Deep Energy-Based NARX Models

Johannes N. Hendriks[1], Fredrik K. Gustafsson[2]
Antônio H. Ribeiro[2], Adrian G. Wills[1],
Thomas B. Schön[2]

[1]The University of Newcastle, Australia
[2]Uppsala University, Sweden

THE UNIVERSITY OF NEWCASTLE AUSTRALIA

UPPSALA UNIVERSITET

Workshop on Nonlinear System Identification Benchmarks
2021

Motivation

Common performance criteria such as **maximum-likelihood** or **prediction-error** criteria usually involve **assumptions** about uncertainty, be they *explicit* or *implicit*

Nonlinear ARX model (Gaussian noise)

- Data model:

$$y_t = f_\theta(y_{t-1}, u_{t-1}) + e_t,$$

where $e_t \sim \mathcal{N}(0, \sigma)$.

Nonlinear ARX model (Gaussian noise)

- Data model:

$$y_t = f_\theta(y_{t-1}, u_{t-1}) + e_t,$$

where $e_t \sim \mathcal{N}(0, \sigma)$.

- $f_\theta \rightsquigarrow$ model structure.

Nonlinear ARX model (Gaussian noise)

- Data model:

$$y_t = f_\theta(y_{t-1}, u_{t-1}) + e_t,$$

  where $e_t \sim \mathcal{N}(0, \sigma)$.
- $f_\theta \rightsquigarrow$ model structure.
- Maximum likelihood estimator:

$$\widehat{\theta} = \arg\min_\theta \sum_{t=1}^{T} \|y_t - f_\theta(y_{t-1}, u_{t-1})\|^2.$$

Energy-based NARX models

- Arbitrary distributions:

$$y_t | (y_{t-1}, u_{t-1}) \sim p_\theta(y_t | y_{t-1}, u_{t-1}),$$

Energy-based NARX models

- Arbitrary distributions:

$$y_t | (y_{t-1}, u_{t-1}) \sim p_\theta(y_t | y_{t-1}, u_{t-1}),$$

- Energy-based model:

$$p_\theta(y_t | y_{t-1}, u_{t-1}) = \frac{e^{g_\theta(y_t, y_{t-1}, u_{t-1})}}{\int e^{g_\theta(\gamma, y_{t-1}, u_{t-1})} \, \mathrm{d}\gamma},$$

Gustafsson, F.K., Danelljan, M., Bhat, G., and Schön,T.B. (2020). Energy-based models for deep probabilistic regression. In Proceedings of the European Conference on Computer Vision (ECCV)

Energy-based NARX models

- Arbitrary distributions:

$$y_t | (y_{t-1}, u_{t-1}) \sim p_\theta(y_t | y_{t-1}, u_{t-1}),$$

- Energy-based model:

$$p_\theta(y_t | y_{t-1}, u_{t-1}) = \frac{e^{g_\theta(y_t, y_{t-1}, u_{t-1})}}{\int e^{g_\theta(\gamma, y_{t-1}, u_{t-1})} \, \mathrm{d}\gamma},$$

Gustafsson, F.K., Danelljan, M., Bhat, G., and Schön,T.B. (2020). Energy-based models for deep probabilistic regression. In Proceedings of the European Conference on Computer Vision (ECCV)

- $g_\theta \rightsquigarrow$ Highly flexible structure: in our case a neural network.

Model training

- Maximum likelihood

$$\widehat{\theta} = \arg\max_{\theta} \sum_{i=1}^{T} - \log p_{\theta}(y_t \mid y_{t-1}, u_{t-1})$$

Model training

- Maximum likelihood

$$\widehat{\theta} = \arg\max_{\theta} \sum_{i=1}^{T} - \log p_{\theta}(y_t \mid y_{t-1}, u_{t-1})$$

$$= \arg\min_{\theta} \sum_{t=1}^{T} \left(-g_{\theta}(y_t, x_t) + \ln \int e^{g_{\theta}(\gamma, x_t)} \, \mathrm{d}\gamma \right)$$

Model training

- Maximum likelihood

$$\widehat{\theta} = \arg\max_{\theta} \sum_{i=1}^{T} -\log p_{\theta}(y_t \mid y_{t-1}, u_{t-1})$$
$$= \arg\min_{\theta} \sum_{t=1}^{T} \left(-g_{\theta}(y_t, x_t) + \ln \int e^{g_{\theta}(\gamma, x_t)} \, \mathrm{d}\gamma \right)$$

- Noise contrastive estimation:

Gutmann, M. and Hyvärinen, A. (2010). Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS), 297–304

Example 1: AR model

$$y_t = 0.95 y_{t-1} + e_t.$$

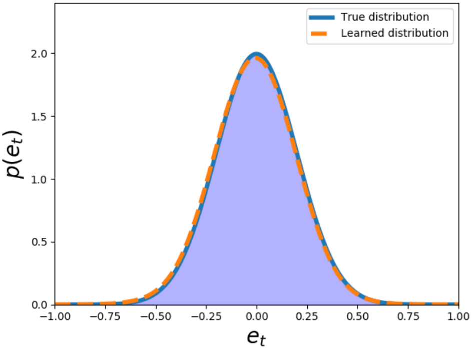


Figure: Gaussian error e_t

Example 1: AR model

$$y_t = 0.95y_{t-1} + e_t.$$

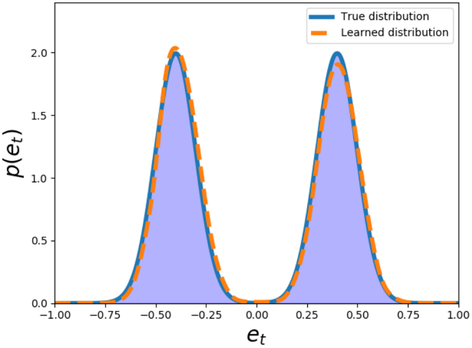


Figure: Gaussian mixture error e_t

Example 1: AR model

$$y_t = 0.95y_{t-1} + e_t.$$

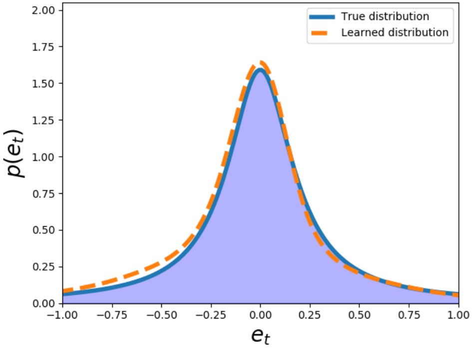


Figure: Cauchy error e_t

Example 1: AR model

$$y_t = 0.95y_{t-1} + e_t.$$

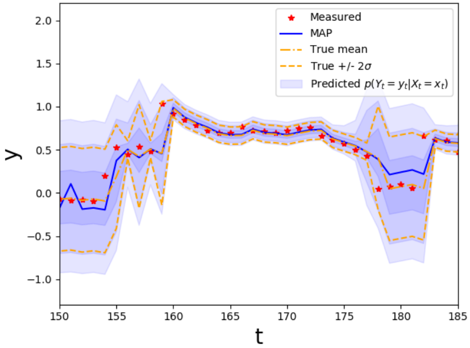
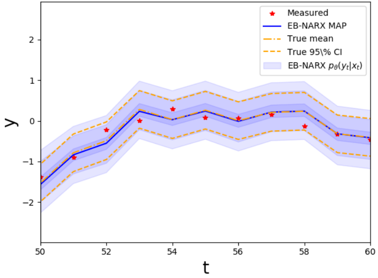


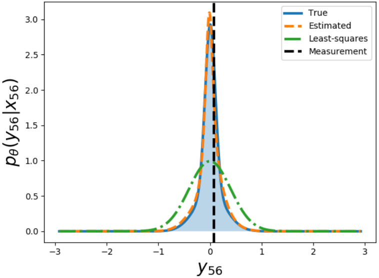
Figure: Gaussian error e_t with conditional variance

Example 2: ARX model

$$y_t = 1.5y_{t-1} - 0.7y_{t-2} + u_{t-1} + 0.5u_{t-2} + e_t,$$



(a) Sequence



(b) t=56

Figure: Estimates of $p_\theta(y_t|x_t)$ for a validation data.

Example 3: nonlinear model

Model:

$$y_t^* = \left(0.8 - 0.5e^{-y_{t-1}^{*2}}\right) y_{t-1}^* - \left(0.3 + 0.9e^{-y_{t-1}^{*2}}\right) y_{t-2}^*$$
$$+ u_{t-1} + 0.2u_{t-2} + 0.1u_{t-1}u_{t-2} + v_t,$$
$$y_t = y_t + w_t$$

Process and output error:

$$v_t \sim \mathcal{N}(0, \sigma_v^2)$$
$$w_t \sim \mathcal{N}(0, \sigma_v^2)$$

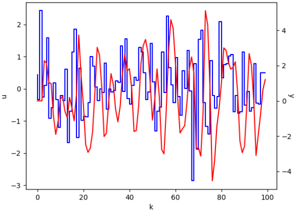


Figure: **System only with process noise.** Input in blue and output in red.

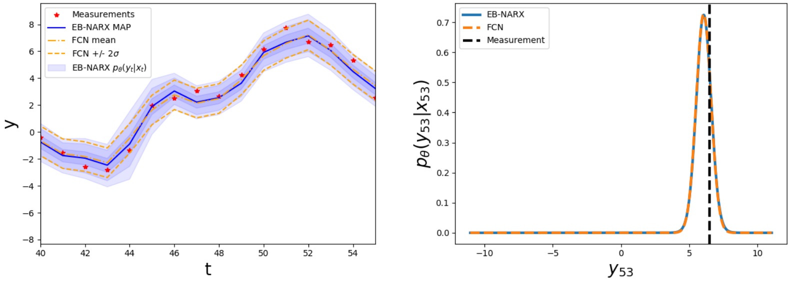
Chen, S., Billings, S.A., and Grant, P.M. (1990). *Non-Linear System Identification Using Neural Networks.* International Journal of Control, 51(6), 1191–1214.

Example 3: nonlinear model

Table: Simulated nonlinear MSE on the validation set for the fully connected network (FCN) NARX model and EB-NARX model

	$N = 100$		$N = 250$		$N = 500$	
	FCN	EB-NARX	FCN	EB-NARX	FCN	EB-NARX
$\sigma = 0.1$	0.122	0.099	0.069	0.070	0.057	0.054
$\sigma = 0.3$	0.398	0.390	0.353	0.354	0.289	0.308
$\sigma = 0.5$	0.860	0.869	0.809	0.822	0.754	0.779

Example 3: nonlinear model



(a) Sequence (b) t=56

Figure: Estimates of $p_\theta(y_t|x_t)$ for a validation data.

Example 4: Coupled Electric Drives

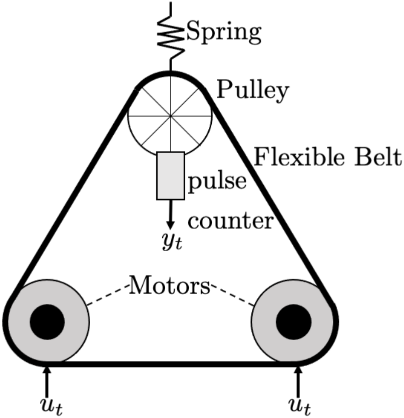


Figure: **Illustration of the CE8 coupled electric drives system**

Wigren, T. and Schoukens, M. (2017). *Coupled electric drives data set and reference models.* Technical Report. Uppsala Universitet, 2017

Example 4: Coupled Electric Drives

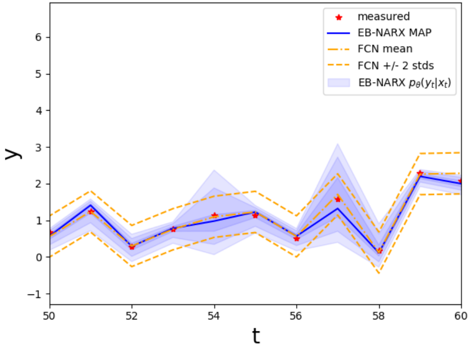


Figure: $p_\theta(y_t|x_t)$ sequence

Example 4: Coupled Electric Drives

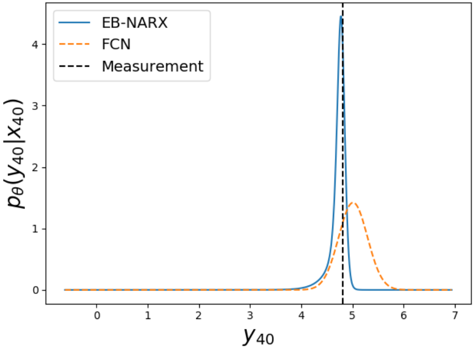


Figure: $t = 40$

Example 4: Coupled Electric Drives

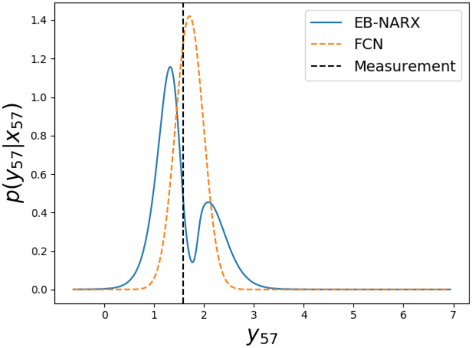


Figure: $t = 57$

Example 4: Coupled Electric Drives

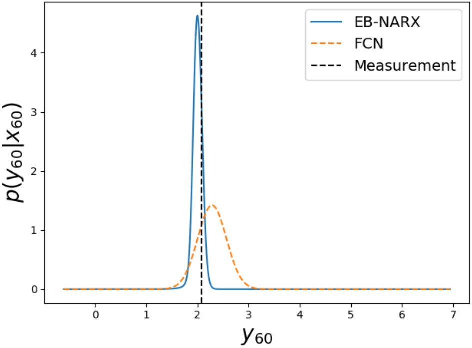


Figure: $t = 60$

Conclusion

- Energy based NARX learns the full conditional distribution rather than the point estimate.

Conclusion

- Energy based NARX learns the full conditional distribution rather than the point estimate.
- The current paper only considers one-step-ahead predictions and not multi-step-ahead predictions.

Conclusion

- Energy based NARX learns the full conditional distribution rather than the point estimate.
- The current paper only considers one-step-ahead predictions and not multi-step-ahead predictions.
- Propagate MAP point estimates *vs* probabilities.

Conclusion

- Energy based NARX learns the full conditional distribution rather than the point estimate.
- The current paper only considers one-step-ahead predictions and not multi-step-ahead predictions.
- Propagate MAP point estimates *vs* probabilities.
- Studying energy-based models for *nonlinear ARMAX*, *output error* and other types of models that can handle more general noise types.

Thank you!

To appear in the 19th IFAC Symposium in System Identification.

arXiv.org **Paper:** https://arxiv.org/abs/2012.04136

Thank you!

To appear in the 19th IFAC Symposium in System Identification.

arXiv.org **Paper:** https://arxiv.org/abs/2012.04136

Code: https://github.com/jnh277/ebm_arx

Thank you!

To appear in the 19th IFAC Symposium in System Identification.

arXiv.org **Paper:** https://arxiv.org/abs/2012.04136

Code: https://github.com/jnh277/ebm_arx

Contact:
johannes.hendriks@newcastle.edu.au
fredrik.gustafsson@it.uu.se
antonio.horta.ribeiro@it.uu.se
adrian.wills@newcastle.edu.au
thomas.schon@it.uu.se