Institute for
Infocomm Research
I²R

POLA: ONLINE TIME SERIES PREDICTION BY ADAPTIVE LEARNING RATES

Wenyu Zhang

Machine Intellection Department

Institute for Infocomm Research

IEEE ICASSP 2021, 6-11 June

Outline

CREATING GROWTH, ENHANCING LIVES

Motivation

Streaming setting:

- Data is collected by continuously monitoring a system
- Data in dynamic environments is subject to concept drift
  - Joint distribution of predictor and response variables changes across time
- Models need to be updated to avoid degrading performance

Motivation

Online time series prediction

- Temporal correlations means that observations cannot assumed to be independently and identically distributed (i.i.d.)
- Focus on deep recurrent neural networks

Motivation

Online time series prediction

- Temporal correlations means that observations cannot assumed to be independently and identically distributed (i.i.d.)
- Focus on deep recurrent neural networks

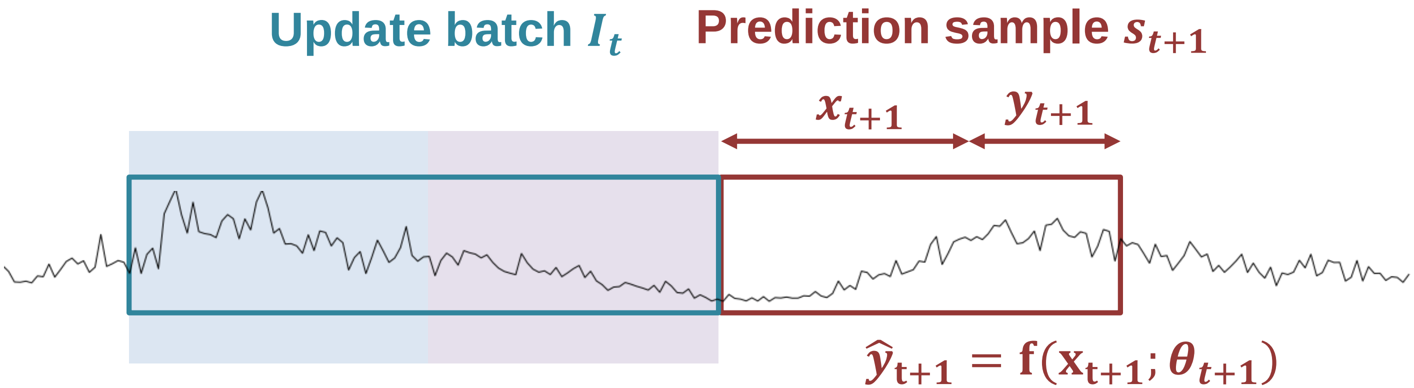
Goal:

- Adapt quickly in dynamic environments without overfitting to current system state or noisy samples, by
- Automatically scheduling the online learning rate of stochastic gradient descent (SGD) algorithm

CREATING GROWTH, ENHANCING LIVES

Problem Setup

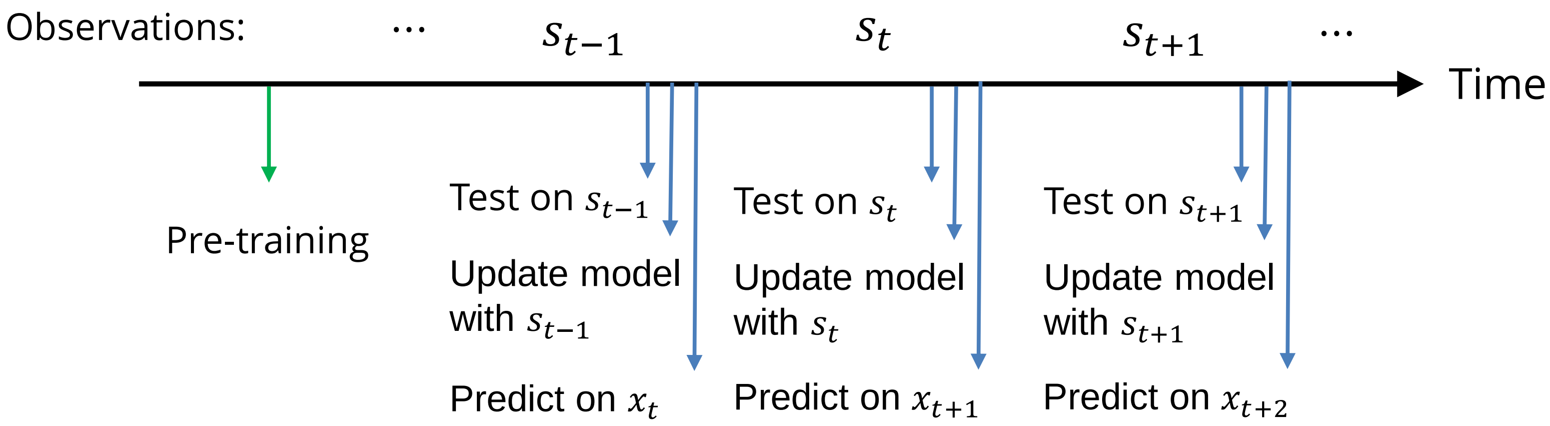
- For process $\{Z_t\}$, observation $z_t \in \mathbb{R}^d$
- At time t, use a historical sequence x_t of length m, to predict an output forecast sequence y_t for the next n time steps
  - $x_t = [z_{t-m+1}, \dots, z_t]$
  - $y_t = [z_{t+1}, \dots, z_{t+n}]$
  - Denote prediction sample $s_t = (x_t, y_t)$
- Online batch size b



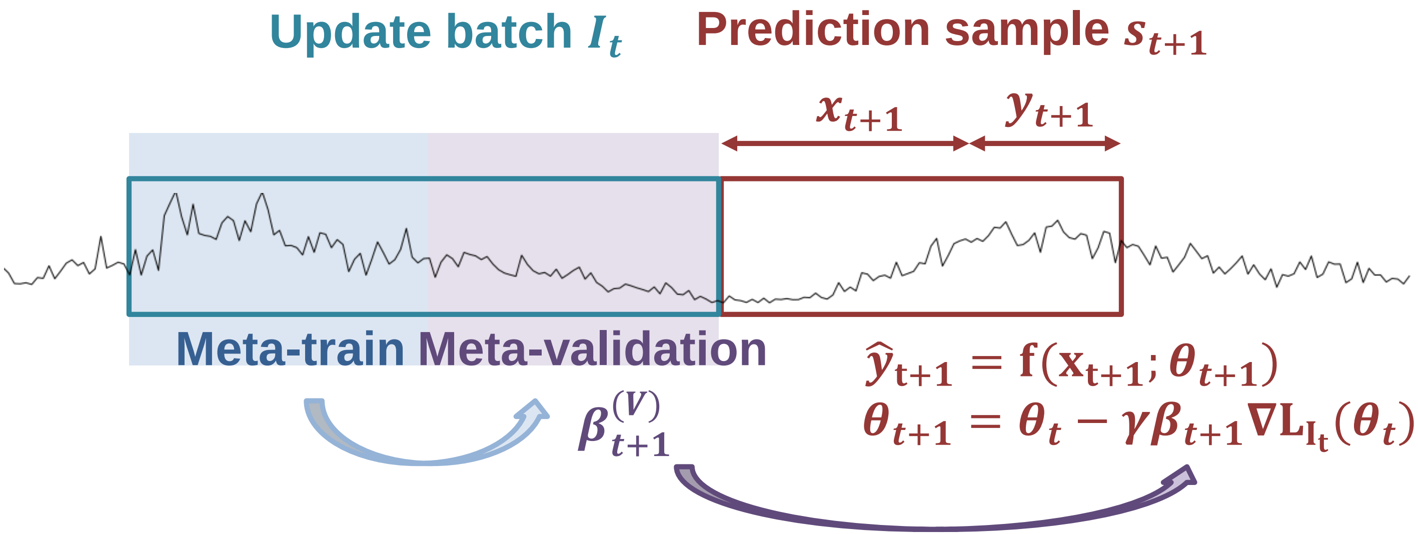
Problem Setup

Example interleaved-test-then-train scheme:

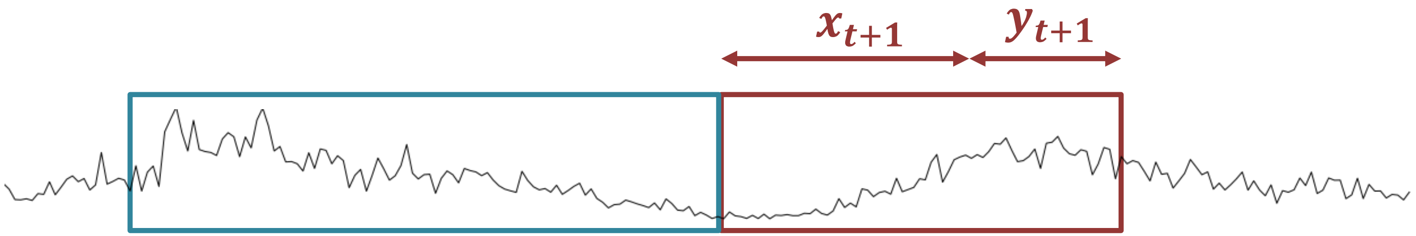
- Forecast length $n = 1$
- Online batch size $b = 1$



Proposed Method: Predicting Online by Learning rate Adaptation (POLA)



Proposed Method: POLA

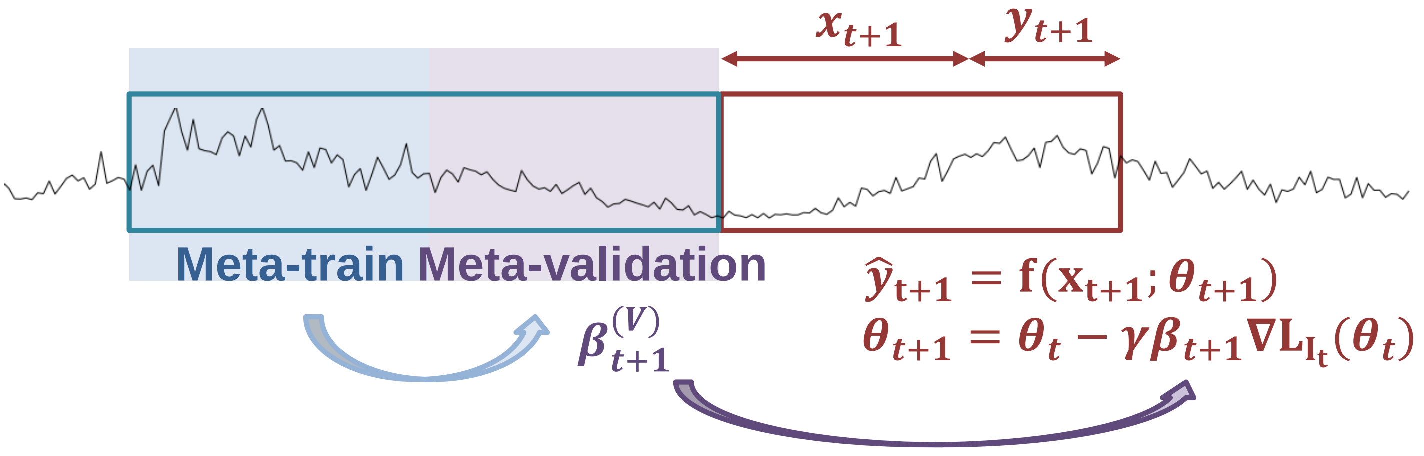


$$\hat{y}_{t+1} = \mathbf{f}(\mathbf{x}_{t+1}; \boldsymbol{\theta}_{t+1})$$
$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma \boldsymbol{\beta}_{t+1} \nabla \mathbf{L}_{\mathbf{I}_t}(\boldsymbol{\theta}_t)$$

Adaptive learning rate

- Maximum learning rate γ
- Learning rate factor $\beta_{t+1} \in [0,1]$
- Learning rate is small if the current update batch is not useful in helping the model adapt

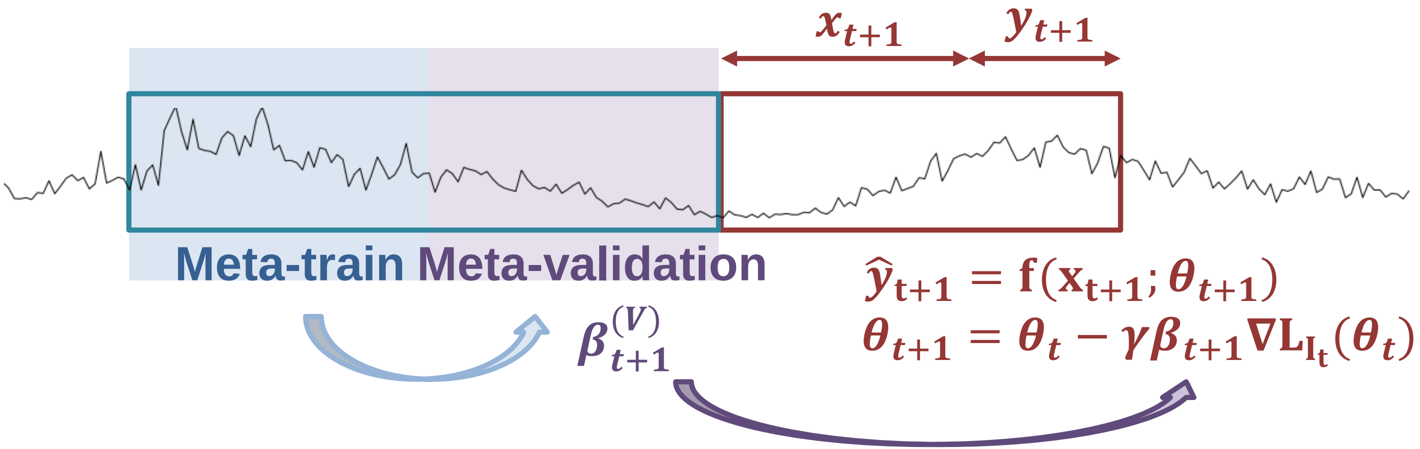
Proposed Method: POLA



Meta-learn the learning rate factor

- Split the update batch to a meta-training and meta-validation set
- Meta-learning sets are a proxy to the training and testing procedure
- Optimize the learning rate factor on the meta-learning sets

Proposed Method: POLA



Implementation

(1) POLA-FS: Search for $\beta_{t+1}^{(V)}$ in a finite set of candidates

(2) POLA-GD: Optimizes for $\beta_{t+1}^{(V)}$ by gradient descent with learning rate η for k steps

Experiments: Datasets

Sunspot

- Monthly sunspot number from January 1749 to July 2020
- Historical data length $m = 48$
- Forecast length $n = 5$

Household Power Consumption

- Daily power consumption (global active power, global intensity, voltage) from December 16, 2006 to November 26, 2010
- Historical data length $m = 28$
- Forecast length $n = 3$

CREATING GROWTH, ENHANCING LIVES

Experiments: Competing Methods

Holt-Winters: exponential smoothing
OR-ELM: online recurrent extreme learning machine

Recurrent neural network

- Pre-trained: no model update in online phase
- FTL: Follow-The-Leader retraining every b steps
- MAML: meta-learning pre-training
- Online-SGD: constant learning rate
- Online-RMSprop: element-wise adaptive learning rate
- WG: adapts SGD learning rate based on whether current sample is outlier or change point

CREATING GROWTH, ENHANCING LIVES

Experiments: Online Prediction Performance

- RNN and time series models

METHOD	NORMALIZED RMSE	
	Sunspot	Power
Holt-Winters	0.991	NA
OR-ELM	0.822	NA
Pre-trained	0.572	0.816
FTL*	0.572	0.820
MAML	1.295	1.023
Online-SGD	0.552	0.775
Online-RMSprop	0.536	0.809
WG	0.552	NA
POLA-FS	<u>0.532</u> $\pm$ 0.002	**0.769** $\pm$ 0.003
POLA-GD	**0.500** $\pm$ 0.002	<u>0.773</u> $\pm$ 0.005

CREATING GROWTH, ENHANCING LIVES

Experiments: Online Prediction Performance

- LSTM and GRU

MODEL	METHOD	NORMALIZED RMSE	
		Sunspot	Power
LSTM	Online-SGD	0.532	0.821
	Online-RMSprop	0.517	**0.794**
	WG	0.532	NA
	POLA-FS	0.534 ± 0.009	0.802 ± 0.005
	POLA-GD	$\mathbf{0.512} \pm 0.006$	0.806 ± 0.071
GRU	Online-SGD	0.526	**0.768**
	Online-RMSprop	0.521	0.786
	WG	0.526	NA
	POLA-FS	0.508 ± 0.002	0.769 ± 0.003
	POLA-GD	$\mathbf{0.489} \pm 0.002$	$\mathbf{0.768} \pm 0.003$

Experiments: Sensitivity Analysis

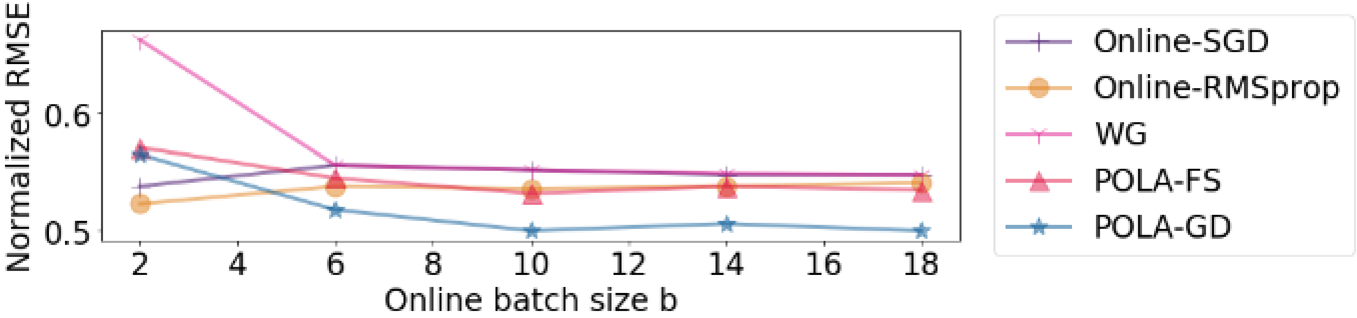
- POLA-GD gradient descent hyperparameters

# STEPS (k)	LEARNING RATE (η)	NORMALIZED RMSE Sunspot	NORMALIZED RMSE Power
1	0.1	0.515	0.773
2	0.1	0.504	0.772
3	0.1	0.500	0.773
1	0.01	0.525	0.777
2	0.01	0.520	0.776
3	0.01	0.516	0.775

CREATING GROWTH, ENHANCING LIVES

Experiments: Sensitivity Analysis

- Online batch size



Summary

Proposed POLA method

- Automatically schedules SGD online learning rate
- Adapts online learning rate by assimilating training and testing procedure with meta-learning

☑ Model-agnostic

☑ Attains overall comparable or better predictive performance over competing methods across multiple datasets and network architectures

CREATING GROWTH, ENHANCING LIVES

Selected References

T. Guo, Z. Xu, X. Yao, H. Chen, K. Aberer and K. Funaya, “Robust Online Time Series Prediction with Recurrent Neural Networks,” *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*.

J. Park and J. Kim, “Online Recurrent Extreme Learning Machine and Its Application to Time-Series Prediction,” *2017 International Joint Conference on Neural Networks (IJCNN)*.

Anusha Nagabandi, I. Clavera, Simin Liu, Ronald S. Fearing, P. Abbeel, S. Levine and Chelsea Finn, “Learning to Adapt in Dynamic, Real-World Environments through Meta-Reinforcement Learning,” *2019 arXiv*.

Geoffrey Hinton, Nitish Srivastava and Kevin Swersky, “Neural Network for Machine Learning.”

CREATING GROWTH, ENHANCING LIVES

Institute for Infocomm Research

I²R

THANK YOU

www.a-star.edu.sg

CREATING GROWTH, ENHANCING LIVES