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]School of Engineering, The University of Newcastle, Australia
[2]Department of Information Technology, Uppsala University, Sweden

The 19th IFAC Symposium on System Identification (SYSID 2021)

Introduction

We consider the problem of learning models for dynamic systems based on observed input-output data $\{(u_t, y_t)\}_{t=1}^{T}$.

Specifically, we assume that the current system output y_t is related to past outputs and past inputs $x_t \triangleq \{y_{t-1}, \ldots, y_{t-D_y}, u_{t-1}, \ldots, u_{t-D_u}\}$.

A common approach is to directly regress y_t from x_t, using a neural network f_θ that is trained by minimizing the mean squared error (MSE),

$$\hat{y}_t = f_{\hat{\theta}}(x_t),$$
$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} \|y_t - f_\theta(x_t)\|^2.$$

Introduction

From a probabilistic perspective, this MSE approach corresponds to minimizing the negative log-likelihood $-\sum_{t=1}^{T} \log p_\theta(y_t|x_t)$ for a fixed-variance Gaussian model $p_\theta(y_t|x_t) = \mathcal{N}(y; f_\theta(x_t), \sigma^2)$ of the conditional distribution $p(y_t|x_t)$.

A fixed-variance, unimodal Gaussian model $p_\theta(y_t|x_t) = \mathcal{N}(y; f_\theta(x_t), \sigma^2)$ is however fairly restrictive, and could give a poor approximation of the true distribution $p(y_t|x_t)$ in many practical situations.

In this paper, we instead utilize highly flexible energy-based models $p_\theta(y_t|x_t)$, enabling $p(y_t|x_t)$ to be learned directly from the available data.

Energy-Based NARX Models

We model the distribution $p(y_t|x_t)$ with a conditional energy-based model (EBM),

$$p_\theta(y_t|x_t) = \frac{e^{g_\theta(y_t,x_t)}}{\int e^{g_\theta(\gamma,x_t)}\,\mathrm{d}\gamma}, \tag{1}$$

where g_θ is a neural network that maps any pair (y_t, x_t) to a scalar $g_\theta(y_t, x_t) \in \mathbb{R}$.

The EBM $p_\theta(y_t|x_t)$ is directly specified via the neural network g_θ, which provides a highly flexible class of functions. This enables $p_\theta(y_t|x_t)$ to model a wide range of distributions, including heavy-tailed, asymmetric or multimodal ones.

Since $p_\theta(y_t|x_t)$ in (1) is an EBM that relies on a nonlinear combination of past outputs and inputs x_t, we refer to this as an energy-based NARX (EB-NARX) model.

Energy-Based NARX Models - Training

We model the distribution $p(y_t|x_t)$ with a conditional energy-based model (EBM),

$$p_\theta(y_t|x_t) = \frac{e^{g_\theta(y_t,x_t)}}{\int e^{g_\theta(\gamma,x_t)}\,\mathrm{d}\gamma},$$

where g_θ is a neural network that maps any pair (y_t, x_t) to a scalar $g_\theta(y_t, x_t) \in \mathbb{R}$.

The neural network $g_\theta(y_t, x_t)$ can be trained using various methods for fitting a density $p_\theta(y_t|x_t)$ to observed data $\{(y_t, x_t)\}_{t=1}^{T}$.

Generally, the most straightforward such method is probably to minimize the negative log-likelihood $\mathcal{L}(\theta) = -\sum_{t=1}^{T} \log p_\theta(y_t|x_t)$, which for the EBM $p_\theta(y_t|x_t)$ is given by,

$$\mathcal{L}(\theta) = \sum_{t=1}^{T} \log\left(\int e^{g_\theta(\gamma,x_t)}\,\mathrm{d}\gamma\right) - g_\theta(y_t, x_t).$$

We model the distribution $p(y_t|x_t)$ with a conditional energy-based model (EBM),

$$p_\theta(y_t|x_t) = \frac{e^{g_\theta(y_t,x_t)}}{\int e^{g_\theta(\gamma,x_t)}\,\mathrm{d}\gamma},$$

where g_θ is a neural network that maps any pair (y_t, x_t) to a scalar $g_\theta(y_t, x_t) \in \mathbb{R}$.

The integral $\int e^{g_\theta(\gamma,x_t)}\,\mathrm{d}\gamma$ is generally intractable, preventing exact evaluation of $\mathcal{L}(\theta)$, but can be approximated using numerical integration techniques.

We instead employ noise contrastive estimation (NCE) to train $g_\theta(y_t, x_t)$.

Energy-Based NARX Models - Training using NCE

$$p_\theta(y_t|x_t) = \frac{e^{g_\theta(y_t,x_t)}}{\int e^{g_\theta(\gamma,x_t)}\,\mathrm{d}\gamma}$$

Noise contrastive estimation (NCE) entails learning to discriminate between observed data examples and samples drawn from a noise distribution.

Specifically, $g_\theta(y_t, x_t)$ is trained by minimizing the cost function $L(\theta) = -\frac{1}{T}\sum_{t=1}^{T} L_t(\theta)$,

$$L_t(\theta) = \log \frac{\exp\left(g_\theta(y_t^{(0)}, x_t) - \log q(y_t^{(0)}|y_t)\right)}{\sum_{m=0}^{M} \exp\left(g_\theta(y_t^{(m)}, x_t) - \log q(y_t^{(m)}|y_t)\right)},$$

where $y_t^{(0)} \triangleq y_t$, and $\{y_t^{(m)}\}_{m=1}^{M}$ are M samples drawn from a noise distribution $q(y|y_t)$ that depends on the true output y_t.

Energy-Based NARX Models - Training using NCE

$$L(\theta) = -\frac{1}{T}\sum_{t=1}^{T} L_t(\theta), \quad L_t(\theta) = \log \frac{\exp\left(g_\theta(y_t^{(0)}, x_t) - \log q(y_t^{(0)}|y_t)\right)}{\sum_{m=0}^{M} \exp\left(g_\theta(y_t^{(m)}, x_t) - \log q(y_t^{(m)}|y_t)\right)},$$

$$y_t^{(0)} \triangleq y_t, \quad \{y_t^{(m)}\}_{m=1}^{M} \sim q(y|y_t) \text{ (noise distribution).}$$

Effectively, $L(\theta)$ is the softmax cross-entropy loss for a classification problem with $M+1$ classes (which of the $M+1$ values $\{y_t^{(m)}\}_{m=0}^{M}$ is the true output y_t?).

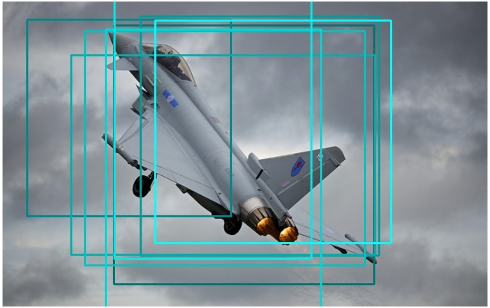
The noise distribution $q(y|y_t)$ is a mixture of K Gaussians centered at y_t,

$$q(y|y_t) = \frac{1}{K}\sum_{k=1}^{K} \mathcal{N}(y; y_t, \sigma_k^2 I).$$

Energy-Based NARX Models - Training using NCE

$$L(\theta) = -\frac{1}{T}\sum_{t=1}^{T} L_t(\theta), \quad L_t(\theta) = \log \frac{\exp\left(g_\theta(y_t^{(0)}, x_t) - \log q(y_t^{(0)}|y_t)\right)}{\sum_{m=0}^{M} \exp\left(g_\theta(y_t^{(m)}, x_t) - \log q(y_t^{(m)}|y_t)\right)},$$

$$y_t^{(0)} \triangleq y_t, \quad \{y_t^{(m)}\}_{m=1}^{M} \sim q(y|y_t) \text{ (noise distribution)}.$$



Energy-Based NARX Models - Prediction

We model the distribution $p(y_t|x_t)$ with a conditional energy-based model (EBM),

$$p_\theta(y_t|x_t) = \frac{e^{g_\theta(y_t,x_t)}}{\int e^{g_\theta(\gamma,x_t)}\,\mathrm{d}\gamma},$$

where g_θ is a neural network that maps any pair (y_t, x_t) to a scalar $g_\theta(y_t, x_t) \in \mathbb{R}$.

Given x_t at test-time, we predict a point estimate $\hat{y}_t$ by maximizing $p_\theta(y_t|x_t)$,

$$\hat{y}_t = \operatorname*{argmax}_{y_t} p_\theta(y_t|x_t) = \operatorname*{argmax}_{y_t} g_\theta(y_t, x_t).$$

Since there is no guarantee that $p_\theta(y_t|x_t)$ is unimodal, we evaluate $g_\theta(y_t, x_t)$ for a range of values y_t and then refine the best of these via T steps of gradient ascent,

$$y_t \leftarrow y_t + \lambda \nabla_{y_t} g_\theta(y_t, x_t).$$

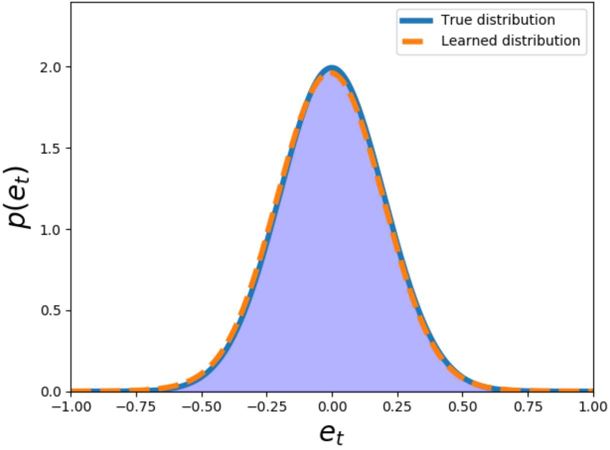
Examples

We provide several examples which illustrate the utility of the EB-NARX model when applied to data from dynamic systems. These examples include both simulated linear and non-linear data, as well as real data from the CE8 coupled electric drives data set.

For the linear examples, qualitative comparisons are made between the estimated and true distributions. We also compare EB-NARX with a fully-connected network (FCN).

Python code for these examples is available at `github.com/jnh277/ebm_arx`.

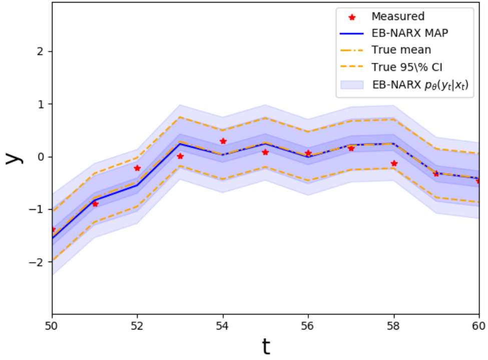
$$y_t = 0.95y_{t-1} + e_t.$$



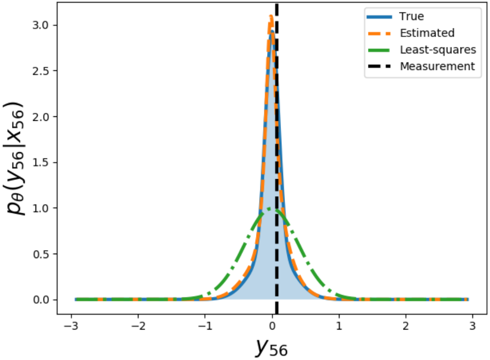
Gaussian error e_t.

Examples - Linear ARX

$$y_t = 1.5y_{t-1} - 0.7y_{t-2} + u_{t-1} + 0.5u_{t-2} + e_t, \quad e_t \sim 0.6\mathcal{N}(0, 0.1^2) + 0.4\mathcal{N}(0, 0.3^2).$$



(a) Sequence



(b) t=56

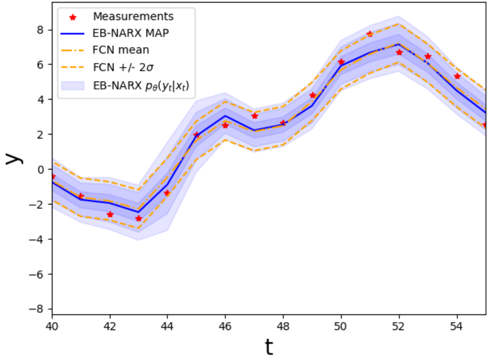
True and estimated $p(y_t|x_t)$ for validation data.

$$
\begin{aligned}
y_t^* = & \left(0.8 - 0.5e^{-y_{t-1}^{*2}}\right) y_{t-1}^* - \left(0.3 + 0.9e^{-y_{t-1}^{*2}}\right) y_{t-2}^* \\
& + u_{t-1} + 0.2u_{t-2} + 0.1u_{t-1}u_{t-2} + v_t, \quad v_t \sim \mathcal{N}(0, \sigma^2) \\
y_t = & y_t^* + w_t, \quad w_t \sim \mathcal{N}(0, \sigma^2).
\end{aligned}
$$

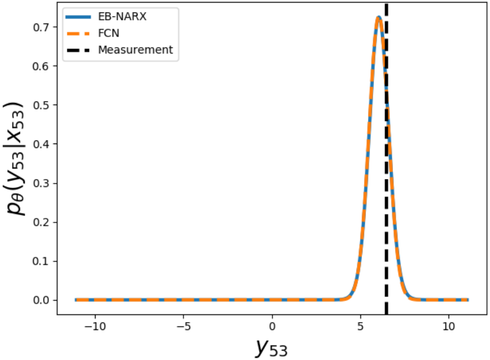
Table 1: Validation set MSE for the fully-connected network (FCN) and EB-NARX model, trained on datasets generated with different noise levels σ and lengths (N).

	$N = 100$		$N = 250$		$N = 500$	
	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

Examples - Simulated Nonlinear Problem



(c) Sequence



(d) t=53

Estimates of $p(y_t|x_t)$ for validation data.

Examples - Real Data: Coupled Electric Drives

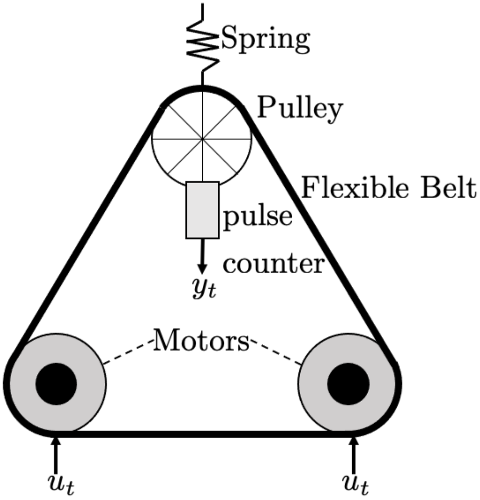
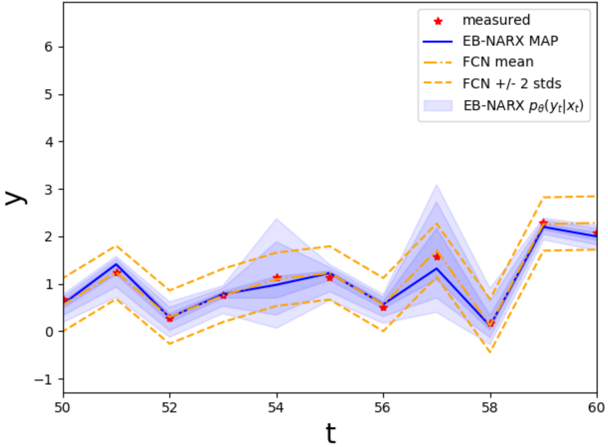


Illustration of the CE8 coupled electric drives system.

Examples - Real Data: Coupled Electric Drives



Estimates of $p(y_t|x_t)$ for a sequence of validation data.

Conclusion

We directly learned the full conditional distribution $p(y_t|x_t)$ for dynamic systems using energy-based models, thus demonstrating their potential within system identification.

Our EB-NARX model $p_\theta(y_t|x_t)$ could learn both very simple and more complex distributions directly from observed data.

We have only considered one-step-ahead prediction. It is not clear how to best propagate $p_\theta(y_t|x_t)$ for multi-step-ahead prediction.

We have only considered NARX models. Future work could explore how to best extend the approach to other model types.