ESN for DA

Quentin BRUANT, Barbara DALENA (CEA/IRFU)

Future Circular Collider– France/Italy

November 4, 2024

Thanks to the work of M.Casanova (CEA/IRFU-Politecnico), M.Giovanozzi (CERN), L.Bonaventura (Politecnico di Milano)

Table of contents

1. Dynamic Aperture (DA)

Table of contents

Dynamic Aperture

Bonded movement phase-space

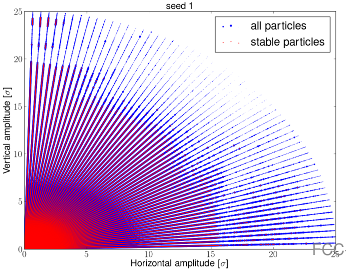
Bonded movement phase-space region

$$\mathcal{A}(N) = \int\int\int\int \chi(x_1, x_2, p_{x_1}, p_{x_2}) dx_1 dx_2 dp_{x_1} dp_{x_2}$$

$$\chi(x_1, x_2, p_{x_1}, p_{x_2}) = \begin{cases} 1 \text{ if the motion starting at } (x_1, x_2) \\ \text{with momentum } (p_{x_1}, p_{x_2}) \text{ is bounded after } N \text{ turns} \\ 0 \text{ else} \end{cases}$$

Illustration with $(p_{x_1}, p_{x_2}) = (0, 0) \Rightarrow$

$$DA = \left(\frac{2\mathcal{A}(N)}{\pi^2}\right)^{\frac{1}{4}}$$



DA and Scaling Law

Scaling Law (SL)[1]

$$DA(N) = \frac{b}{\left(\mathcal{B} \ln \frac{N}{N_0}\right)^{\kappa}}$$

SL derivative

$$\frac{dDA}{dN}(N) = -\frac{bk}{N}\mathcal{B}^{-k} \ln\left(\frac{N}{N_0}\right)^{-k-1}$$

2. Echo State Network and DA

Table of contents

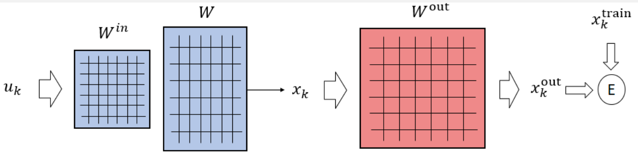
Echo State Network

- Recurrent Neural Network (RNN) using the Reservoir Computing approach.
- Fast training
- Usually less prone to over-fit
- Predictive algorithm

Universal approximator of Dynamical Systems [5]

Shallow ESN

Shallow ESN



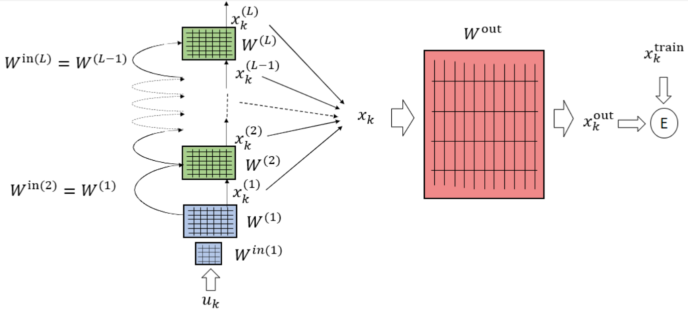
$$x_k = \left(1 - a\frac{\Delta t}{c}\right) x_{k-1} + \frac{\Delta t}{c} f(W^{in} u_k + W x_{k-1})$$

$$x_k^{out} = g(W^{out}[x_k; u_k])$$

$$W^{out} = \text{argmin} \left\{ \frac{1}{M} \sum_{i=0}^{M} \left(\sum_{k=Bl+1}^{k_{train}} (x_{ik}^{out} - x_{ik}^{train})^2 - \beta ||w_i^{out}||^2 \right) \right\}$$

$$W^{out} = X^{train} X^T (XX^T + \beta I)^{-1}$$

Deep ESN



$$x_k^{(l)} = \left(1 - a\frac{\Delta t}{c}\right) x_{k-1}^{(l)} + \frac{\Delta t}{c} f\left(W^{(l-1)} x_k^{(l-1)} + W^{(l)} x_{k-1}\right), l \geq 1$$

Echo State Property (ESP)

Echo State Property [4]: assure stability for zero-input case

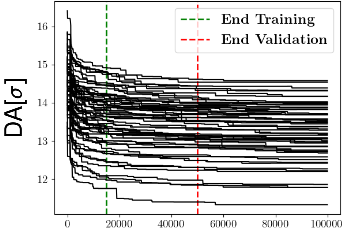
$$\tilde{W} = \frac{\Delta t}{c}|W| + \left(1 - a\frac{\Delta t}{c}\right) I \quad (3)$$

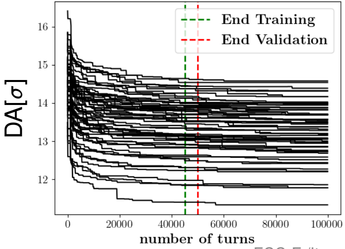
$$\rho(\tilde{W}) < 1 \wedge f : \mathbb{R}^{N_r} \to \mathbb{R}^{N_r}, x \mapsto \tanh x \quad (4)$$

Sufficient, not necessary

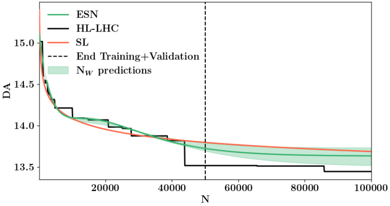
Data set & splitting

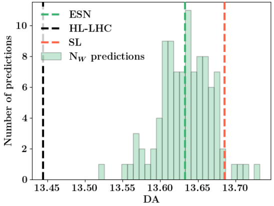
- High-Luminosity LHC, simulated
- 60 configurations→ 60 distributions in magnetic lattice
- nominal parameters ($\sqrt{s}$ = 14 TeV, $\beta^*_{x,y}$ = 15cm, $\epsilon_{x,y}$ = 3.75 µm)
- DA is shown as normalized to the σ of the beam





Ensemble approach





Ensemble approach

- Generate randomly N_w pair of (W_{in}, W)
- Train and validate all the pairs on the same data set and find a set of hyperparameters minimizing the RRMSE [1] on the validation set for all configurations
- For each configuration, compute the RRMSE w.r.t the data.
- Normalize the prediction on the number N_w of pairs (W_{in}, W)

[1]
$$RRMSE^{set}(x^{target}, x^{out}_{mean})[\%] = \frac{100}{N_w}\sum_{i=1}^{N_w}\sqrt{\frac{\sum_{k=1}^{k_{set}}(x^{out,i}_{mean,k} - x^{target,i}_{k})^2}{\sum_{k=1}^{k_{set}}(x^{target,i}_{k})^2}}$$

Models

SL alone

Benchmark. Fit on the training and validation sets

SL ALL

In use today. SL fitted on all the data set ($N \approx 1 \times 10^5$).

ESN only

Basic model
ESN trained on the train set, hyperparameters' optimization on the validation and predictions on test set

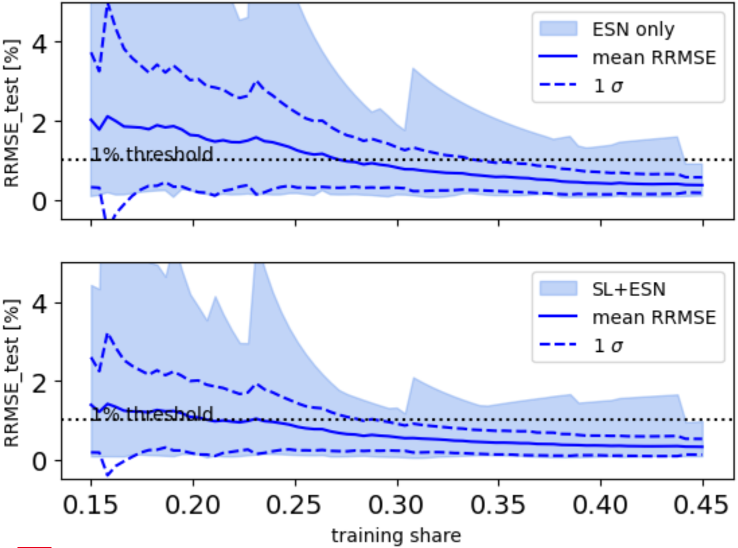
SL+ESN

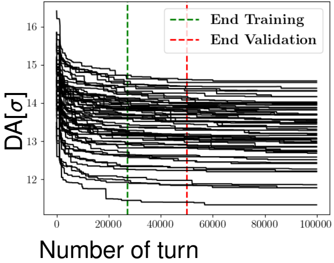
Most evolved model
SL fitted on the training and validation set and on the ESN predictions on the test set

3. Results

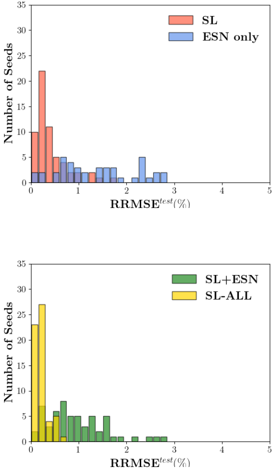
Table of contents

Evolution of error w.r.t training share

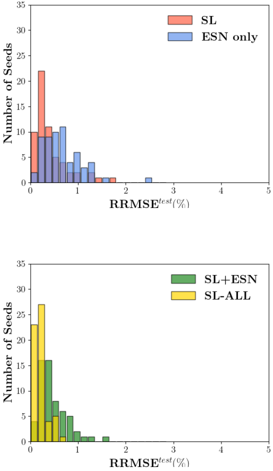




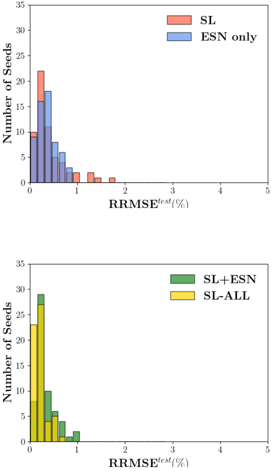
Model Comparison on Projection



(a) 21%

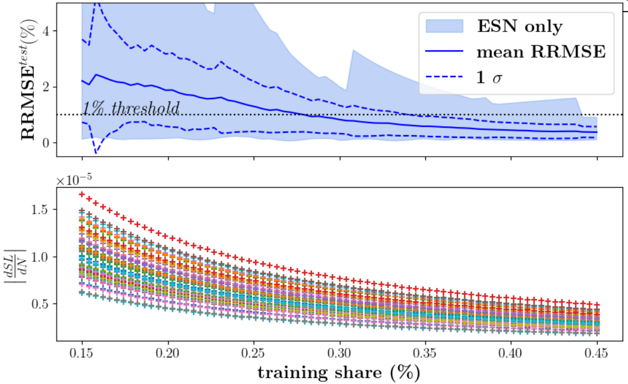


(b) 33%



(c) 45%

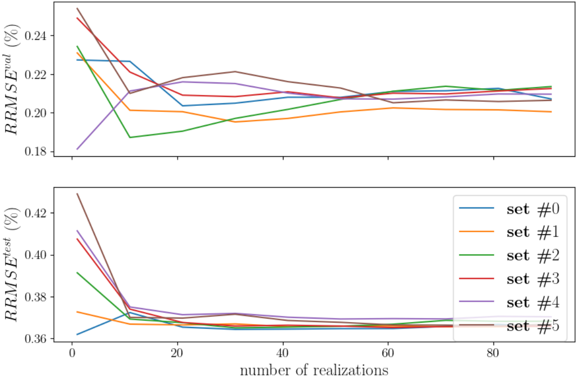
Linking data splitting to SL derivative



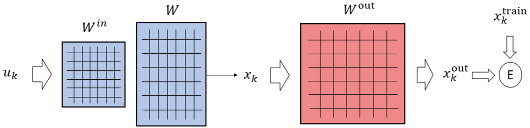
$$\frac{dDA}{dN}(N) = -\frac{bk}{N}\mathcal{B}^{-k}\ln\left(\frac{N}{N_0}\right)^{-k-1}$$

- the 1% mean RRMSE threshold is achieved around the 30% training mark
- Matches with $\left|\frac{dSL}{dN}\right| \approx 5 \times 10^{-6}$
- Matches with $\sigma\left(\left|\frac{dSL}{dN}\right|\right) \approx 2 \times 10^{-6}$

Random variability control



Most stable value for realizations ≈ 60



Parameters: Training share: 45%, Validation: 5%

Hyperparameters (probabilistic systematic optimization)

Main objective: Use Bayesian Optimization to accelerate the finding of the optimum set of hyperparameters
Tool: hyperopt library

Set #	L	N_r	BI	dt	β	ρ	Time(s)	$RRMSE_{max}^{test}$	$RRMSE_{mean}^{test}$
1	2	20	2	0.171	1.387	1.293	14801.7	2.261	0.925
2	2	15	2	0.138	0.379	0.537	18039.4	1.518	0.499
3	1	30	2	0.111	0.618	0.298	7557.7	1.901	0.552
4	1	20	2	0.025	1.877	1.259	6249.9	1.336	0.449
5	1	10	2	0.432	0.561	0.181	10438.5	2.674	1.112

Table: Set of optimized hyperparameters *via* hyperopt

L	N_r	BI	dt	β	ρ	$RRMSE_{max}^{test}$	$RRMSE_{mean}^{test}$
1	20	0	0.009	0.0224	0.19	0.892	0.365

Table: "by-hand" optimized set of parameters

Conclusion

Data Splitting

- The main threshold (1% mean RRMSE) is attained with ≈30% training and no great improvement beyond.

link to SL derivative

- Possible to link Scaling Law derivative to data splitting for automation purposes

Hyperparameters

- hyperopt library find similar set than "by-hand" optimization, the caveat being its sequential implementation taking too much time for a full automation

Perspectives

Data Splitting

- Investigate the outlier configurations

Link to SL derivative

- Try the automation of the splitting based on the 2 parameters described
- Investigate the link to second derivative

Hyperparameters

- Working on a custom-made script for a full parallel optimization

Thank you !

CEA SACLAY
91 191 Gif-sur-Yvette Cedex
France
quentin.bruant@cea.fr
barbara.dalena@cea.fr

References I

[1] A. Bazzani et al. "Advances on the modeling of the time evolution of dynamic aperture of hadron circular accelerators". In: *Phys. Rev. Accel. Beams* 22 (10 Oct. 2019), p. 104003. DOI: 10.1103/PhysRevAccelBeams.22.104003. URL: https://link.aps.org/doi/10.1103/PhysRevAccelBeams.22.104003.

[2] J. Bergstra, D. Yamins, and D. D. Cox. "Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures.". In: 30th International Conference on Machine Learning (ICML 2013). 2013, I-115 to I–23.

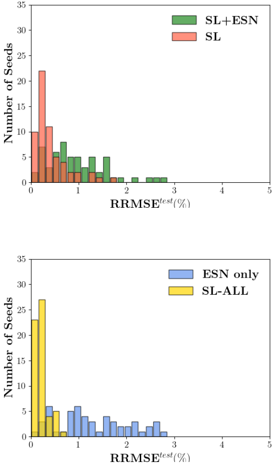
[3] Eric Brochu, Vlad M. Cora, and Nando de Freitas. *A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning*. 2010. arXiv: 1012.2599 [cs.LG].

References II

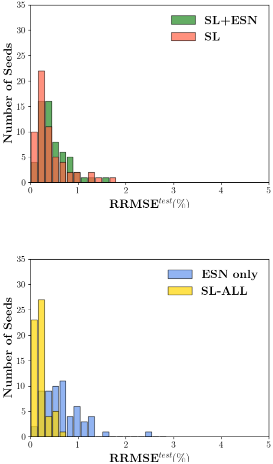
[4] M. Casanova et al. "Ensemble Reservoir Computing for Dynamical Systems: Prediction of Phase-Space Stable Region for Hadron Storage Rings". In: *The European Physical Journal Plus* 138 (2023). DOI: 10.1140/epjp/s13360-023-04167-y. URL: https://doi.org/10.1140/epjp/s13360-023-04167-y.

[5] Lyudmila Grigoryeva and Juan-Pablo Ortega. *Echo state networks are universal*. 2018. arXiv: 1806.00797 [cs.NE].

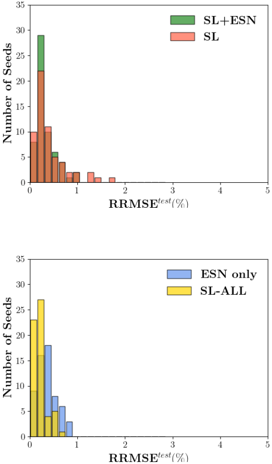
Annex



(a) 21%



(b) 33%



(c) 45%

Annex

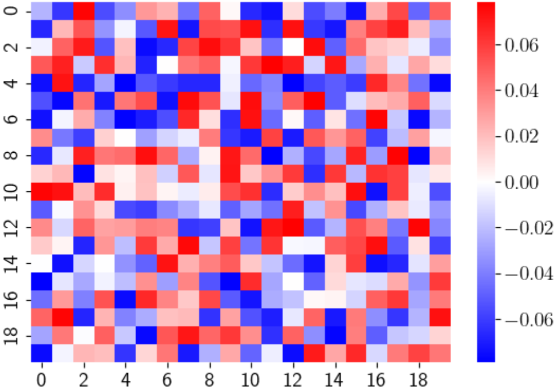


Figure: Reservoir matrix

Annex

Use of Tree Parzen Estimator (TPE)[2, 3].
Bayesian Optimization
Main goal:

$$x_{opt} = f(x)$$

$$P(y|x) \propto P(x|y)P(y)$$

TPE models $P(x|y)$ and $P(y)$ by building "probability-density" functions:

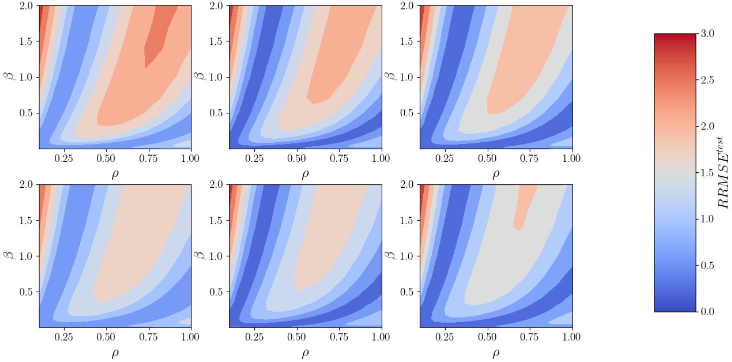
$$P(x|y) = \begin{cases} l(x) \text{ if } y < y* \\ g(x) \text{ if } y > y* \end{cases}$$

where $y* = \min\{f(x_i), 1 \leq i \leq n\}$ But no specific model for $P(y)$ is required

Hyperparameters (mapping)

$$W^{out} = X^{train} X^T (XX^T + \beta I)^{-1}$$

$$\rho = \rho(\tilde{W})$$



Scan of (β, ρ) set of hyperparameters (L = 1, N_r=20, BI=0, dt=0.009)

From Left to Right, Top to Bottom: Training share: 35%, 37%, 39%, 41%, 43%, 45%