• Process models are the cornerstone for the successful deployment of advanced control and real-time optimisation routines.

• Models often need to be customised with plant data to fit the actual processes, but fully replacing process knowledge by so-called deep learning is not a good idea.

• Data-driven models often match plants just at the current or past operation but predict unphysical responses far from these.

**The problem**

**The solution**

• UVa developed a systematic methodology for grey-box modelling that allows the engineer to transfer the available process knowledge into a model.

• The methodology builds upon nonlinear data reconciliation, driven by basic first-principles equations, and constrained regression to seek for additional experimental equations among process variables.

• In this novel machine-learning framework, several desirable physical features can be enforced on the data-driven models.

---

This project receives funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 723575.

www.spire2030.eu/copro
Estimation of grey-box models with constraints

The problem

Pure machine learning is not enough

Recent computational advances opened the door to the development of computer-aided systems which support plant operators to solve complex decision problems in real time. However, traditional rigorous physical-chemical models used in process design are not always suitable for their usage within real-time decision support systems because of their excessive computational complexity or because they cannot be fitted to the actual plant data well enough.

This limitation can be overcome to some extent thanks to the increased amount of data due to the industrial advances in digitalisation, and the new techniques for machine learning. However, machine learning relies on data of good quality and in large quantity, and data from industrial process systems often lacks both: plants usually operate around the same set-points and tight production constraints prevent running experiments for data collection in other regions of the operational space. Hence, using prediction models that were exclusively built from historical data may lead to unreliable models, even if the data has been pre-treated and standard regularisation techniques are used in the fitting problem.

The solution

A two-stage modelling methodology

In the scientific process-systems community, the agreed way to approach the above problem is to build so-called grey-box models. These are hybrid models that are partly built from first principles and partly from plant data, with the aim of combining the advantages of both approaches: physical coherence and accurate match with the observed plant outputs.

The systematic methodology for grey-box process modelling proposed by UVa is structured into two stages: 1) estimation and 2) constrained regression.

The first stage starts from a set of basic first-principle laws of the process (e.g. mass and energy balances) and a large set of process data recorded from sensors, usually pre-treated to exclude plant stops and other operation outliers. Then, a dynamic data reconciliation problem is formulated (with this incomplete model) and solved via nonlinear programming. Numerical integrators and nonlinear optimisation algorithms are the involved tools in this stage. The outcome is a set of estimates of all variables over time which is as coherent as possible with the basic physics of the process.

Note that the model is still incomplete at this point, i.e., more inputs than the actual are needed to compute the outputs. Therefore, a constrained regression problem is stated and solved in a second stage, in order to find the unknown relationships between some process variables from experimental data, hence completing the model. Here, two main features make the difference:

- The data for regression is a combination of virtual data for the rest of the variables that are generated in the estimation stage.
- Unconstrained regression is extended to include physical knowledge on the candidate models, such as (local) bounds on the outputs (e.g. positivity), constraints on the model slope or curvature, etc.

If candidate experimental sub-models are of polynomial structure, this constrained regression problem can be handled efficiently via sum-of-squares (SOS) programming (convex optimisation).

The summary

Coherent and reliable grey-box models

The two-stage systematic methodology proposed by UVa provides a framework where the knowledge that the engineer may have about the process is explicitly included in the data-driven modelling procedure. The objective is to avoid wrong predictions that contradict the known physics of the process, which would invalidate the suggestions provided by decision-support systems that are based on such models.

The SOS-constrained regression provides well-behaved polynomial models even in the presence of scarce measurements, which is a limiting factor in other machine-learning techniques. It can be combined with other standard regularisation approaches that balance model complexity with the accuracy of the fit to the data in the objective function.

The developers

Dr. José Luis Pitarch
Supervision and Process Control research group, Systems Engineering and Automatic Control Dpt. Industrial Engineering School Universidad de Valladolid 47011 Valladolid, Spain jose.pitarch@autom.uva.es

Prof. Dr. César de Prada
Supervision and Process Control research group, Systems Engineering and Automatic Control Dpt. Industrial Engineering School & Institute of Sustainable Processes Universidad de Valladolid 47011 Valladolid, Spain prada@autom.uva.es

Absolute LS regression error for modelling a heat-transfer coefficient.

<table>
<thead>
<tr>
<th>Method</th>
<th>Training</th>
<th>Validation</th>
<th>Total</th>
<th>Fit deterioration</th>
<th>Extrapolation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential regularised</td>
<td>13.448</td>
<td>14.282</td>
<td>27.73</td>
<td>-</td>
<td>Unreliable</td>
</tr>
<tr>
<td>SOS constrained</td>
<td>14.751</td>
<td>13.362</td>
<td>28.113</td>
<td>1.36%</td>
<td>Acceptable</td>
</tr>
<tr>
<td>First principles</td>
<td>–</td>
<td>–</td>
<td>37.361</td>
<td>25.78%</td>
<td>Acceptable</td>
</tr>
</tbody>
</table>