Exercise: water heater model identification (real data)

Consider a water heater whose input is the resistor voltage (measured in percentage) and output is the water temperature (measured in Celsius degrees). Input-output measurements of this system, to be modeled for prediction, have been collected in two different datasets in the MATLAB `heater.mat` file.

Problem:

1) Identify from experimental data several ARX, OE and NARX models of different orders and delays (assume $n_a = n_b = n_f \leq 5$, $n_k \leq 5$), using only one dataset and looking for models guaranteeing satisfactory characteristics of whiteness of the residuals associated to this first dataset (please specify the threshold to suitably pass the whiteness test).

2) Compare only the models that guarantee satisfactory residual characteristics in the previous step, by considering the following two criteria, both applied to the other dataset not used for identification:

- the Root Mean Square Error $RMSE = \sqrt{\frac{1}{N - N_0} \sum_{t=N_0+1}^{N} [y(t) - \hat{y}(t)]^2}$, where $y(t)$ = measured output, $\hat{y}(t)$ = predicted output and N_0 is a time instant to be suitably chosen after which the transient is past;
- the Rissanen's Minimum Description Length Criterion (MDL).

Among all these models coming from different classes, choose the two best trade-offs between:
- $RMSE$ and model complexity n, with $n = n_a + n_b$ (for ARX and NARX models) or $n = n_f + n_b$ (for OE models),
- MDL and model complexity n,

and report the corresponding linear models in predictor form with their block diagrams and transfer functions.

Main steps:

1) Load the `heater.mat` file, that contains the estimation dataset ES (`ue`, `ye`: 2000 input-output samples) and the validation dataset VS (`uv`, `yv`: 1000 input-output samples).

2) Remove the mean value from all the measured signals.

3) Using the unbiased ES dataset, identify several linear models of different orders and delays:
- $ARX(n_a, n_b, n_k)$, with $n_a = n_b = 1, \ldots, 5$ and $n_k = 1, \ldots, 5$
- $OE(n_b, n_f, n_k)$, with $n_b = n_f = 1, \ldots, 5$ and $n_k = 1, \ldots, 5$

4) Using the unbiased ES dataset, perform the whiteness residual test of each identified model, reporting into a table the number of residuals sufficiently outside the confidence interval of the corresponding Autocorrelation plot: if this number is greater than a reasonable threshold to be suitably selected, then the model is wasted; otherwise, the model is considered for further analysis.

5) Using the unbiased VS dataset, consider the $RMSE$ and MDL model selection criteria, reporting in two tables and plotting in two figures the corresponding values: compare only the models that passed the whiteness residual test to choose the best trade-off between $RMSE$ (or MDL) and model complexity n.

6) Using the unbiased ES dataset, identify several nonlinear neural models of different orders and delays $NARX(n_a, n_b, n_k)$, with $n_a = n_b = 1, \ldots, 5$ and $n_k = 1, \ldots, 5$, using different values of the number r of basis functions (neurons) in the range $[5, 10]$.

7) Using the unbiased VS dataset, consider the $RMSE$ and MDL model selection criteria, reporting in two tables and plotting in two figures the corresponding values, and choose the best trade-off between $RMSE$ (or MDL) and model complexity n.