Online time-series forecasting with adaptive learning rates

Proposed by Ericsson

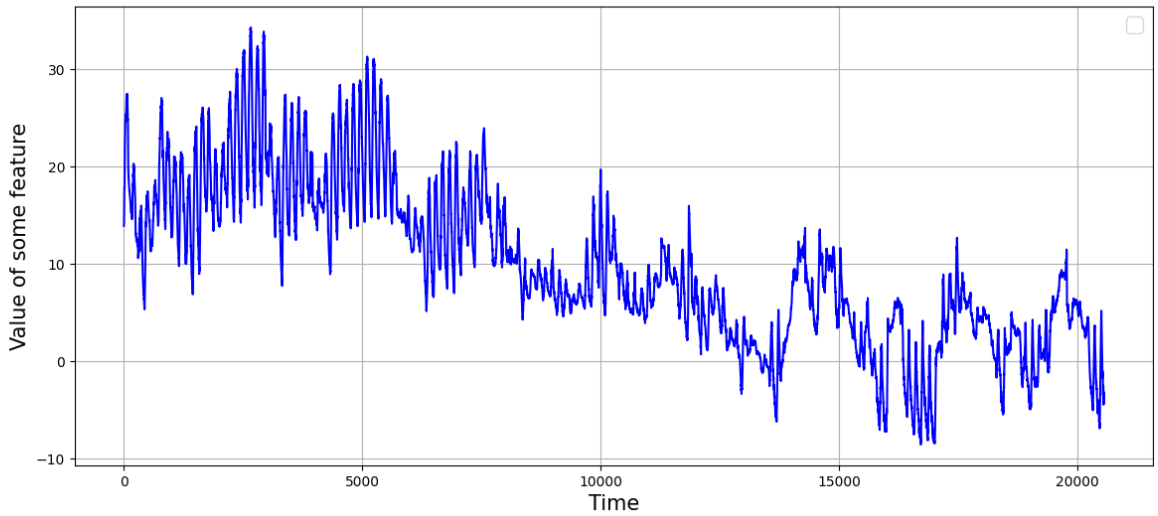
Team Ericsson

August 25, 2023

Contents

Introduction

Given time series data $\{t_i, x_i\}$, online learning methods predict values at future time steps given only the very last samples as input.



Our goal is to develop a general method to automate the selection of the learning rate.

Offline v. Online

- Offline learning: use a training dataset $D_0 = \{\mathrm{x}_i, \mathrm{y}_i\}_{1 \leq i \leq M}$ to train a model $f_\theta(\mathrm{x})$.
- To determine the hyperparameters θ, solve an optimization problem on the loss $\mathcal{L}$:

$$\theta^* = \arg\min_\theta \mathcal{L}(\theta)$$

using a gradient descent:

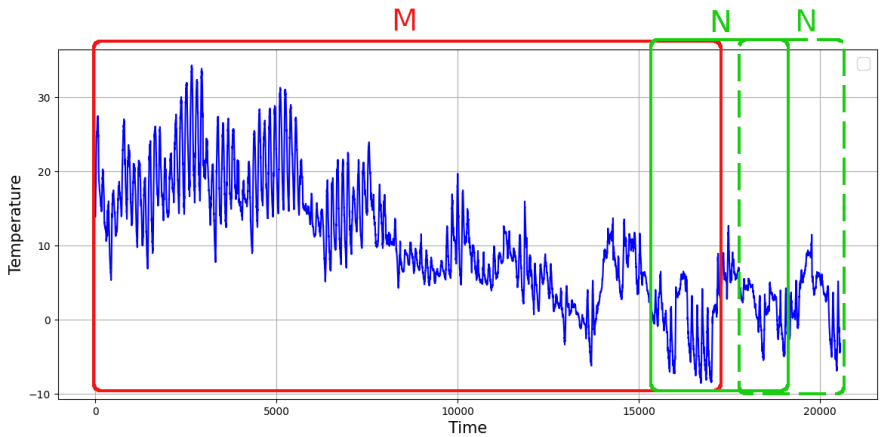
$$\theta_k = \theta_{k-1} - \alpha \nabla \mathcal{L}(\theta_{k-1})$$

where the learning rate α has to be tuned.

- How to choose α automatically while learning on-the-fly?

The specific problem statement

1. Train an initial model on a data set D_0 of M data,
2. Keep training our model on $N \ll M$ incoming data 'on-the-fly',
3. Use the updated model to predict either the following point or use P successive predictions to predict the P next points.



Questions raised:

- Model? Loss function?
- Updating procedure for α?

Multilayer Perceptron (MLP)

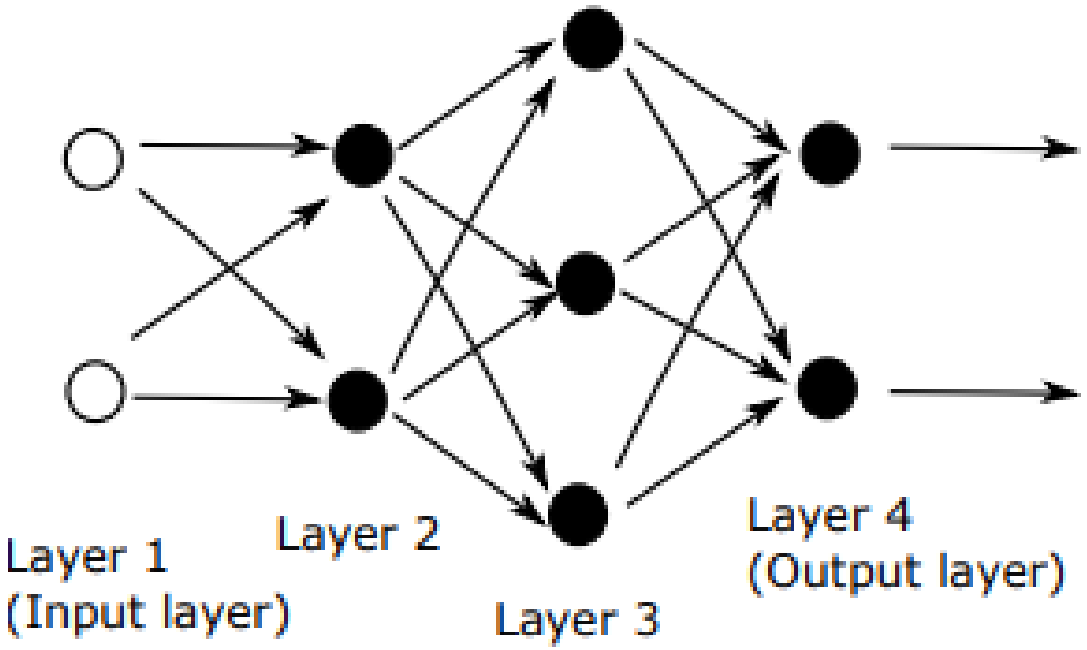


Figure: Example of a multilayer perceptron

Long Short Term Memory (LSTM)

- LSTM networks are a type of recurrent neural network (RNN) that help to carry over information over many timesteps [1][2].

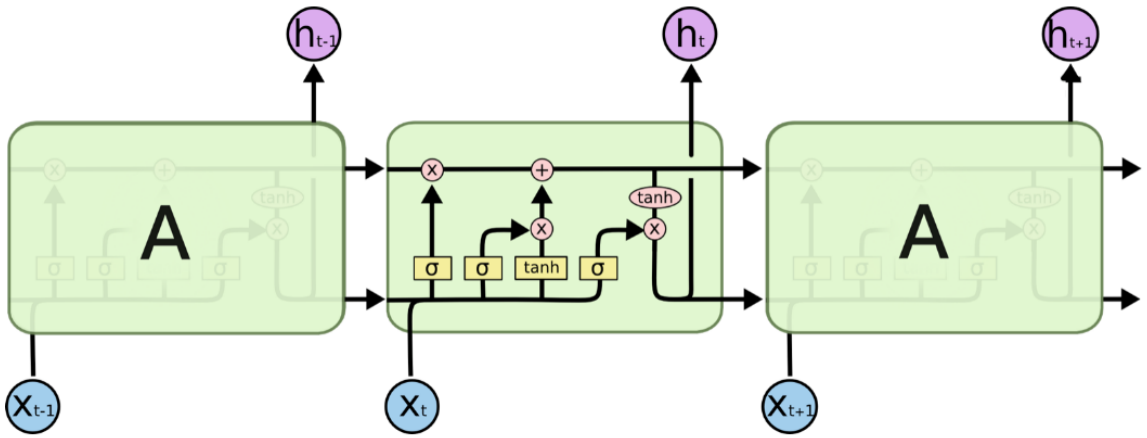


Figure: LSTM units arranged in series

The adaptive learning rate: Hypergradient Descent [3] and ADAM [4]

- Learning rate update

$$\alpha_k = \alpha_{k-1} + \beta \nabla \mathcal{L}(\theta_{k-1}) \nabla \mathcal{L}(\theta_{k-2})$$

- This update is a version of gradient descent for α as $\alpha_k = \alpha_{k-1} - \beta \frac{\partial \mathcal{L}(\theta_{k-1})}{\partial \alpha}$. Apply the chain rule to compute $\frac{\partial \mathcal{L}(\theta_{k-1})}{\partial \alpha} = \nabla \mathcal{L}(\theta_{k-1}) \frac{\partial \theta_{k-1}}{\partial \alpha}$.
- There is an analogous update rule for the ADAM optimizer in which learning parameter α is updated as above.

Jena Climate temperature data set

Temperatures at the Max Planck Institute from 2009 to 2016.

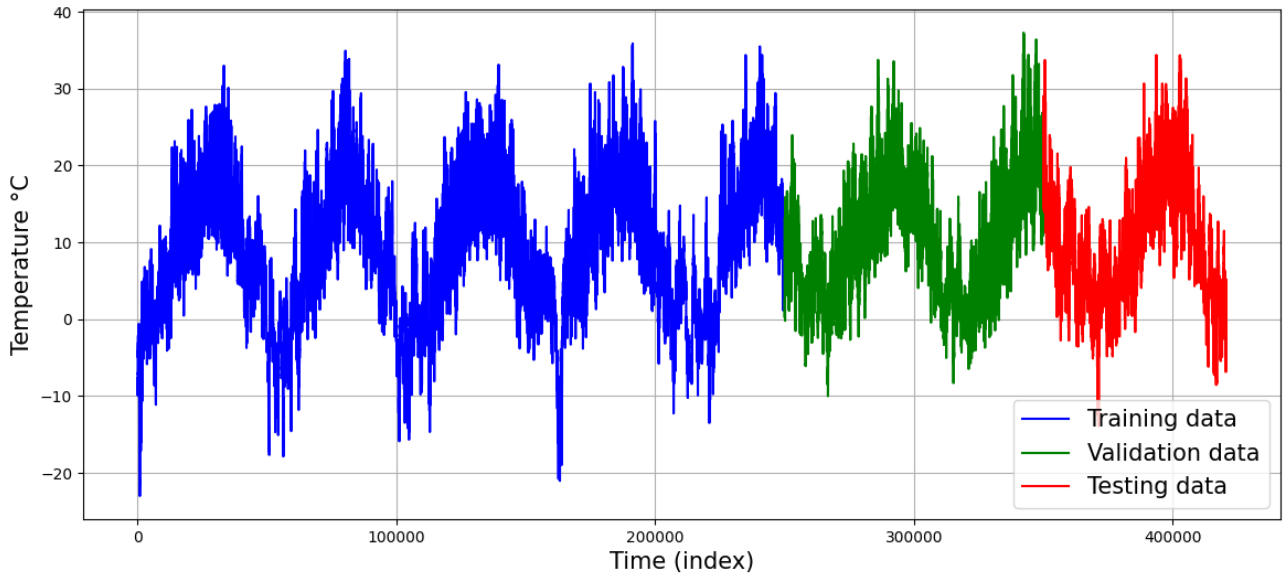


Figure: Splitting of the Jena Climate data set

We split this data set into a training set, a validation set to validate the decrease of the loss, and a test set for prediction.

One-step predictions on the Jena data set using a MLP

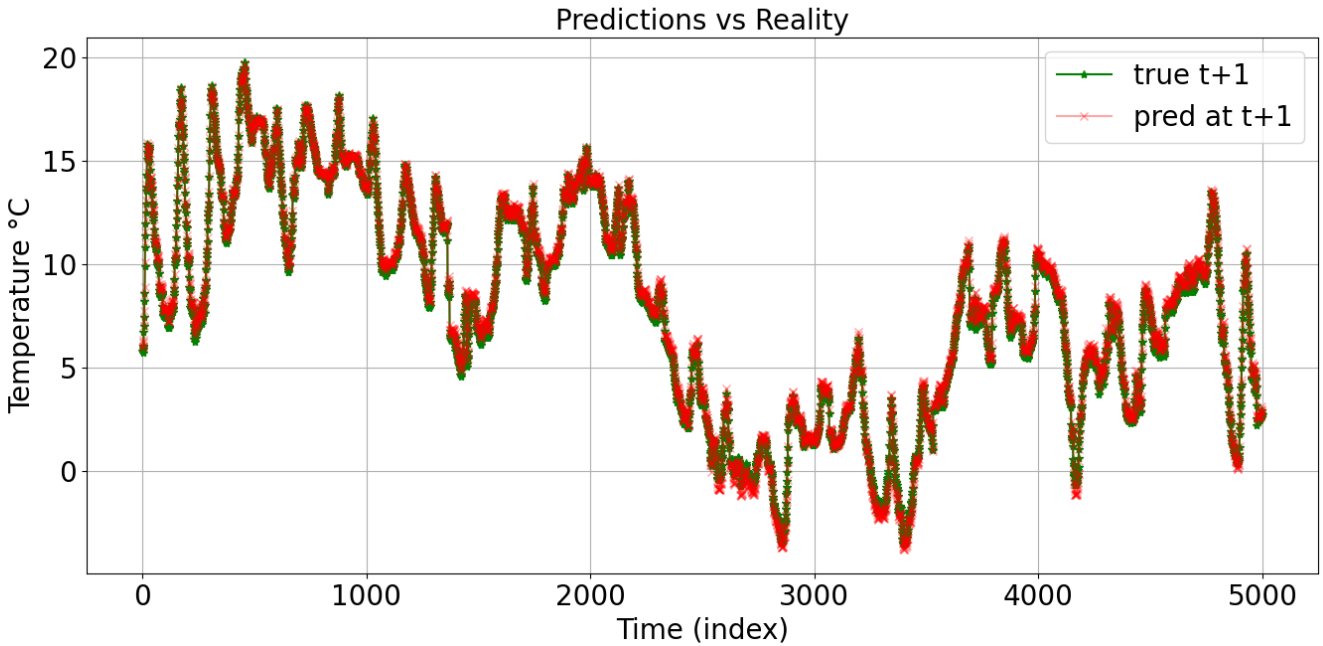


Figure: Mean residual of -0.78944.

Twenty-step predictions on the Jena data set using a MLP

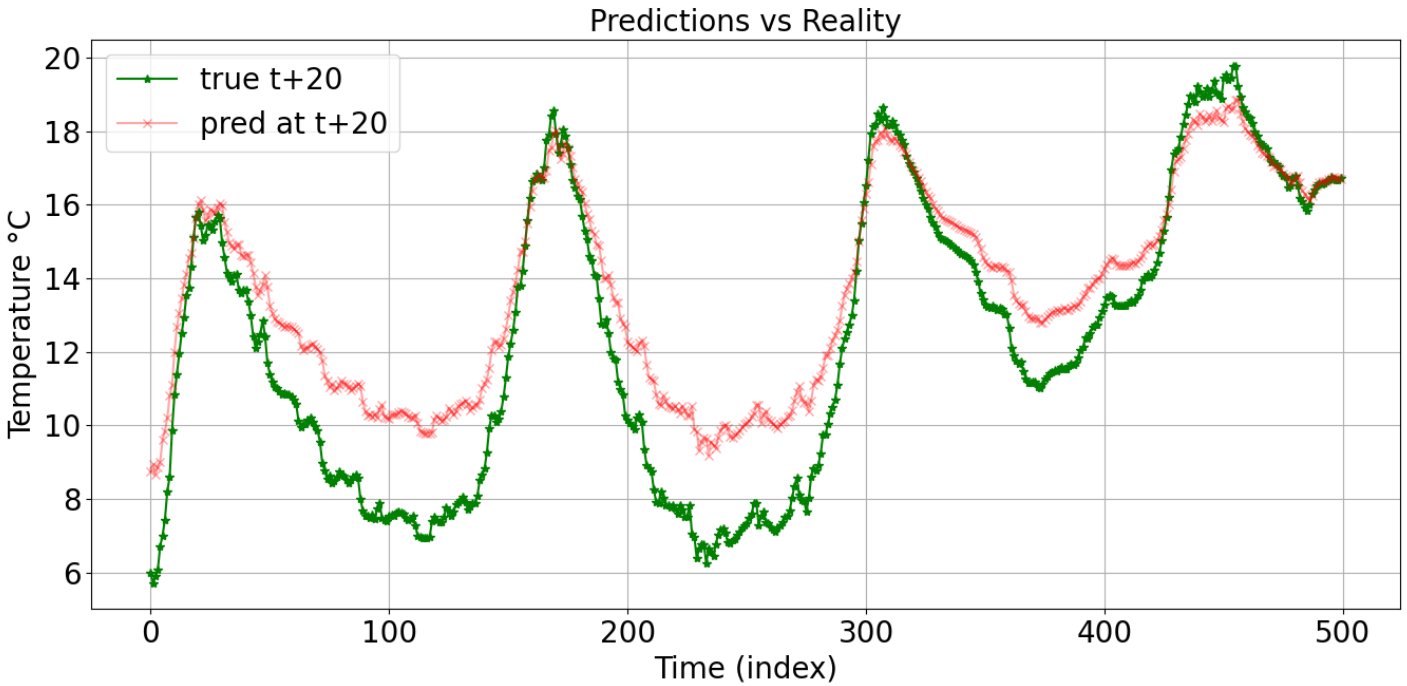


Figure: Predicting over 20 steps with mean residue -1.3150.

One-step predictions on the Jena data set using a LSTM

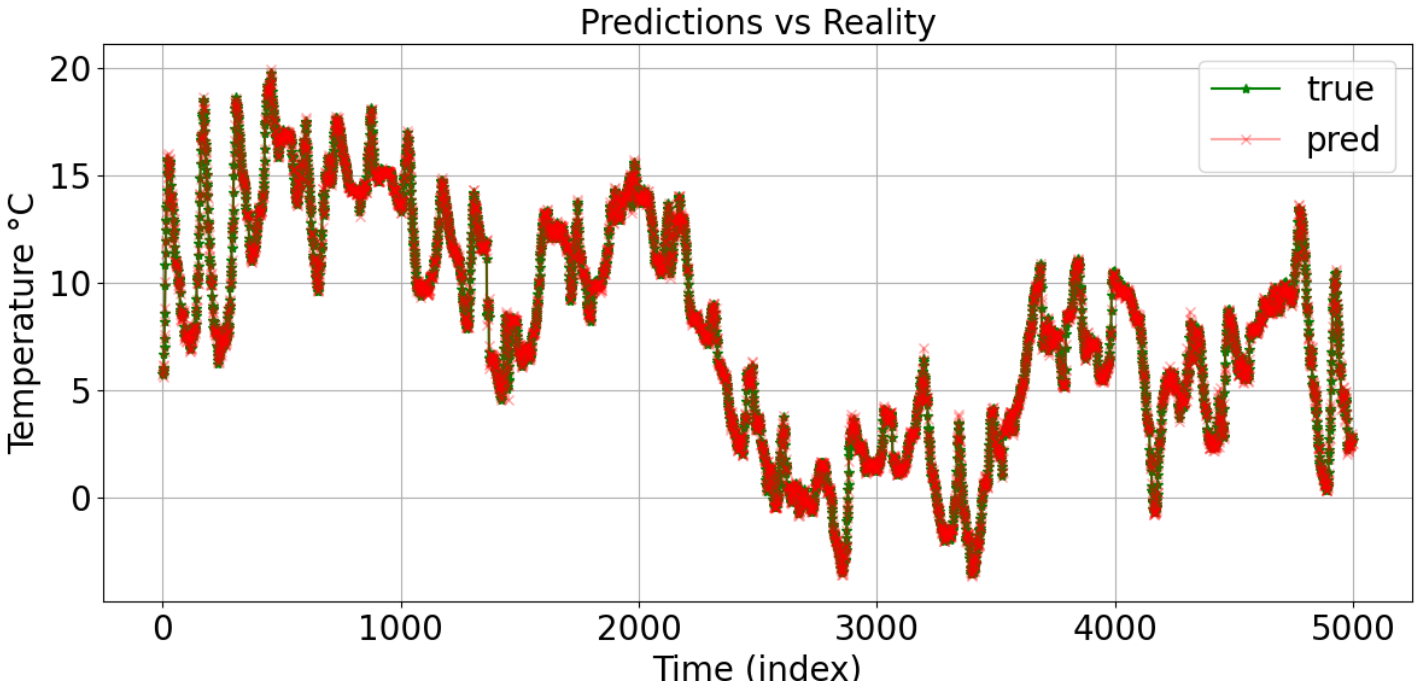


Figure: Mean residual of 0.012199.

Twenty-step predictions on the Jena data set using a LSTM

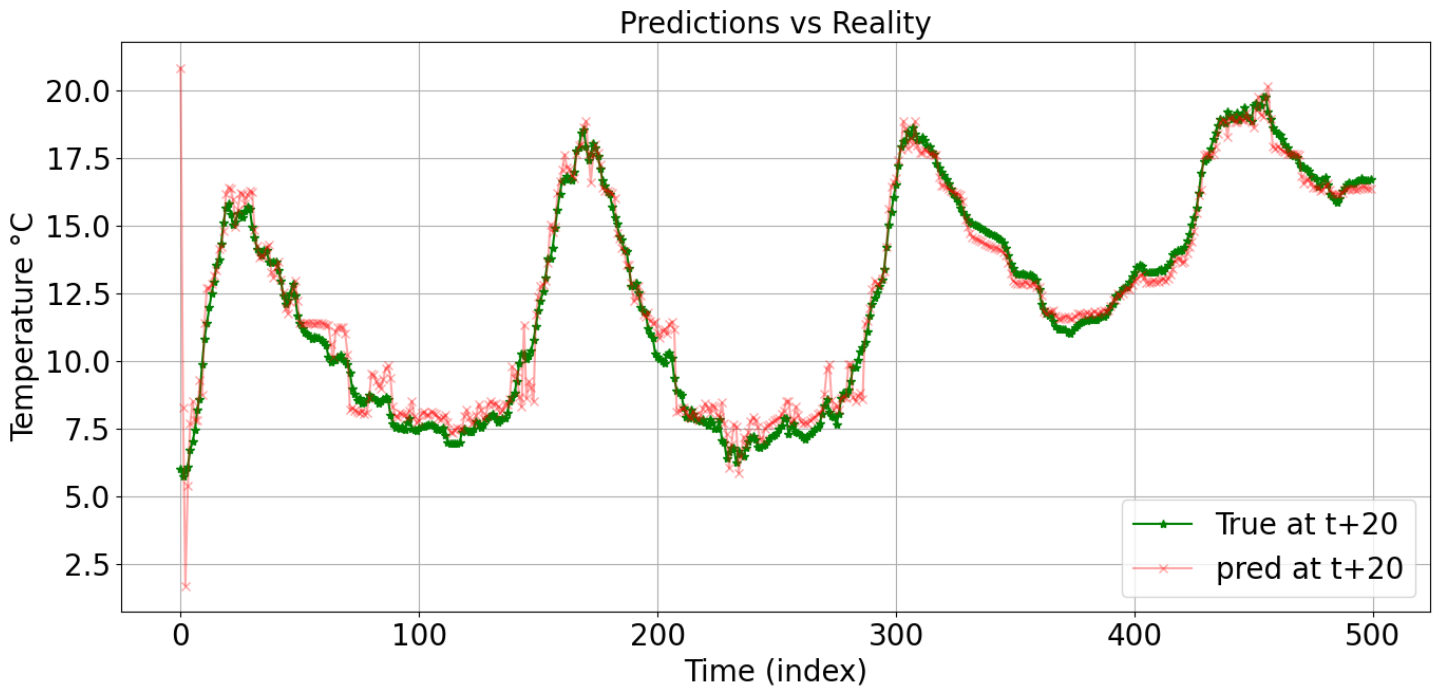
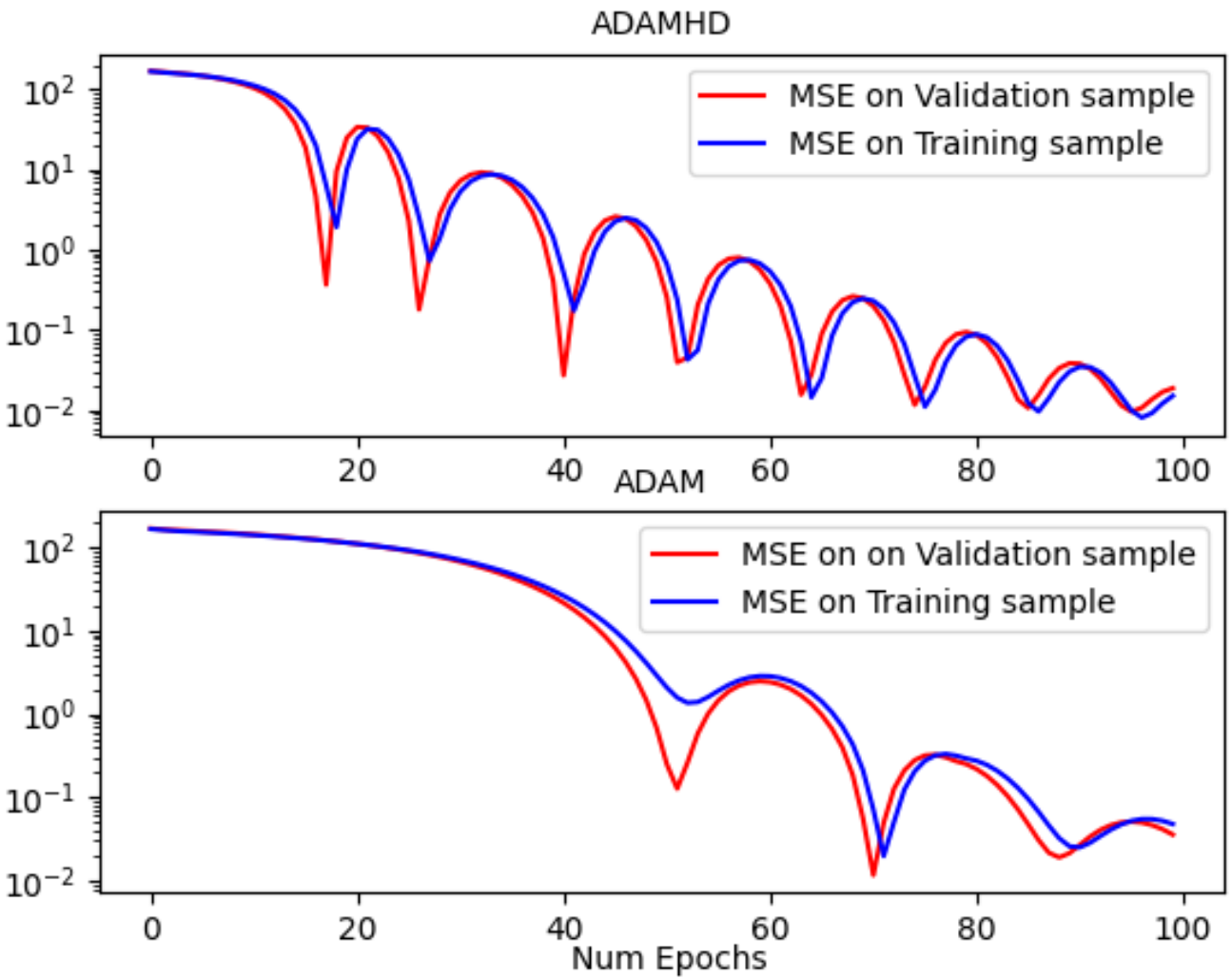


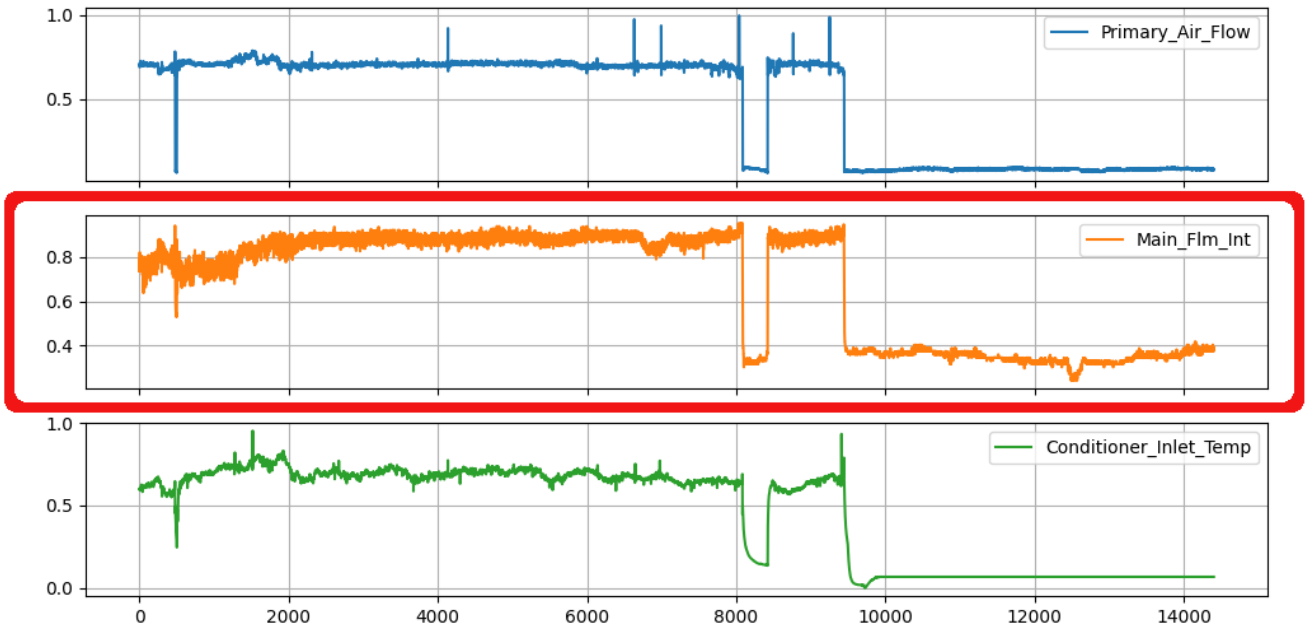
Figure: Predicting over 20 steps with mean residue -0.14160.

Comparison of losses with/without HD



A real-world data set

This data set is extracted from a coal-fired power plant, over 10 days with a measurement each minute.



The data is pre-normalized, among the 12 features available, we chose to focus on the main flame intensity.

One-step predictions on the coal burner data set: MLP

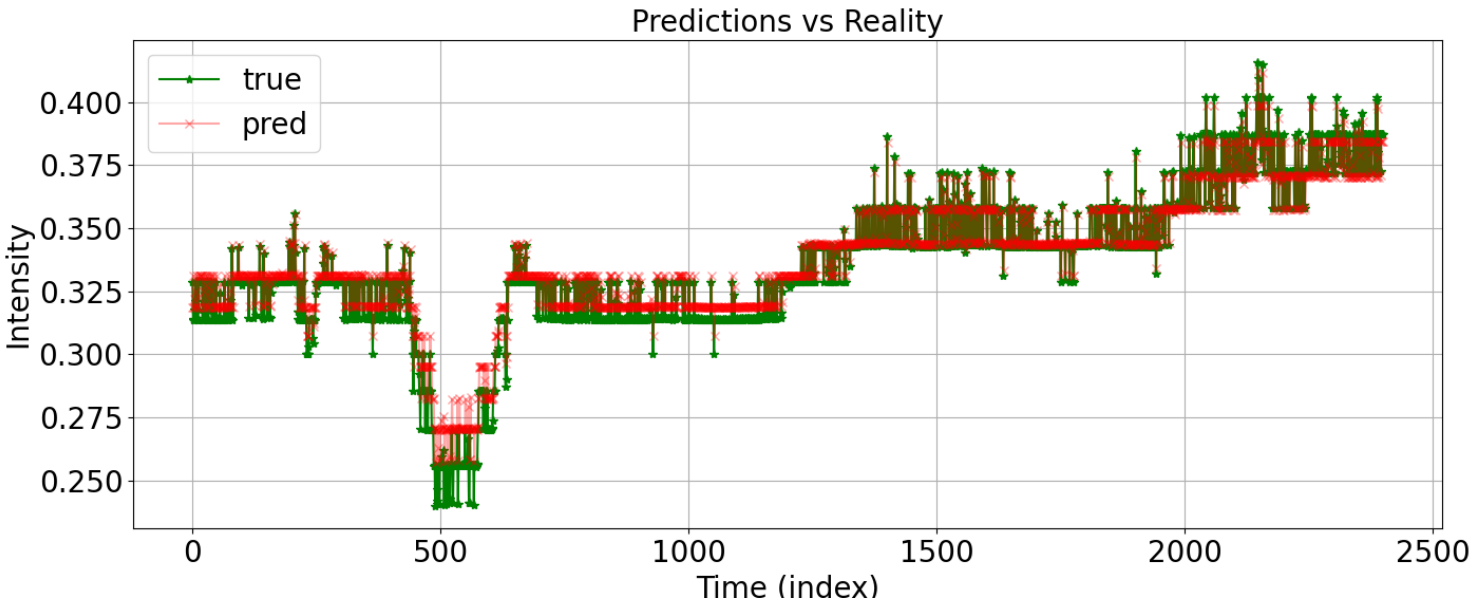


Figure: The one step predictions using a MLP

One-step predictions on the coal burner data set: LSTM

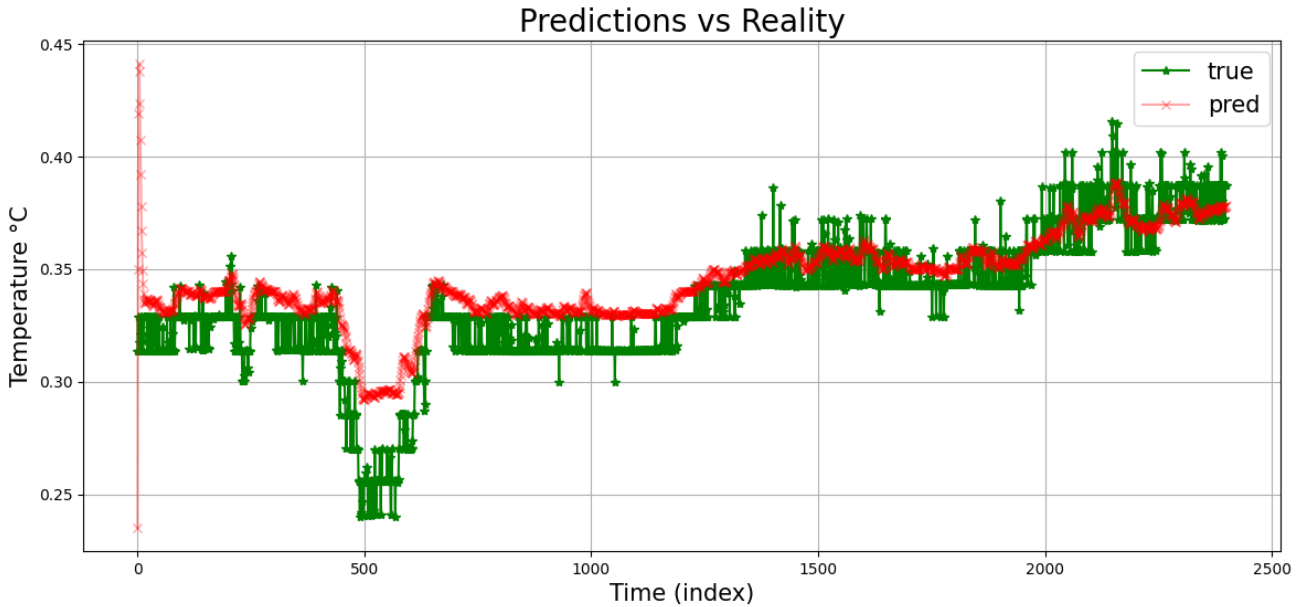
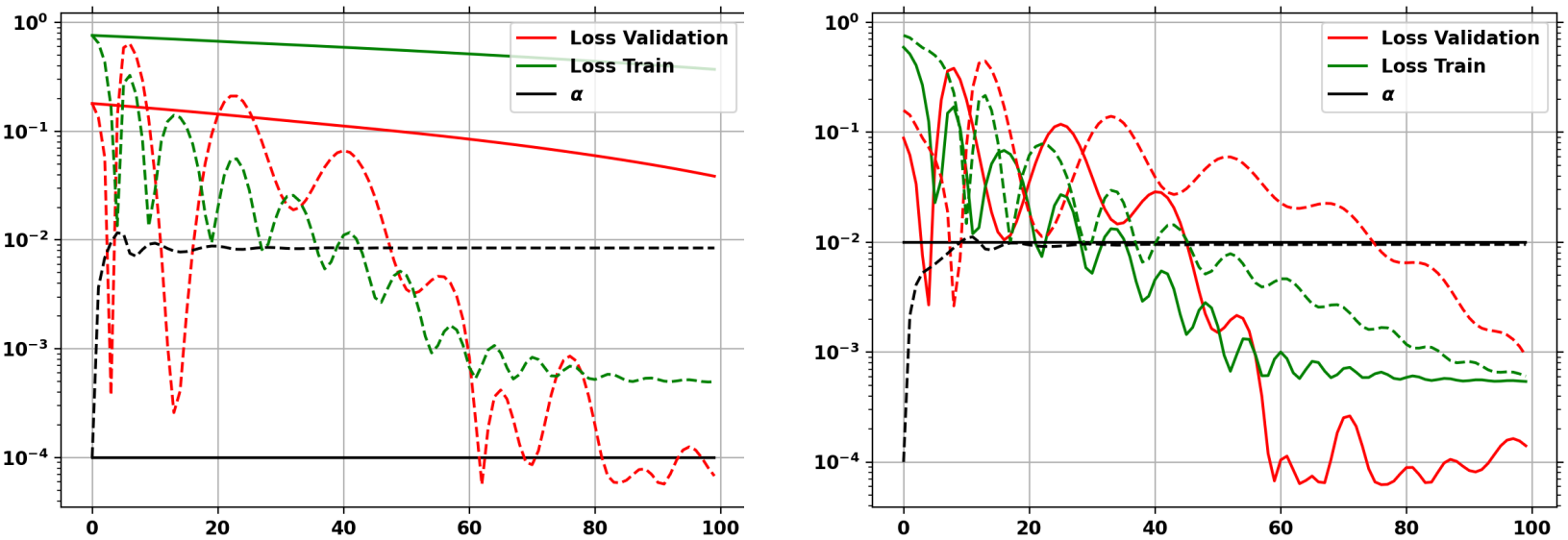


Figure: The one step predictions using a LSTM

Comparison of losses with/without HD



Dotted lines: with HD. Solid lines: without HD.

Summary and directions for future work

- We implemented a automatized method for modifying the learning rate on-the-fly based on HyperGradient descent.
- Our time-series LSTM could be improved so as to deal with the erratic behaviour of the coal burner data set.

Thank you!

References

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, pp. 1735–80, Dec. 1997.

[2] B. Lindemann, T. Müller, H. Vietz, N. Jazdi, and M. Weyrich, "A survey on long short-term memory networks for time series prediction," *Procedia CIRP*, vol. 99, pp. 650–655, 2021, ISSN: 2212-8271.

[3] A. G. Baydin, R. Cornish, D. M. Rubio, M. Schmidt, and F. Wood, "Online learning rate adaptation with hypergradient descent," in *International Conference on Learning Representations*, 2018.

[4] D. P. Kingma and J. Ba, *Adam: A method for stochastic optimization*, 2017. arXiv: 1412.6980 [cs.LG].