C = [c_i] \in \mathbb{R}^N$. In all panels $c_i \sim$ i.i.d. $\mathcal{N}(0, 1)$. Total MC is computed as the ratio between the sum of MC_τ's up to $1.5 \times N$ terms and reservoir size N.

OSM+

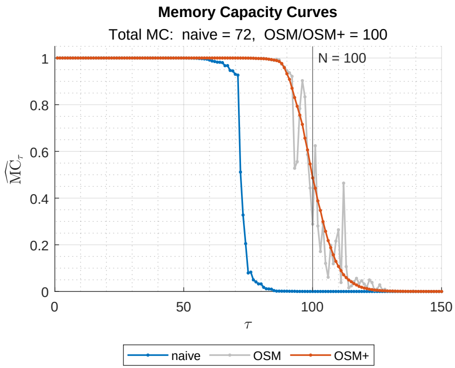
- Jaeger (2002) already proved that MC_τ must be **monotonic decreasing** in τ.
- One issue with OSM is that it yields MC curves which need not be monotonic.

- We also propose an *improved* version of OSM, called **OSM+**:
  - Since MC_τ is neutral to $\boldsymbol{C}$, we can simply *resample it* (keeping A fixed)
  - Draw $\{\boldsymbol{C}_1, \ldots, \boldsymbol{C}_L\}$ and estimate memory using

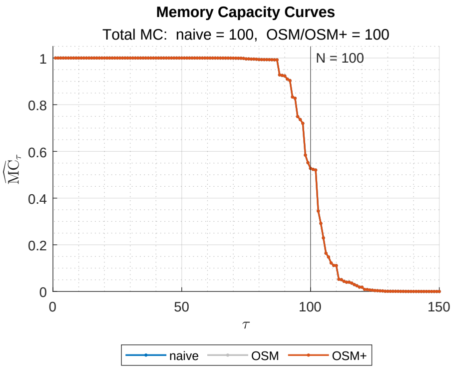
$$\boxed{\mathrm{MC}_\tau = \frac{1}{L}\sum_{\ell=1}^{L}\left[V(\boldsymbol{C}_\ell)V(\boldsymbol{C}_\ell)^\top\right]_{\tau,\tau},} \tag{8}$$

where $V(\boldsymbol{C}_\ell)^\top$ are the right singular vectors of Krylov matrix $K_{m,\ell} = K_m(A, \boldsymbol{C}_\ell)$

Naive vs OSM vs OSM+



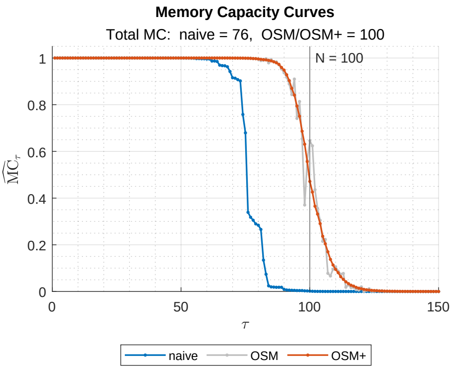
(a) $A \sim$ i.i.d. $\mathcal{U}(-1, 1)$



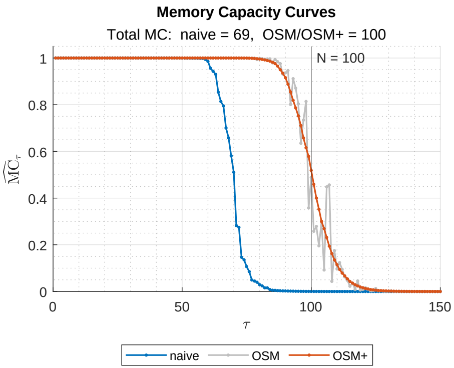
(b) $A \sim \mathcal{O}$

Figure: Memory capacity curves of LESNs with $A = [A_{i,j}] \in \mathbb{R}^{N\times N}$, $\rho_A = 0.9$, and $C = [c_i] \in \mathbb{R}^N$. In all panels $c_i \sim$ i.i.d. $\mathcal{N}(0, 1)$. C is resampled $K = 1000$ times and normalized to compute the average memory curve. Total MC is computed as the ratio between the sum of MC_τ's up to $1.5 \times N$ terms and reservoir size N.

Naive vs OSM vs OSM+



(a) $A \sim$ i.i.d. $\mathcal{N}(0,1)$



(b) $A \sim$ i.i.d. $sp\mathcal{N}(0,1,0.1)$

Figure: Memory capacity curves of LESNs with $A = [A_{i,j}] \in \mathbb{R}^{N\times N}$, $\rho_A = 0.9$, and $C = [c_i] \in \mathbb{R}^N$. In all panels $c_i \sim$ i.i.d. $\mathcal{N}(0,1)$. C is resampled $K = 1000$ times and normalized to compute the average memory curve. Total MC is computed as the ratio between the sum of MC_τ's up to $1.5 \times N$ terms and reservoir size N.

Table of Contents

Conclusion

In this work:

1. Focused explaining and providing solutions for what we call the linear memory gap by demonstrating that this discrepancy arises due to numerical artifacts that have been overlooked in previous studies.

2. Our findings suggest that previous efforts to optimize memory capacity for linear recurrent networks may have been plagued with numerical artifacts, leading to incorrect results.

3. Propose robust techniques for the accurate estimation of memory capacity, which result in full memory results for linear RNNs, as should be generically expected.

Coming Soon: For *nonlinear* ESNs (RNNs), the bound $1 \leq \mathrm{MC} \leq N$ is actually **sharp** i.e. memory capacity is input-dependent!

Thank You

References I

P. Barancok and I. Farkas. Memory capacity of input-driven echo state networks at the edge of chaos. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, pages 41–48, 2014.

A. Basak and M. Rudelson. The circular law for sparse non-Hermitian matrices. *The Annals of Probability*, 4(47): 2359–2416, 2019.

R. Couillet, G. Wainrib, H. Sevi, and H. T. Ali. The asymptotic performance of linear echo state neural networks. *Journal of Machine Learning Research*, 17(178):1–35, 2016.

J. Dambre, D. Verstraeten, B. Schrauwen, and S. Massar. Information processing capacity of dynamical systems. *Scientific reports*, 2(514), 2012. doi: doi:10.1038/srep00514.

I. Farkas, R. Bosak, and P. Gergel. Computational analysis of memory capacity in echo state networks. *Neural Networks*, 83:109–120, 2016. ISSN 18792782. doi: 10.1016/j.neunet.2016.07.012.

C. Gallicchio. Short-term memory of Deep RNN. 2018.

C. Gallicchio, A. Micheli, and L. Pedrelli. Deep reservoir computing: a critical experimental analysis. *Neurocomputing*, (April):87–99, 2017.

S. Ganguli, D. Huh, and H. Sompolinsky. Memory traces in dynamical systems. *Proceedings of the National Academy of Sciences of the United States of America*, 105(48):18970–5, dec 2008. ISSN 1091-6490.

L. Gonon, L. Grigoryeva, and J.-P. Ortega. Memory and forecasting capacities of nonlinear recurrent networks. *Physica D*, 414(132721):1–13., 2020.

A. Goudarzi, S. Marzen, P. Banda, G. Feldman, M. R. Lakin, C. Teuscher, and D. Stefanovic. Memory and information processing in recurrent neural networks. Technical report, 2016. URL `https://arxiv.org/pdf/1604.06929.pdf`.

L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. Optimal nonlinear information processing capacity in delay-based reservoir computers. *Scientific Reports*, 5(12858):1–11, 2015. doi: 10.1038/srep12858.

References II

L. Grigoryeva, J. Henriques, L. Larger, and J.-P. Ortega. Nonlinear memory capacity of parallel time-delay reservoir computers in the processing of multidimensional signals. *Neural Computation*, 28:1411–1451, 2016.

M. Hermans and B. Schrauwen. Memory in linear recurrent neural networks in continuous time. *Neural Networks*, 23 (3):341–55, apr 2010. ISSN 1879-2782.

H. Jaeger. Short term memory in echo state networks. *Fraunhofer Institute for Autonomous Intelligent Systems. Technical Report.*, 152, 2002.

H. Jaeger and H. Haas. Harnessing Nonlinearity: Predicting Chaotic Systems and Saving Energy in Wireless Communication. *Science*, 304(5667):78–80, 2004. doi: 10.1126/science.1091277.

P. Lax. *Functional Analysis*. Wiley-Interscience, 2002.

L. Livi, F. M. Bianchi, and C. Alippi. Determination of the edge of criticality in echo state networks through Fisher information maximization. 2016.

M. Matthews and G. Moschytz. The identification of nonlinear discrete-time fading-memory systems using neural network models. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 41(11): 740–751, 1994. ISSN 10577130. doi: 10.1109/82.331544. URL `http://ieeexplore.ieee.org/document/331544/`.

M. B. Matthews. *On the Uniform Approximation of Nonlinear Discrete-Time Fading-Memory Systems Using Neural Network Models*. PhD thesis, ETH Zürich, 1992. URL `https://www.research-collection.ethz.ch:443/handle/20.500.11850/140592`.

G. Meurant and J. Duintjer Tebbens. *Krylov Methods for Nonsymmetric Linear Systems*. Springer International Publishing, 2020.

A. Rodan and P. Tino. Minimum complexity echo state network. *IEEE Transactions on Neural Networks*, 22(1): 131–44, jan 2011. ISSN 1941-0093.

References III

A. Rodan and P. Tino. Simple deterministically constructed cycle reservoirs with regular jumps. *Neural Computation*, 24(7):1822–1852, 2012.

T. Tao. *Topics in Random Matrix Theory*. American Mathematical Society, 2012.

T. Tao, V. Vu, and M. Krishnapur. Random matrices: Universality of ESDs and the circular law. *The Annals of Probability*, 38(5):2023–2065, 2010.

P. Tino. Asymptotic Fisher memory of randomized linear symmetric echo state networks. *Neurocomputing*, 298:4–8, 2018. ISSN 18728286. doi: 10.1016/j.neucom.2017.11.076. URL `https://doi.org/10.1016/j.neucom.2017.11.076`.

P. Tino and A. Rodan. Short term memory in input-driven linear dynamical systems. *Neurocomputing*, 112:58–63, 2013.

E. E. Tyrtyshnikov. How bad are Hankel matrices? *Numerische Mathematik*, 67(2):261–269, 1994.

P. Verzelli, C. Alippi, and L. Livi. Echo State Networks with self-normalizing activations on the hyper-sphere. *Scientific Reports*, 9(13887), 2019.

O. White, D. Lee, and H. Sompolinsky. Short-term memory in orthogonal neural networks. *Physical Review Letters*, 92(14):148102, apr 2004. ISSN 0031-9007.

P. M. Wood. Universality and the circular law for sparse random matrices. *The Annals of Probability*, 22(3): 1266–1300, 2012.

F. Xue, Q. Li, and X. Li. The combination of circle topology and leaky integrator neurons remarkably improves the performance of echo state network on time series prediction. *PloS one*, 12(7):e0181816, 2017. ISSN 1932-6203. doi: 10.1371/journal.pone.0181816. URL `http://www.ncbi.nlm.nih.gov/pubmed/28759581http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5536322`.

Memory: Orthogonal A, Gaussian C, More Terms

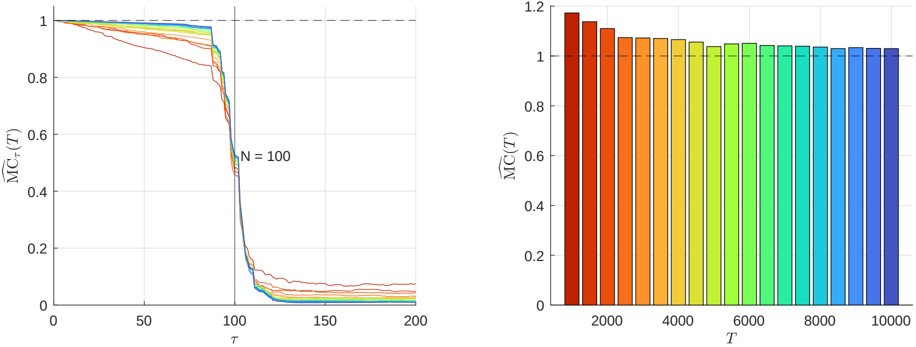


Figure: All memory curves $\widehat{\mathrm{MC}}_\tau(T)$ are computed for $\tau \in \{0, 1, \ldots, \mathbf{5N}\}$ but plot is shortened for clarity.

Orthogonal connectivity matrix $A = [A_{i,j}] \in \mathbb{R}^{N\times N}$ of size $N = 100$ and spectral radius $\rho(A) = 0.9$.

Monte Carlo Simulation Bias

- For simplicity, suppose that the ESN system is **regular**, i.e. $\Gamma_{\mathbf{x}} = \mathbb{I}_N$, and that $\mathbb{E}(z_t) = 0$ and $\mathrm{Var}(z_t) = 1$.
- Let $\gamma_{\mathbf{x}z}(\tau) := \mathrm{Cov}(\mathbf{x}_t, z_{t-\tau})$, then

$$\widehat{\mathrm{MC}}_\tau(T) = \left\| \widehat{\gamma_{\mathbf{x}z}}(\tau) \right\|_2^2 = \left\| \frac{1}{T-\tau} \sum_{t=\tau+1}^{T} \mathbf{x}_t \, z_{t-\tau} \right\|_2^2, \tag{9}$$

$$\widehat{\mathrm{MC}}(T) = \frac{1}{\tau_{\max}} \sum_{\tau=0}^{\tau_{\max}-1} \widehat{\mathrm{MC}}_\tau(T). \tag{10}$$

- We show that $\widehat{\mathrm{MC}}_\tau(T)$ is consistent, but has **positive bias** for τ large,

$$B_{MC} := \mathbb{E}\left[\widehat{\mathrm{MC}}_\tau(T)\right] - \mathrm{MC}_\tau = \boxed{\frac{N}{T-\tau}} + \underbrace{\frac{2}{T-\tau} \sum_{j=0}^{\tau} \gamma_{\mathbf{x}z}(j)^\top \gamma_{\mathbf{x}z}(2\tau - j)}_{\text{exponentially decaying in } \tau}.$$

- Even if T is large, summing $\tau_{\max}$ terms of $\{\mathrm{MC}_\tau\}_{\tau=0}^{\infty}$ can yield $\mathrm{MC} > N$!

Example: Cyclic Reservoir I

- Consider a N-dimensional cyclic reservoir with the unscaled orthogonal connectivity matrix

$$\widetilde{A} = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 \\ 1 & 0 & \ddots & 0 & 0 \\ 0 & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & 0 & 0 \\ 0 & \dots & 0 & 1 & 0 \end{pmatrix} \in \mathbb{M}_N,$$

which is rescaled with some $\rho_A < 1$ by setting $A = \rho_A \widetilde{A}$.

- Let $\boldsymbol{C} = \boldsymbol{e}_1$ and note

$$A\boldsymbol{C} = \boldsymbol{e}_2, \quad A^2\boldsymbol{C} = \boldsymbol{e}_3, \dots, A^{N-1}\boldsymbol{C} = \boldsymbol{e}_N,$$

Back

Example: Cyclic Reservoir II

- For a cyclic reservoir with $\boldsymbol{C} = \boldsymbol{e}_1$ it thus holds

$$G_{\mathbf{x}} = \operatorname{diag}\left(\frac{1}{1-\rho_A^{2N}}, \frac{\rho_A^2}{1-\rho_A^{2N}}, \ldots, \frac{\rho_A^{2(N-1)}}{1-\rho_A^{2N}}\right)$$

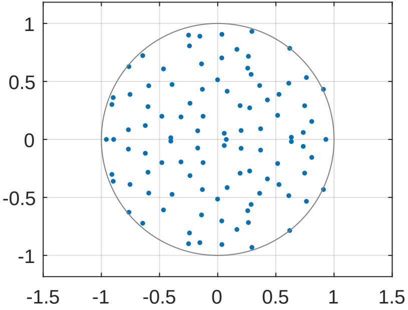
  and hence

$$G_{\mathbf{x}}^{-1} = \operatorname{diag}\left(1-\rho_A^{2N},\ \frac{1-\rho_A^{2N}}{\rho_A^2},\ \ldots\ ,\ \frac{1-\rho_A^{2N}}{\rho_A^{2(N-1)}}\right).$$

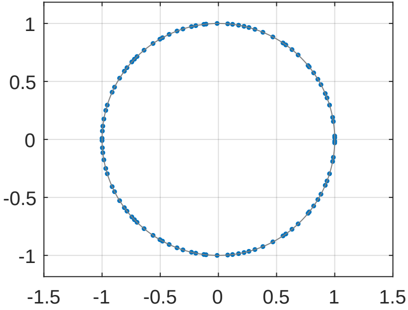
- When N is large, inversion of $G_{\mathbf{x}}$ quickly becomes an ill-conditioned problem (depending on ρ_A).
- In this simple, special setup one can also easily express MC_τ analytically for each $\tau \geq 0$, see also Rodan and Tino (2011).

Back

Eigenvalue Plots



(a) $A_{ij} \sim$ i.i.d. $\mathcal{N}(0,1)$



(b) $A \sim \mathcal{O}$

Figure: Eigenvalues (blue) for random and non-random reservoir matrices and the complex unit circle (gray), $N = 100$. For specification with entries $A_{ij} \sim$ i.i.d. $\mathcal{N}$ matrix is normalized according to circular law rate $N^{-1/2}$.

Back

Proof: Input Mask Memory Neutrality I

Let $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N\}$ be an eigenbasis of A and $\{\lambda_1, \ldots, \lambda_N\}$ be the associated eigenvalues. Denote $\Lambda := \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$, $V := (\boldsymbol{v}_1 | \boldsymbol{v}_2 | \ldots | \boldsymbol{v}_N)$, and

$$V^{-1} = \begin{pmatrix} \boldsymbol{v}_1^* \\ \vdots \\ \boldsymbol{v}_N^* \end{pmatrix},$$

and notice that by the hypothesis of diagonalizability of A one has $A = V \Lambda V^{-1}$.

Using the eigenbasis of A, or using the columns of V, it holds for the input mask that $\boldsymbol{C} = \sum_{i=1}^{N} c_i \, \boldsymbol{v}_i$ with $\boldsymbol{c} := (c_1, \ldots, c_N)^\top$ the vector of coefficients.

Recall that

$$G_{\boldsymbol{x}} = \sum_{j=0}^{\infty} A^j \boldsymbol{C}\boldsymbol{C}^\top (A^j)^\top = \sum_{i,j=1}^{N} \varphi_{i,j} \, \boldsymbol{v}_i \, \boldsymbol{v}_j^*,$$

with $\varphi_{i,j} := (c_i \overline{c}_j)/(1 - \lambda_i \overline{\lambda}_j)$.

Proof: Input Mask Memory Neutrality II

Hence it holds that

$$V^{-1} G_{\boldsymbol{x}} (V^*)^{-1} = \left(\sum_{i,j=1}^{N} \varphi_{i,j} \left(\boldsymbol{v}_k^* \, \boldsymbol{v}_i \, \boldsymbol{v}_j^* \, \boldsymbol{v}_l \right) \right)_{k,l}^{N} = (\varphi_{k,l})_{k,l}^{N} .$$

Finally, using this expression in (5), we can write MC_τ as follows:

$$\begin{aligned}
\mathrm{MC}_\tau &= \boldsymbol{C}^\top (A^\tau)^\top G_{\boldsymbol{x}}^{-1} A^\tau \boldsymbol{C} \\
&= \boldsymbol{C}^\top (V^*)^{-1} (\Lambda^*)^\tau V^* G_{\boldsymbol{x}}^{-1} V \Lambda^\tau V^{-1} \boldsymbol{C} \\
&= \boldsymbol{C}^\top (V^{-1})^* (\Lambda^*)^\tau \left((\varphi_{k,l})_{k,l}^{N} \right)^{-1} \Lambda^\tau V^{-1} \boldsymbol{C} \\
&= \boldsymbol{c}^* (\Lambda^*)^\tau \left((\varphi_{k,l})_{k,l}^{N} \right)^{-1} \Lambda^\tau \boldsymbol{c} \\
&= \boldsymbol{c}^* (\Lambda^*)^\tau \left(\operatorname{diag}(\boldsymbol{c}) \left(\frac{1}{1 - \lambda_k \bar{\lambda}_l} \right)_{k,l}^{N} \operatorname{diag}(\boldsymbol{c}^*) \right)^{-1} \Lambda^\tau \boldsymbol{c}
\end{aligned}$$

Back

Proof: Input Mask Memory Neutrality III

[cont'd]

$$\begin{aligned}
\mathrm{MC}_\tau &= \boldsymbol{c}^*(\Lambda^*)^\tau \mathrm{diag}\left(\boldsymbol{c}^*\right)^{-1}\left(\left(\frac{1}{1-\lambda_k\overline{\lambda}_l}\right)_{k,l}^N\right)^{-1}\mathrm{diag}\left(\boldsymbol{c}\right)^{-1}\Lambda^\tau\boldsymbol{c}\\
&= \boldsymbol{\iota}_N^\top\left(\Lambda^*\right)^\tau\left(\left(\frac{1}{1-\lambda_k\overline{\lambda}_l}\right)_{k,l}^N\right)^{-1}\Lambda^\tau\,\boldsymbol{\iota}_N,
\end{aligned}$$

where $\boldsymbol{\iota}_N = (1,\ldots,1)^\top \in \mathbb{R}^N$. The last equality in the derivation follows from the commutative property of the product of diagonal matrices.

Hence, MC_τ is independent of $\boldsymbol{C}$ for all $\tau \in \mathbb{N}$ under the stated assumptions.

Back

Key References

- **Memory Capacity**
  Jaeger (2002), Matthews (1992), Matthews and Moschytz (1994), Jaeger and Haas (2004)
  - **Linear ESN:** Hermans and Schrauwen (2010), Dambre et al. (2012), Barancok and Farkas (2014), Couillet et al. (2016), Goudarzi et al. (2016), Xue et al. (2017)
  - **Shallow ESN:** White et al. (2004), Farkas et al. (2016), Verzelli et al. (2019)
  - **Deep ESN:** Gallicchio et al. (2017), Gallicchio (2018)
- **Theoretical Properties:**
  Hermans and Schrauwen (2010), Rodan and Tino (2011, 2012), Tino and Rodan (2013), Grigoryeva et al. (2015, 2016), Gonon et al. (2020)
- **Fisher Memory:**
  Ganguli et al. (2008), Tino and Rodan (2013), Livi et al. (2016), Tino (2018)