An adaptive data-driven modelling and optimization framework for complex chemical process design

Thomas Savage,a Hector Fernando Almeida-Trasvina,a Ehecatl Antonio del Río-Chanona,b Robin Smith,a Dongda Zhang,a,b

aCentre for Process Integration, University of Manchester, The Mill, Sackville Street, Manchester, M1 3AL, UK.
bCentre for Process Systems Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK.
dongda.zhang@manchester.ac.uk

Abstract

Current advances in computer-aided chemical process design and synthesis take advantage of surrogate modelling and superstructure optimization techniques. Conventionally, this is completed by using first-principle physical models or data-driven models to replace the original rigorous models for optimization and selection of a specific unit operation. Despite its achievements, this strategy is inefficient when dealing with complex process flowsheets such as utility and refrigeration systems where a large number of unit operations are heavily connected by recycling streams. To address this problem, an integrated data-driven modelling and optimization framework is proposed in this work. The framework first constructs a hybrid machine learning based surrogate model to automatically reduce the system dimensionality and capture the nonlinearity of the underlying chemical process. Then, an efficient optimization algorithm, in specific, evolutionary algorithm, is embedded to identify the optimal solution of this surrogate model. Quality and accuracy of the estimated optimal solution is finally validated against the rigorous process model. Through an iterative approach, optimal operating conditions for the entire process flowsheet are efficiently identified. Furthermore, the novel CryoMan Cascade cycle system for large scale liquefied natural gas manufacturing is used as the case study. This framework is demonstrated to be superior regarding time-efficiency, solution quality, and flexibility over the rigorous model based optimization approach.

Keywords: large-scale chemical process, surrogate modelling, dimensionality reduction, Gaussian processes, artificial neural network.

1. Introduction

Developing disruptive digital technology to enable the design and operation of cost-effective and energy-efficient manufacturing systems is one of the grand research themes under the context of the 4th Industrial Revolution. Given the large amount of data accumulated from process industries, building data-driven models to enable rapid decision-making is of critical importance in order to guarantee the process performance and safety (Bhosekar and Ierapetritou, 2018). This directly triggers the development and application of surrogate modelling technology in the current industry and research community (McBride and Sundmacher, 2019).

Specific to the chemical industry, rigorous process models derived from mass and energy balances allow for accurate determination of states of a system, and can be directly constructed using multiple computer-aided software packages such as Aspen (Bhosekar
However, due to their complexity, the computational time to evaluate a rigorous model can be relatively large (e.g. potentially weeks to months), hence limiting their applications for process control and optimization (del Rio-Chanona et al., 2018; McBride and Sundmacher, 2019). As a result, surrogate models have been adopted to replace rigorous models and help with identifying optimal operating conditions.

However, at present, surrogate models are predominantly used to replace single unit operations (Henao and Maravelias, 2011; Quirante, Javaloyes and Caballero, 2015), whilst their applicability in terms of substituting more complex systems such as entire process flowsheets has not been well explored. Therefore, in this study, different cutting-edge surrogate models are tested to simulate and optimize a novel liquefied natural gas (LNG) refrigeration cycle, with their performance thoroughly compared against the optimal solution identified through the rigorous model. The structure of this paper is organized as follows. Section 2 introduces the LNG production process. The approach to optimize the rigorous model and the associated primary challenges are also explained. Section 3 details the construction and optimization of different surrogate models. Section 4 summarizes the results of these surrogate models and compares them with the rigorous model’s verification. A thorough discussion regarding the advantages of combining Gaussian processes and partial least squares for large scale complex chemical process simulation and optimization is presented in the Results and Discussion section.

2. Introduction to the CryoMan Cycle
Commercial scale production of LNG involves the use of large, complex and energy-intensive refrigeration cycles. The costs associated with the energy for refrigerant compression (shaft work energy) dominate the overall operating costs of the LNG plant. The CryoMan Cascade cycle recently developed by (Almeida-Trasvina and Smith, 2018) (shown in Fig. 1(a)) is a novel refrigeration configuration that can save significant energy demand in shaft work compared to current commercial processes. In this process, within the precooing cycle, a ‘heavy’ mixed refrigerant provides cooling in a series of two multi-stream heat exchangers (MSHEs); three stages are employed for refrigerant compression. In the liquefaction cycle, the ‘light’ mixed refrigerant is first fed into a liquid-vapour separator. The resulting outlet streams are partially mixed with each other to create the two refrigerant streams that provide cooling in a series of two MSHEs. Refrigerant compression is also carried out in three compression stages. In both precooing and liquefaction cycles, a pump after the second compression stage pumps any condensed refrigerant to the compressor discharge pressure.

Figure 1: The CryoMan Cascade cycle (a), and its KPLS based surrogate model (b).
An adaptive driven modelling and optimization framework for complex chemical process design

The rigorous model of the CryoMan Cascade cycle was implemented in Aspen HYSYS v8.2 and subsequently linked to MATLAB. This resulted in a callable function taking 31 inputs and producing 20 outputs. The inputs include the refrigerant mass flowrates, refrigerant compositions, discharge pressure, refrigerant evaporating pressures, MSHE outlet temperatures, compression ratios and refrigerant split fractions (where applicable) for both precooling and liquefaction cycles. The outputs from the process are eight values of shaft work demand, four MSHE approach temperatures, vapour fractions of four streams (to assess wetness at inlet of compressors), and four compression ratios.

The optimization problem for the rigorous model is defined as follows:

\[
\min_{\varphi} \left( \sum_{i=1}^{N} W_i \right) / m_{LNG}
\]

s.t. \( \Delta T_{\text{min}} \geq 2^\circ\text{C}; \; P_{\text{rat}} \leq 3.5; \; \sum_{j=1}^{m} x_j = 1, \; x_j \in X^\text{MR}; \; V F_{\text{ref}} = 1; \; \varphi_{lb} \leq \varphi \leq \varphi_{ub} \)

where the specific shaft work, defined as the sum of the individual shaft works \( W_i \) divided by the mass flow rate of LNG \( m_{LNG} \), is the objective function to be minimized. The 31 inputs to the rigorous model are represented by \( \varphi \) with the set of inputs having corresponding upper and lower bounds \( \varphi_{ub} \) and \( \varphi_{lb} \), respectively. Constraints include minimum approach temperatures for MSHEs \( \Delta T_{\text{min}} \), maximum compression ratios \( P_{\text{rat}} \) to discourage mechanical damage to compressors, valid molar compositions represented by \( x_j \), and no wetness within compressors by constraining vapour fractions at the inlet of compressors \( V F_{\text{ref}} \) to zero. An evolutionary algorithm (EA) is first employed for the optimization of the rigorous model. The way on which constraints are dealt with in stochastic optimization is by applying penalties to invalid solutions. Successive Quadratic Programming (SQP) is next used to identify a local optimal solution around the best candidate resulting from the stochastic optimization.

3. Developing Surrogate Models

Directly optimizing the rigorous model is time consuming (over 17 hours per run), whilst the real-time decision-making of a commercial LNG plant is often around once per 4 hours. Therefore, surrogate models are used to resolve this challenge. Different from using surrogate models to substitute a unit operation, building a surrogate model to replace an entire process flowsheet is more challenging due to the high nonlinearity of the underlying process and high dimensionality of the involved design variables. To guarantee success, several surrogate model structures were proposed and their performance was thoroughly compared in this work. Furthermore, taking advantage of both supervised (dealing with nonlinearity) and unsupervised (dealing with dimensionality reduction) machine learning techniques is another strategy proposed in this work, as this may greatly simplify the process complexity for surrogate model construction and meanwhile obtain high quality optimal solutions.

3.1. Artificial Neural Network based surrogate models

To consider the nonlinearity of this highly interconnected process flowsheet, different structures of artificial neural networks (ANN) based surrogate models are constructed. These include: (1) a single ANN directly simulating the entire process (31 inputs and 20 outputs); (2) an ANN framework comprising 2 independent ANNs, one simulating the 8 shaft works, 4 compression ratios and 4 vapour fractions given the 31 inputs, and the other simulating the 4 temperature differences given the same inputs; (3) an ANN framework comprising 5 separate ANNs, one simulating the 8 shaft works, 4 compression ratios and 4 vapour fractions, and the other 4 each of which only simulating a specific...
temperature difference. The reasoning behind designing different ANN model structures is thoroughly discussed in the Results and Discussion section.

3.2. Kriging Partial Least Squares based surrogate model

In the method described by (Bouhlel et al., 2016) to integrate partial least squares (PLS) into Gaussian Processes (GPs), PLS is performed in the data and the PLS weights (projected spaces of inputs and outputs) are used to construct a lower dimensional covariance matrix over which a GP model is constructed. By reducing the model dimensionality it allows for easier determination of the minimum log-likelihood resulting in a lower computational time cost, and more accurate mapping of between the input space and the output space. This integrated modelling strategy (embedding PLS into GP), namely KPLS, is adopted in this research to complete the construction of the GP based surrogate model superstructure. 7 GP models are constructed in this work to account for the 7 most important outputs. The structure of this model is shown in Fig. 1(b).

3.3. Data generation and selection

Three different datasets were generated with differing qualities and time costs in order to create the surrogate models: Dataset 1: 5000 random data points. Data points in this dataset were generated randomly within 10 minutes. Despite the negligible time cost, this dataset consists of mostly invalid solutions due to the nonlinearity and complexity of the LNG production process; Dataset 2: 500 refined data points. The input bounds were slightly more constricted to result in an increased proportion of valid solutions; Dataset 3: 350 high-quality data points. This dataset consists completely of valid solutions and was created by discarding invalid solutions. Despite the high accuracy, it takes 50 minutes to generate this dataset, thus the time cost is much higher than Dataset 1.

As the accuracy of a surrogate model heavily relies on the data quality, Dataset 1 was found incapable of constructing an accurate surrogate model. However, having a large size of data is also a prerequisite to build an accurate surrogate model, thus purely using Dataset 3 (i.e. only using feasible points) is not enough. Hence, Datasets 2 and 3 were combined (well mixed) to provide global representation of the rigorous model as well maintain a relatively large proportion of valid solutions. For a highly nonlinear system with 31 inputs and 20 outputs, 850 data points is not necessarily large. The ratio between feasible and infeasible data points in the combined dataset is approximate 3:1. Although extra high-quality data can be generated, the overall surrogate model construction time will be increased, diminishing its advantage over the rigorous model.

A specifically designed evolutionary algorithm was employed to optimize the surrogate models, taking advantage of tournament selection and single point crossover. A mutation rate of 3% and selection percentage of 98% were used over a population size of 100. The optimization scheme was chosen to run for 500 generations. Moreover, a linear penalty is implemented in order to guide the optimization initially towards a valid solution space. Finally, the ANN based surrogate models were implemented using PyTorch 1.2.0 and the KPLS based surrogate models were implemented using the SMT toolbox by (Bouhlel et al., 2019), both within Python 3.7.3 on a Windows 10 operating system. Total construction of the KPLS models takes on average 2 minutes using 850 data points, and that for the ANN based surrogate models takes approximately 5 minutes.

4. Results, Discussion and Conclusions

4.1. Results of ANN based surrogate model frameworks

As shown in Fig. 2, after validation with the rigorous model, it is found that although the single ANN model can well predict total shaft work and vapour fractions, its predicted
An adaptive driven modelling and optimization framework for complex chemical process design

MSHE approach temperatures (i.e. temperature constraints) are infeasible and greatly deviate from the rigorous model’s verification result. The single large ANN is here unable to capture the nonlinearity of the process, and its accuracy is verified to be low. A straightforward improvement to the single ANN surrogate model is to split this large ANN into a two separate ANNs to deal with the nonlinear temperature constraints. In other words, a surrogate model framework can be built to include two independent ANNs, one only predicting the four highly nonlinear MSHE approach temperatures and the other estimating the remaining 16 outputs. Breaking a surrogate model up into a series of parallel sub-models may increase accuracy in capturing specific nonlinearities that are particularly prominent in a complex process flowsheet. Hence, a second surrogate model framework is constructed to further separate the MSHE approach temperature ANN into four separate ANNs, each predicting an individual approach temperature. This will allow the overall surrogate model to gain better accuracy still with regards to the individual approach temperatures, and ensure that a prediction made using the surrogate model remains feasible when validated with the rigorous model.

Indeed, through verification, the second framework (consisting of 5 ANNs) successfully meets all the constraints, hence resulting in feasible solutions (Fig. 2). Nonetheless, the optimal total shaft work identified in the third surrogate model is 160 MW, much higher than the optimal solution (144 MW) identified using the rigorous model. Hence, the surrogate model is still not efficient in terms of process optimization. Overall, the results from the three ANN based surrogate models suggest that due to the high nonlinearity and dimensionality of the underlying process and limited amount of data, it may not be ideal to use ANNs to construct surrogate models to simulate a whole process flowsheet.

![Figure 2: Plot of validated approach temperatures for 4 different surrogate model structures. Infeasible MSHE approach temperatures have values below the dotted red line.](image)

4.2. Results of the KPLS superstructure model

Due to the system having a relatively large number of inputs and outputs, unsupervised learning techniques are adopted to reduce the dimension of the solution space (i.e. reducing the impact of the curse of dimensionality) for the construction and optimization of surrogate models. In addition, after assessing the rigorous optimization problem described in Section 2, it was observed that there exists no constraint on individual shaft works within the refrigeration cycle, thus using the summation of individual shaft works (i.e. total shaft work, one output) to substitute for the 8 individual shaft works (8 outputs) can potentially simplify the surrogate model construction. If the individual shaft works
did need to be accessed after gaining a solution from the surrogate model with a reduced output such as this, the solution can be directly entered into the rigorous model to gain the relevant process information. Likewise, it was also observed from the ANN based surrogate models that the majority of data regarding 3 out of the 4 vapour fractions and 3 out of the 4 shaft works exist in feasible solution space all the time (i.e. redundant constraints). Due to this fact, it is decided to neglect to model these outputs in the first place to further reduce the dimension of the outputs. From Fig. 2, it is seen that the KPLS superstructure model (consisting of 7 GPs) satisfies all the practical constraints. Furthermore, its predicted total shaft work is verified to be accurate (147.4 MW verified by the Aspen model), and is comparable to the optimal result obtained using the rigorous optimization (144.4 MW). In addition, individual shaft works are found to be feasible after verification. Most importantly, the full time to train and optimize this surrogate model only takes around 10 minutes (i.e. 2 minutes for model construction and 8 minutes for stochastic optimization) as opposed to 17 hours spent when optimizing the rigorous model. This directly demonstrates the superiority and practical advantages of using the integrated KPLS modelling strategy for the optimization and real-time decision-making of high dimensional complex chemical systems (e.g. entire process flowsheet).

5. Conclusions

To conclude, Gaussian processes and artificial neural networks can be used as building blocks to construct efficient surrogate models to represent highly nonlinear systems such as an entire chemical process flowsheet. However, selection of surrogate model structure, fidelity of available data, and amount of data can greatly affect the accuracy and efficiency of surrogate models. Moreover, unsupervised learning techniques can also be taken advantage of to reduce the high dimensionality encountered in large scale processes simulation. Through the use of data-driven models and efficient stochastic optimization algorithms, it is possible to greatly reduce the computational time cost and meanwhile identify a high quality optimal solution for the operation of large scale systems.

References