Nonlinear systems identification

Mario Milanese, Carlo Novara

Dipartimento di Automatica e Informatica
Politecnico di Torino

Nonlinear regression systems

- Consider a nonlinear system in regression form:

$$y^{t+1} = f(w^t) + d^{t+1}$$

where:

- w^t : regressor. It defines the system structure:

$$w^t = \left[y^t \ y^{t-1} \dots u^t \ u^{t-1} \dots\right]^T \quad \Leftrightarrow \quad \text{NARX}$$

$$w^t = \left[f\left(w^t\right) f\left(w^{t-1}\right) \dots u^t \ u^{t-1} \dots\right]^T \quad \Leftrightarrow \quad \text{NOE}$$

$$w^t = \left[y^t \ y^{t-1} \dots u^t \ u^{t-1} \dots d^t \ d^{t-1} \dots\right]^T \quad \Leftrightarrow \quad \text{NARMAX}$$

- u : input signal.
- d : noise acting on the system.

Nonlinear regression systems

- The predictor of system f is defined as:

$$\hat{y}^{t+1} = f(w^t)$$

where:

$$w^t = \left[y^t \ y^{t-1} \ldots u^t \ u^{t-1} \ldots\right]^T \quad \Leftrightarrow \quad \text{NARX}$$

$$w^t = \left[\hat{y}^t \ \hat{y}^{t-1} \ldots u^t \ u^{t-1} \ldots\right]^T \quad \Leftrightarrow \quad \text{NOE}$$

$$w^t = \left[y^t \ y^{t-1} \ldots u^t \ u^{t-1} \ldots \varepsilon^t \ \varepsilon^{t-1} \ldots\right]^T \quad \Leftrightarrow \quad \text{NARMAX}$$

$\varepsilon^t = y^t - \hat{y}^t$: prediction error

Making inferences from data

- Suppose that a finite number of noise corrupted measurements have been generated by a system described by a regression function f^o:

$$y^{t+1} = f^o(w^t) + d^{t+1}, \quad t = 1,2,\ldots,N$$

- Suppose that the function f^o is not known.

Nonlinear identification problem:

Find and estimate $\hat{f}$ of f^o: $\hat{f} \cong f^o$

Making inferences from data

- Problems :
  - ***for a given estimate*** $\hat{f} \cong f^o$

    evaluate the identification error $\left\| f^o - \hat{f} \right\|$
  - ***find an estimate*** $\hat{f} \cong f^o$

    "minimizing" the identification error

- The identification error cannot be exactly evaluated since f^o and w^t are not known.

- Need of prior assumptions on f^o and d^t for deriving finite bounds on inference error.

Parametric approach

- Typical assumptions in literature:
  - on system: $f^o \in \Psi(\theta) = \left\{ f(w,\theta) = \sum_{i=1}^{r} \alpha_i \sigma(w,\beta) \right\}$
  - on noise: iid stochastic

- Functional form of f^o:
  - derived from physical laws
  - σ_i: "basis" function (polynomial, sigmoid,..).

- Parameters θ are estimated by means of the Prediction Error (PE) method.

Parametric approach

- Predictor: $$\hat{y}^{t+1} = f(w^t, \theta) = \sum_{i=1}^{r} \alpha_i \sigma\left(w^t, \beta_i\right)$$

- Given N noise-corrupted measurements of y^t, w^t:

$$\begin{aligned} y^2 &= f(w^1, \theta) + \varepsilon^2 \\ y^3 &= f(w^2, \theta) + \varepsilon^3 \\ &\vdots \\ y^{N+1} &= f(w^N, \theta) + \varepsilon^{N+1} \end{aligned} \quad \Longrightarrow \quad Y = F\left(\theta\right) + D_\varepsilon$$

Measured output (Y)

Known function of θ (F)

Prediction Errors (D_ε)

Parametric approach

- Given the measurements equation:

$$Y = F(\theta) + D_\varepsilon$$

It is possible to estimate θ by means of the Prediction Error (PE) method:

$$\hat{\theta}^{LS} = \arg\min_{\theta} V_N(\theta)$$

$$V_N(\theta) = \frac{1}{N} D_\varepsilon^T D_\varepsilon = \frac{1}{N}[Y - F(\theta)]^T [Y - F(\theta)]$$

Problem: $V_N(\theta)$ is in general non-convex.

Parametric approach

- If possible, physical laws are used to obtain the parametric representation of $f(w,\theta)$.

- When the physical laws are not well known or too complex, black-box parameterizations are used.

Fixed basis parameterization

Polinomial, trigonometric, etc.

Tunable basis perametrization

Neural networks

Fixed basis functions

$$f(w,\theta) = \sum_{i=1}^{r} \alpha_i \sigma_i(w) \qquad \theta = [\alpha_1 \cdots \alpha_r]^T$$

$\sigma_i(w)$: Basis functions

- **Problem:** Can σ_i's be found such that:

$$f(w,\theta) \xrightarrow[r\to\infty]{} f^o(w) \qquad ?$$

Fixed basis functions

- For continuous f^o, bounded $W \subset \Re^n$ and σ_i polynomial of degree i (Weierstrass):

$$\lim_{r \to \infty} \sup_{w \in W} \left| f^o(w) - f(w, \theta) \right| = 0$$

Polynomial models

Fixed basis functions

$$f(w,\theta) = \sum_{i=1}^{r} \alpha_i \sigma_i(w) \qquad \theta = [\alpha_1 \cdots \alpha_r]^T$$

- NARX models: PE estimation of θ is a linear problem:

$$Y = L\theta + D_\varepsilon$$

$$L = \begin{bmatrix} \sigma_1(w^1) & \cdots & \sigma_r(w^1) \\ \vdots & \ddots & \vdots \\ \sigma_1(w^N) & \cdots & \sigma_r(w^N) \end{bmatrix} \qquad Y = \begin{bmatrix} y^2 \\ \vdots \\ y^{N+1} \end{bmatrix}$$

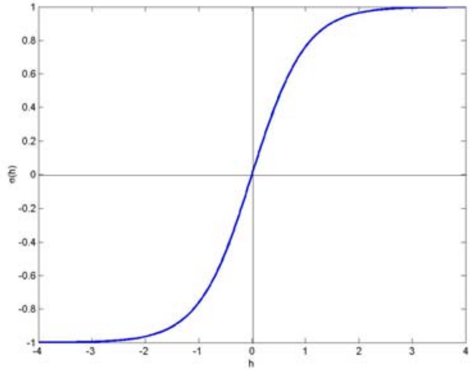
- Least squares solution: $\hat{\theta}^{LS} = \left(L^T L\right)^{-1} L^T Y$

Tunable basis functions

$$f(w,\theta)=\sum_{i=1}^{r}\alpha_i\sigma(w,\beta_i)$$

$$\theta=\left[\alpha_1\cdots\alpha_r\ \beta_{11}\cdots\beta_{rq}\right]^T,\quad \beta_i\in\Re^q$$

- One of the most common tunable parameterization is the one-hidden layer sigmoidal neural network.



Parametric models

- Model structure choice:
  - Basis functions
  - Number of Basis functions
  - Number of regressors

The complexity of such problem can be exponential in n

- **Problem:** curse of dimensionality

  The number of parameters needed to obtain "accurate" models may grow exponentially with the dimension n of regressor space.

More relevant in the case of fixed basis functions

Tunable basis functions

- Under suitable regularity conditions on the function to approximate, the number of parameters required to obtain "accurate" models grows linearly with n.

- Estimation of θ requires to solve a non-convex minimization problem (even for NARX models).

↓

Trapping in local minima