NeurIPS 2019 Vancouver
33rd Conference on Neural Information Processing Systems

Shape and Time Distortion Loss for Training Deep Time Series Forecasting Models

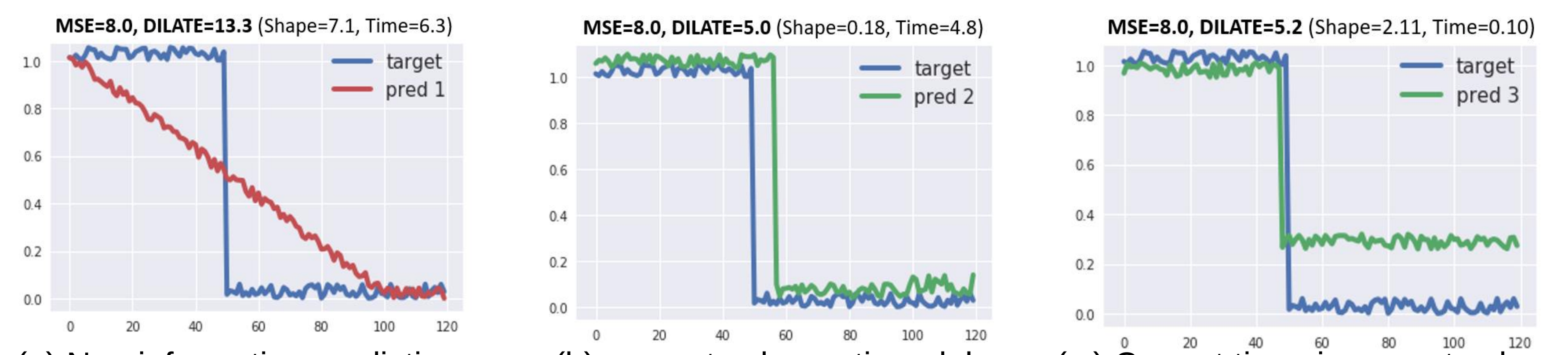
Vincent Le Guen[1,2], Nicolas Thome[2]
[1] EDF R&D, Chatou, France
[2] CEDRIC, Conservatoire National des Arts et Métiers, Paris, France

Context

Multi-step and **non stationary** time series forecasting (with sudden changes)
Important in many contexts, *eg* anticipate future drops of electricity production

Task-dependent metrics for evaluating forecasts (*e.g.* Time Distortion Index [1], ramp score, Hausdorff) often **non differentiable**

⇒ **Mean Squared Error (MSE) as a surrogate training loss** for most state-of-the-art models [2,3,4]
MSE however ill-adapted to favour interesting *vs* naïve forecasts…

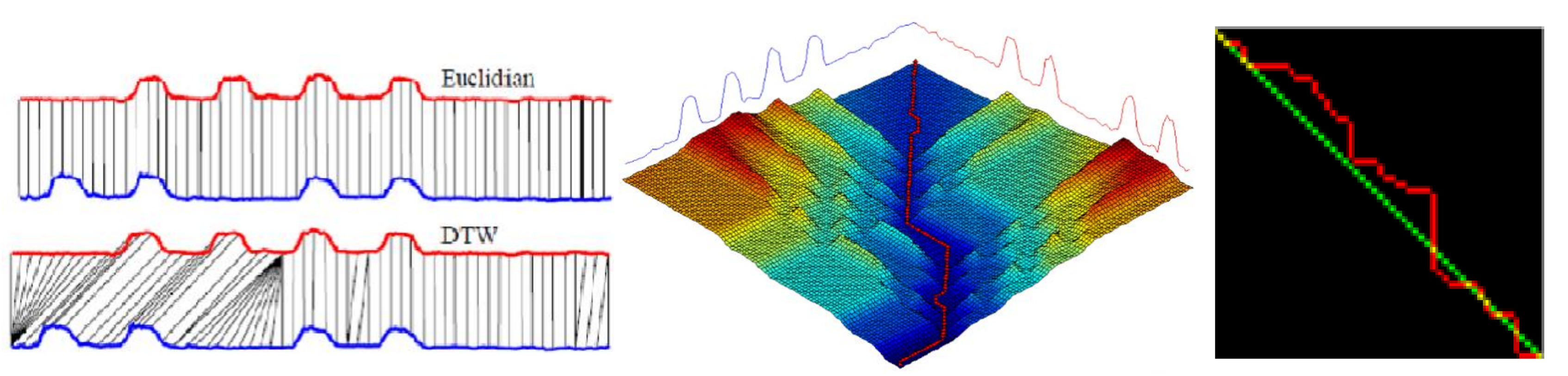


(a) Non informative prediction (b) accurate shape, time delay (c) Correct time, inaccurate shape

<u>**Motivation:**</u> **differentiable** loss function for training **deep models** for **precise shape** and **temporal change** detection

Shape Loss

Based on **Dynamic Time Warping (DTW)** that computes optimal alignment $\mathbf{A}^*$ between time series:



MSE *vs* DTW loss — Pairwise cost matrix Δ — Optimal alignment $\mathbf{A}^*$

We use the soft-DTW [6] with $\min_\gamma(a_1,...,a_n) = -\gamma \log(\sum_{i=1}^{n} \exp(-a_i/\gamma))$

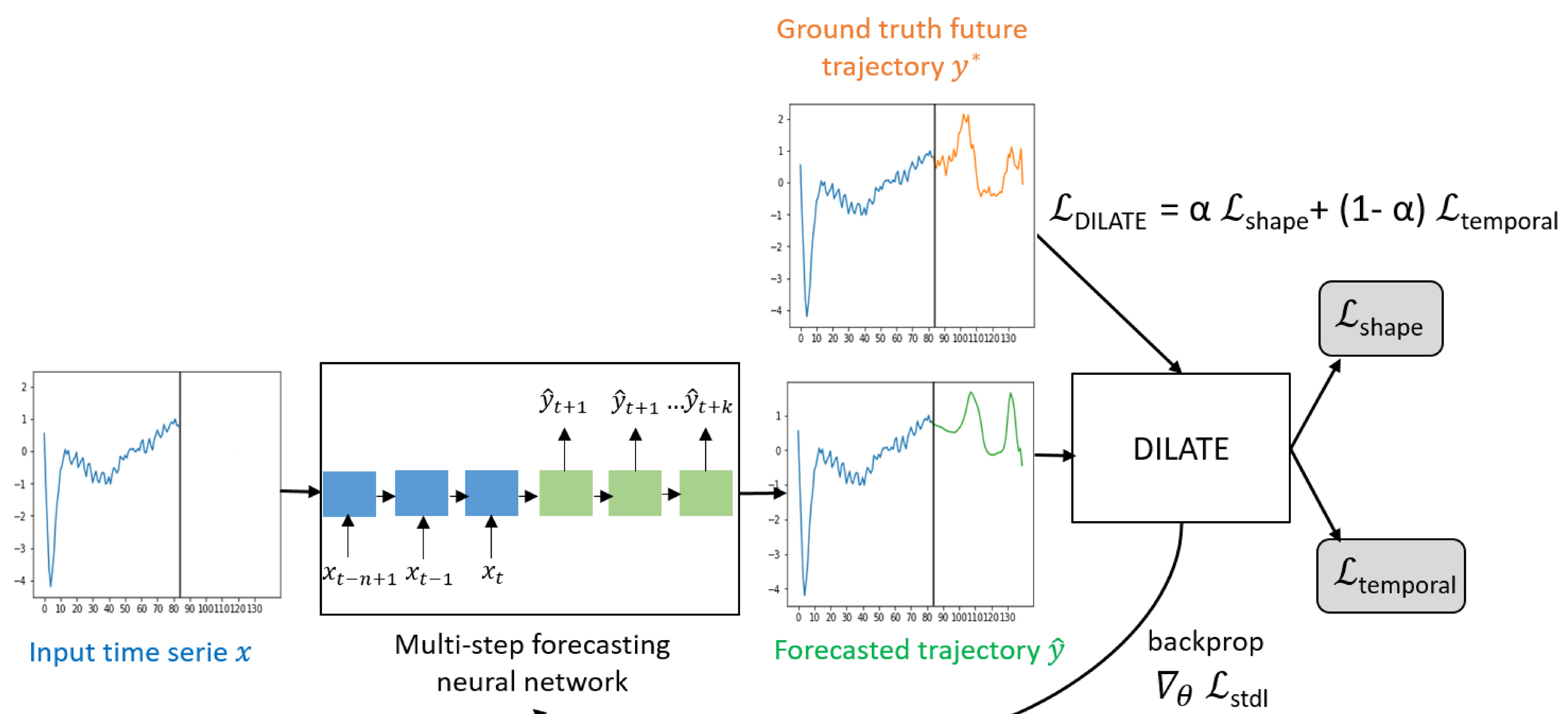
to get our **differentiable shape loss**: $\mathbf{A}^* = \underset{\mathbf{A}\in\mathcal{A}_{k,k}}{\arg\min} \left\langle \mathbf{A}, \mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i)\right\rangle$

$$\mathcal{L}_{shape}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = DTW_\gamma(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i)$$

DILATE (DIstortion Loss with shApe and TimE)

Multi-step time series forecasting: predict the future k-steps trajectory $\hat{\mathbf{y}}_i = (\hat{\mathbf{y}}_i^1, ..., \hat{\mathbf{y}}_i^k) \in \mathbb{R}^{d\times k}$ given input sequence $\mathbf{x}_i = (\mathbf{x}_i^1, ..., \mathbf{x}_i^n) \in \mathbb{R}^{p\times n}$

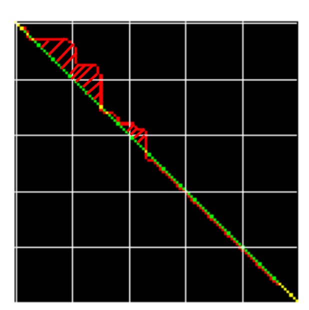
$$\mathcal{L}_{DILATE}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = \alpha\, \mathcal{L}_{shape}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) + (1-\alpha)\, \mathcal{L}_{temporal}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i)$$



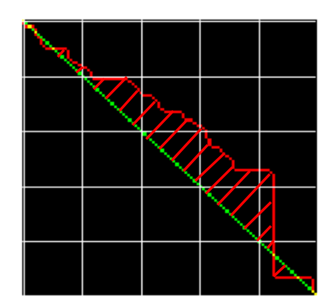
Temporal Loss

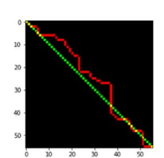
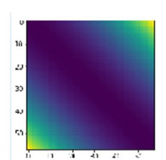
Quantify the deviation of optimal path $\mathbf{A}^*$ from the main diagonal with the Time Distortion Index (TDI) [1]:

Small red area:
⇒ small temporal distortion



Large red area:
⇒ large temporal distortion



$$TDI(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = \langle \mathbf{A}^*, \mathbf{\Omega}\rangle = \left\langle \underset{\mathbf{A}\in\mathcal{A}_{k,k}}{\arg\min} \left\langle \mathbf{A}, \mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i)\right\rangle, \mathbf{\Omega} \right\rangle = \Big\langle$$  ,  $$\Big\rangle$$

<u>**Challenge:**</u> **differentiating TDI:** replace $\mathbf{A}^*$by its smooth approximation:

$$\mathbf{A}^*_\gamma = \nabla_\Delta DTW_\gamma(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) = 1/Z \sum_{\mathbf{A}\in\mathcal{A}_{k,k}} \mathbf{A} \exp^{-\frac{\langle \mathbf{A}, \mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i)\rangle}{\gamma}}$$

$$\mathcal{L}_{temporal}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i) := \langle \mathbf{A}^*_\gamma, \mathbf{\Omega}\rangle = \frac{1}{Z} \sum_{\mathbf{A}\in\mathcal{A}_{k,k}} \langle \mathbf{A}, \mathbf{\Omega}\rangle \exp^{-\frac{\langle \mathbf{A}, \mathbf{\Delta}(\hat{\mathbf{y}}_i, \overset{*}{\mathbf{y}}_i)\rangle}{\gamma}}$$

Efficient implementation

Direct computation of $\mathcal{L}_{shape}$ and $\mathcal{L}_{temporal}$ **intractable** ($|\mathcal{A}_{k,k}| = O(exp(k^2))$)

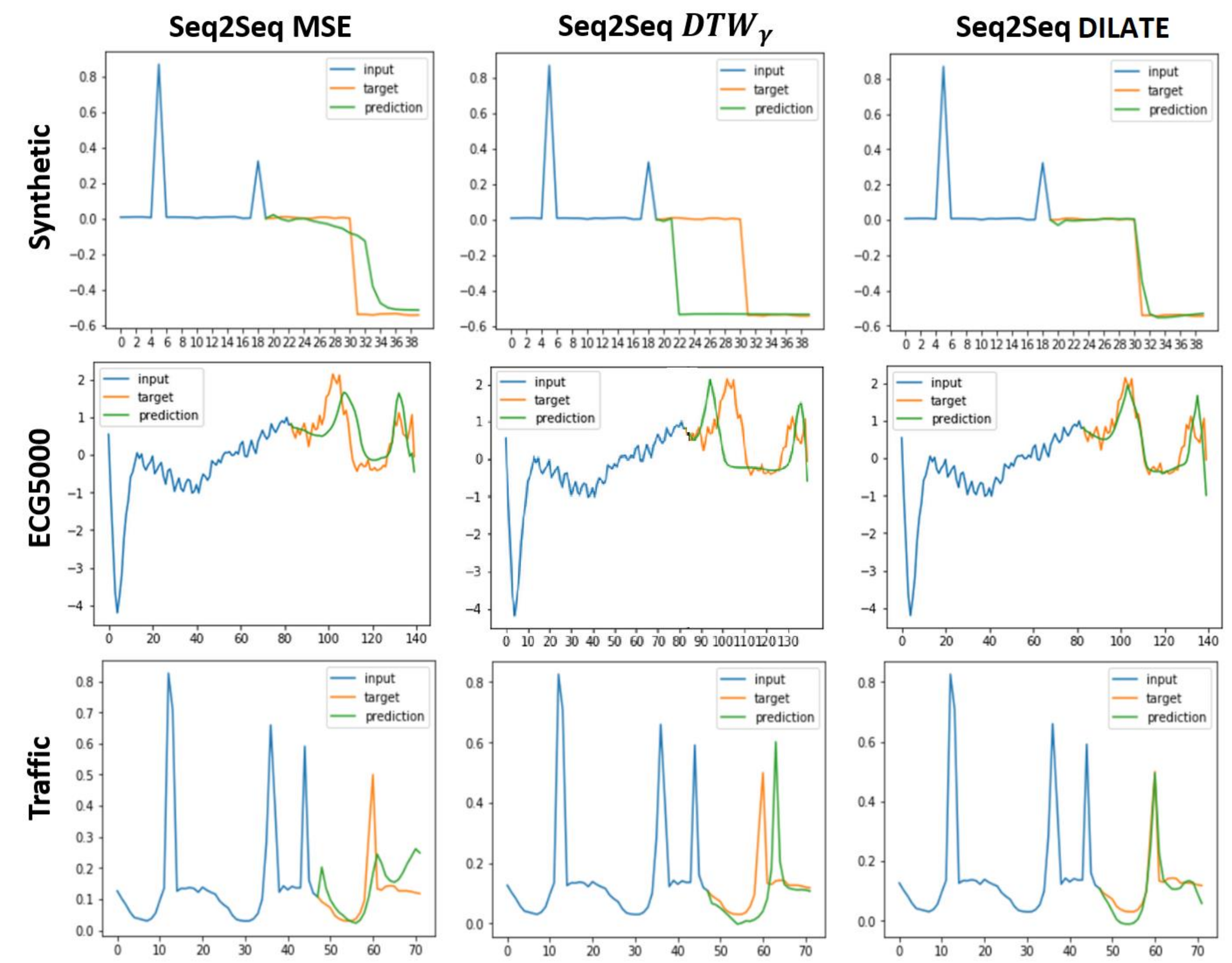
⇒ <u>**Solution:**</u> dynamic programming with custom forward/backward implementation (Pytorch)

Experiments

3 various datasets: Synthetic, ECG 5000, Traffic
Evaluate *k*-steps future trajectories (k=20 for Synthetic, 57 for ECG, 24 for Traffic)

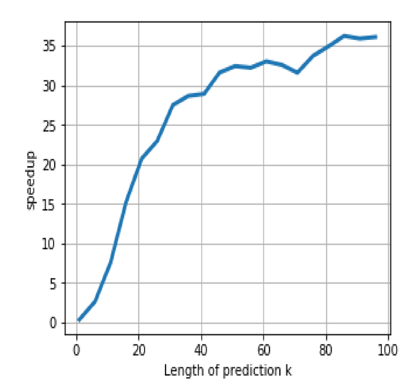
		Fully connected network (MLP)			Recurrent neural network (Seq2Seq)		
Dataset	Eval	MSE	DTW_γ [13]	DILATE (ours)	MSE	DTW_γ [13]	DILATE (ours)
Synth	MSE	**1.65 ± 0.14**	4.82 ± 0.40	**1.67± 0.184**	**1.10 ± 0.17**	2.31 ± 0.45	**1.21 ± 0.13**
	DTW	38.6 ± 1.28	**27.3 ± 1.37**	32.1 ± 5.33	**24.6 ± 1.20**	**22.7 ± 3.55**	**23.1 ± 2.44**
	TDI	15.3 ± 1.39	26.9 ± 4.16	**13.8 ± 0.712**	17.2 ± 1.22	20.0 ± 3.72	**14.8 ± 1.29**
ECG	MSE	**31.5 ± 1.39**	70.9 ± 37.2	37.2 ± 3.59	**21.2 ± 2.24**	75.1 ± 6.30	30.3 ± 4.10
	DTW	19.5 ± 0.159	18.4 ± 0.749	**17.7 ± 0.427**	17.8 ± 1.62	17.1 ± 0.650	**16.1 ± 0.156**
	TDI	**7.58 ± 0.192**	38.9 ± 8.76	**7.21 ± 0.886**	8.27 ± 1.03)	27.2 ± 11.1	**6.59 ± 0.786**
Traffic	MSE	**0.620 ± 0.010**	2.52 ± 0.230	1.93 ± 0.080	**0.890 ± 0.11**	2.22 ± 0.26	**1.00 ± 0.260**
	DTW	24.6 ± 0.180	**23.4 ± 5.40**	**23.1 ± 0.41**	24.6 ± 1.85	**22.6 ± 1.34**	**23.0 ± 1.62**
	TDI	**16.8 ± 0.799**	27.4 ± 5.01	**16.7 ± 0.508**	**15.4 ± 2.25**	22.3 ± 3.66	**14.4± 1.58**

⇒ DILATE loss **better when evaluated on shape (DTW) and time (TDI),** equivalent when evaluated on MSE



State-of-the-art comparison: DILATE training can improve SOTA deep forecasting models (*e.g.* TT-RNN [4]) on shape and time metrics

Eval loss		LSTNet-rec [30]	TT-RNN [60, 61]	Seq2Seq DILATE
Euclidian	MSE (x100)	1.74 ± 0.11	**0.837 ± 0.106**	1.00 ± 0.260
Shape	DTW (x100)	42.0 ± 2.2	25.9 ± 1.99	**23.0 ± 1.62**
	Ramp (x10)	9.00 ± 0.577	6.71 ± 0.546	**5.93 ± 0.235**
Time	TDI (x10)	25.7 ± 4.75	17.8 ± 1.73	**14.4 ± 1.58**
	Hausdorff	**2.34 ± 1.41**	**2.19 ± 0.125**	**2.13 ± 0.514**



Speedup compared to auto-diff

References

[1] L. Vallance et al, Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric, *Solar Energy*
[2] Laptev, Time-series extreme event forecasting with NNs at Uber, ICML'17
[3] Yu, Long-term forecasting using tensor-train RNNs, Arxiv
[4] Deep state space models for time series forecasting, *NeurIPS'18*
[5] Cuturi et al, Soft-DTW, *ICML'17*