A novel Deep Neural Network architecture for non-linear system identification

Università degli Studi di Padova

Luca Zancato, Alessandro Chiuso

University of Padova

Department of Information Engineering, University of Padova

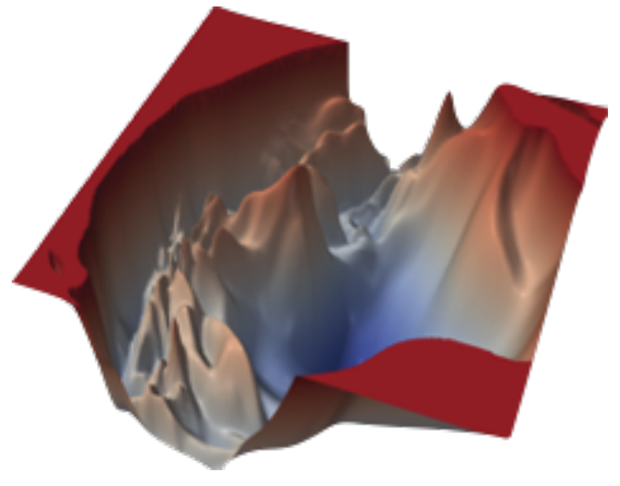
Problem

What:

Non-linear system identification with parametric models: Deep Neural Networks

Why:

- DNNs are universal approximators
- Favorably scales on the large data regime



DNNs Loss Landscape [1]

Challenges:

- Overfitting
- Interpretability

How:

- Inductive bias (on the architecture)
- Regularization (on the loss function)
- Differentiable automatic complexity selection based on available data
- Optimize model and regularization loss with standard Deep Learning primitives

Notation

$u_t \rightarrow$ Non-linear system $\rightarrow$ $y_t = \overbrace{F_0(z_t^-)}^{\text{One-step-ahead predictor}} + \underbrace{e_t}_{\text{Innovation sequence}}$ $\qquad z_t := \begin{bmatrix} y_t \\ u_t \end{bmatrix}$

Goal

Find: $\hat{F} \approx F_0$ given N data from the true system

Remark: F_0 depends on z_t^- (infinite past)

$$(1) \qquad \hat{F} = \arg \min_{F \in \mathscr{F}} \frac{1}{N} \sum_{t=1}^{N} (y_t - F(z_t^-))^2 + \lambda P(F)$$

Where $\mathscr{F}$ is the model class and $P(F)$ is a penalty function

Modeling assumption

$\boxed{\text{fading memory systems can be uniformly approximated on compact sets}}$

$\Downarrow$

We shall consider Neural Networks model class (universal approximators)

Remark:

Fading memory $\implies F_0(z_t^-) \approx F_0(\varphi_{t,T}(z))$

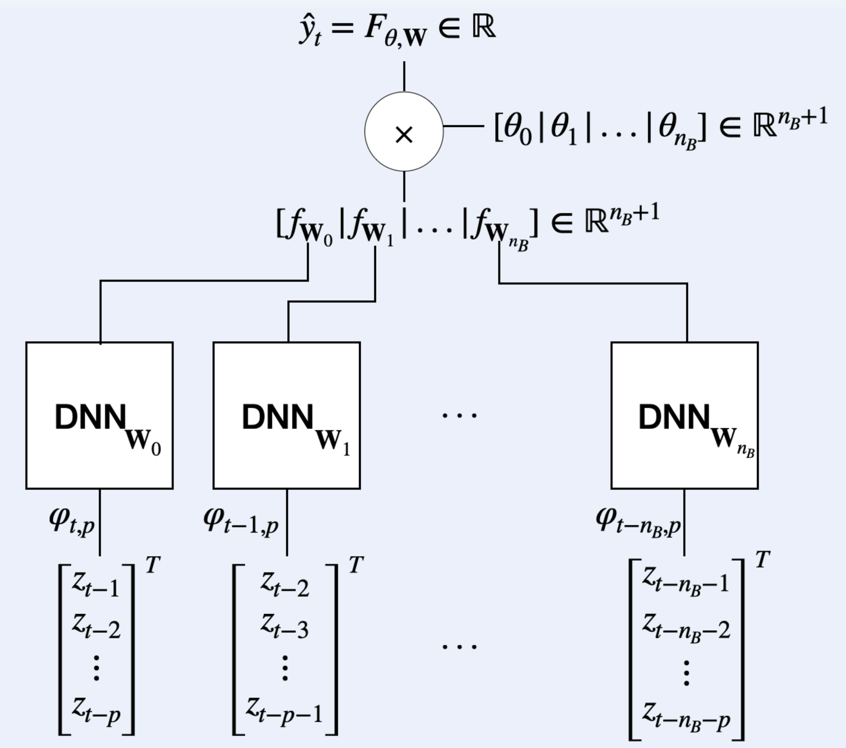
where $\varphi_{t,T}(z)$ is a finite (length T) yet arbitrarily long window of past data w.r.t. t

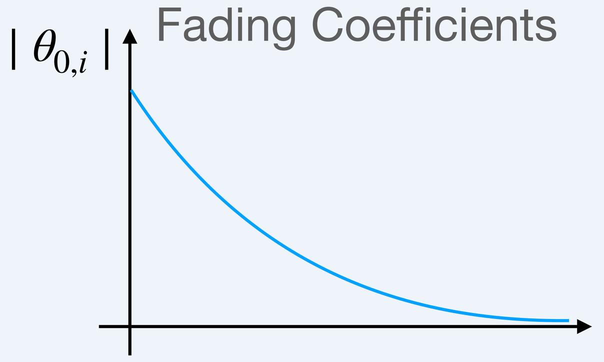
Approach

Fading Memory network architecture

We design a block-structured architecture to encode fading memory [2]

Model class $\qquad (2) \qquad F_{\theta,\mathbf{W}}(z_t^-) = \sum_{i=0}^{n_B} \theta_i f_{\mathbf{W}_i}(\varphi_{t-i,p}(z))$





Architecture Parameters:

- n_B number of blocks
- p size of "elementary" regressor $\varphi_{t-i,p}(z)$
- block weights $\mathbf{W} = \{\mathbf{W}_0, \mathbf{W}_1, \dots, \mathbf{W}_{n_B}\}$
- recombination parameters $\theta_0, \theta_1, \dots, \theta_{n_B}$

Remark: n_B large enough to capture the true memory and should **not** be chosen to face a bias-variance trade-off

Fading Regularization

Model structure design:

- How to choose the number of blocks n_B?
- Automatically choose the right complexity so that only relevant past is considered?

Solution:

Fading Regularization: "large enough" n_B (larger than the true memory) and automatically select the best model complexity to avoid overfitting.

Joint posterior optimization

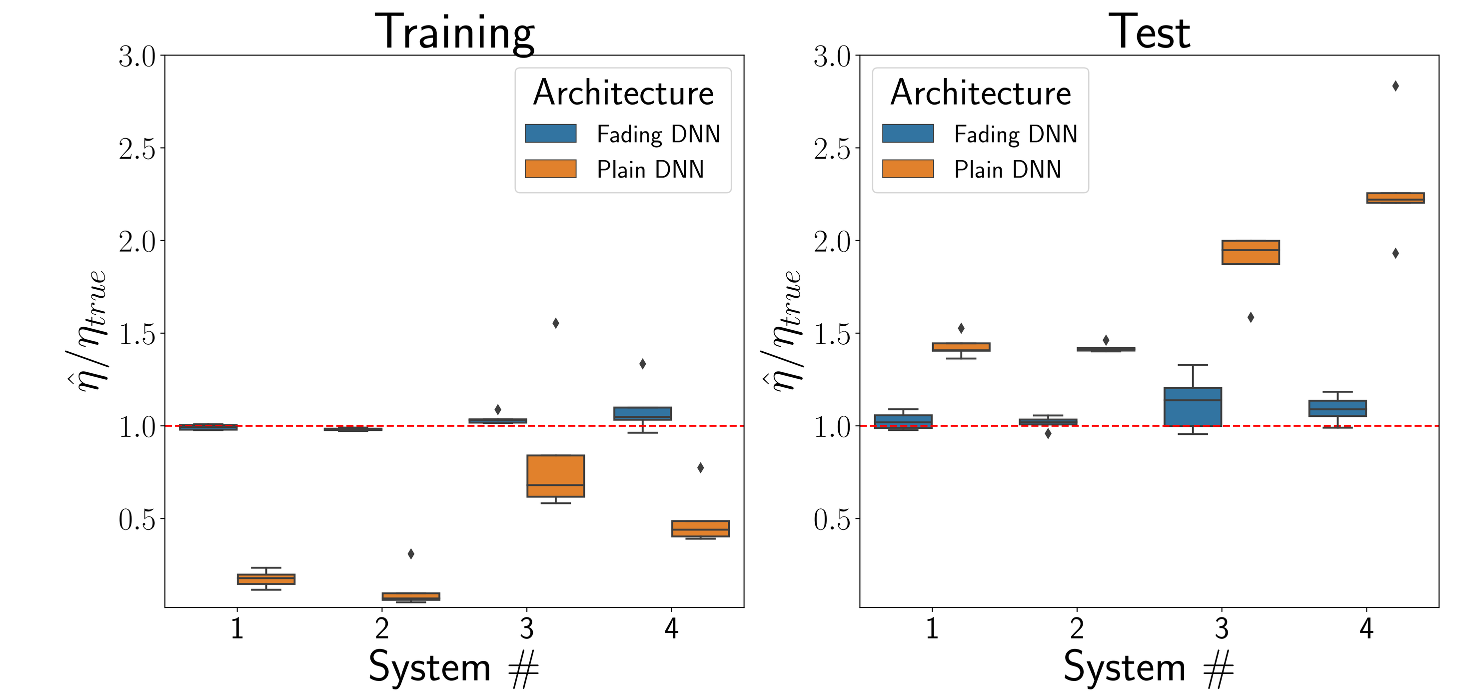
In a Bayesian setting the optimal parameter set is obtained through *maximum a-posteriori*:

$$(3) \qquad \hat{\theta}, \hat{\mathbf{W}}, \hat{\lambda}, \hat{\kappa} = \arg \min_{\theta, \mathbf{W}, \lambda \in (0,1), \kappa > 0} \frac{||Y - \hat{Y}_{\theta,\mathbf{W}}||^2}{\eta^2} + \log(\eta^2) - \log(p(\mathbf{W})) - \log(p_{\lambda,\kappa}(\theta))$$

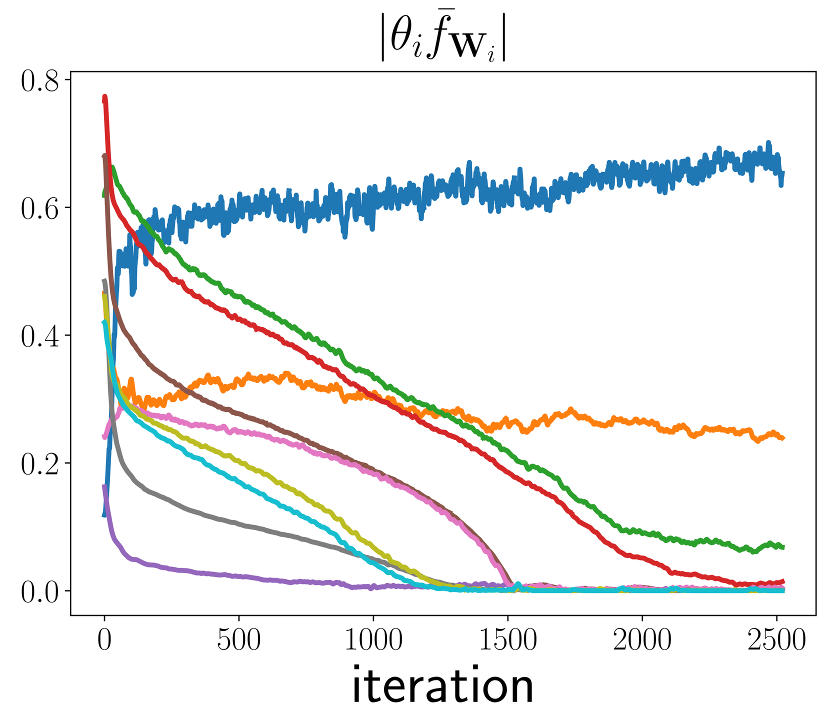
Proposition 1: *The following is an upper bound on the marginal log likelihood associated to the posterior in equation (3) with marginalization taken only w.r.t. θ:*

$$(4) \qquad \mathscr{U} := \frac{1}{\eta^2} ||Y - \hat{Y}_{\theta,\mathbf{W}}||^2 + \theta^\top \Lambda^{-1} \theta + \log p(\mathbf{W}) + \log |F_{\mathbf{W}} \Lambda F_{\mathbf{W}}^\top + \eta^2 I|$$

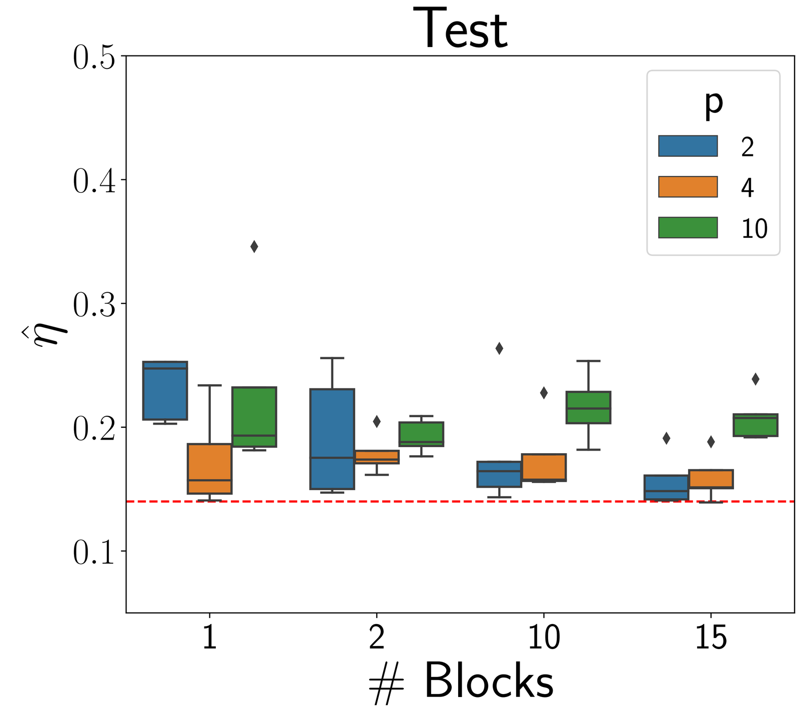
Experiments



Fading DNN vs Plain DNN: Box plots obtained from 20 independent runs with N=10k. The closer to the dashed line the better.



Block's relevance during optimization.



Effects of window's length: Box plots obtained from 20 independent runs with N=10k on System 4

Models	N = 400		N = 1000		N = 10k		N = 100k	
	Train	Test	Train	Test	Train	Test	Train	Test
GP model from Pillonetto et al.[2]	0.14	0.27	0.13	0.19	0.14	0.17	-	-
Our architecture w/o SO regularization	0.02	0.49	0.03	0.45	0.07	0.23	0.12	0.20
Our complete architecture	0.10	0.32	0.15	0.22	0.16	0.17	0.15	0.15

Comparison with SOTA: optimal innovation variance: $\eta^2 = 0.14$

Take-Home Message

1. Block-structured architecture
2. Regularization for automatic complexity selection
3. Over-parametrized vs non-parametric models: good in mid-large data regimes

SCAN ME

REFERENCES:

[1] Hao Li et al. Visualizing the loss landscape of neural nets. NeurIPS 2018

[2] Pillonetto et al. A New Kernel-Based Approach for Nonlinear System Identification, IEEE Transactions on Automatic Control, 2011