Accelerating Steam Cracking Simulations with Surrogate-Assisted Parameter Estimation

Qiming Zhao,a,b Dong Qiu,a,b Kexin Bi,c Tong Qiua,b\*

aDepartment of Chemical Engineering, Tsinghua University, Beijing 100084, China

bBeijing Key Laboratory of Industrial Big Data System and Application, Tsinghua University, Beijing, China

cSchool of Chemical Engineering, Sichuan University, Sichuan 610065, China

ilmaple@163.com

Abstract

The paper presents a method that accelerates steam cracking simulations by employing machine learning surrogates for parameter estimation. This approach utilizes Latin hypercube sampling to create initial simulation datasets. It applies Bayesian ridge regression to kernel approximations of the data, enabling efficient prediction of simulation parameters with measurable uncertainty. Compared to traditional iterative techniques, this method speeds up simulations by reducing the average number of iterations by 4.33. Data-driven models using pre-computed simulation data have been shown to enhance the efficiency of simulators based on first principles significantly.

**Keywords**: Steam Cracking, Parameter Estimation, Surrogate Modeling

* 1. Introduction

Steam cracking is the primary process for producing light olefins. Research in this field depends on process simulation models, including rigorous and data-driven ones, to enable yield optimization by adjusting operating conditions. Rigorous models use equations based on first principles, allowing for broad applicability. On the other hand, data-driven models operate as black boxes capable of approximating input-output relationships but have constrained predictive power during extrapolation (Zhao et al., 2023). It is expected to use data-driven techniques to enhance rigorous simulations, which can speed up calculations or improve model fidelity.

When solving complex simulations numerically, iterative methods help manage the challenge of measuring specific parameters. Sequential iterations can be time-consuming, but choosing initial values wisely can accelerate convergence. Due to the nonlinear and dynamic nature of real-world processes, algorithms generally rely on heuristics for parameter initialization, which are customized for particular operating conditions and model frameworks. An alternative approach is using machine learning to provide better initial values of parameters. This can be done by constructing a surrogate model using a dataset of offline simulations that were pre-computed in parallel by the rigorous model.

This paper presents a surrogate-assisted approach to parameter estimation, which uses a regression model to guide the selection of parameters. A simulation database is built from parallel pre-computation of simulations. Using a Bayesian ridge regression model allows for estimating parameters with quantifiable uncertainties, thereby reducing the iterations required to attain similar levels of precision. Measuring uncertainty enables users to set simulation search limits informed by probabilistic understanding.

* 1. Methods

Steam cracking process efficiency is impacted by feedstock quality and furnace performance. Critical parameters like feedstock and steam flow rates, temperature, and pressure must be monitored to evaluate industrial production efficiency (Fakhroleslam et al., 2020). We developed a first-principles simulator for the radiation section of an industrial steam cracker that processes light hydrocarbons. The simulator only considers axial changes in temperature and pressure, resulting in simplified one-dimensional profiles. Once the reaction network and furnace settings are established, the simulator takes input such as feedstock compositions and operating conditions.

Integral calculations are made during simulation to achieve the desired COT and COP. Two parameters, dTg and VPR, are adjusted at each iteration. dTg measures fuel adjustments needed within furnace control systems to meet COT targets, while VPR reflects pressure adjustments for desired COP. When started from fixed parameters, a simulation needs multiple sequential iterations to reach an acceptable error threshold.

The proposed surrogate-assisted parameter estimation approach is applied to the simulator mentioned above, with feedstocks assumed to contain only ethane, propane, and n-butane. It aims to approximate the relationship between input variables, estimated parameters, and convergence loss. The approach consists of two phases: surrogate construction and surrogate-assisted simulation, as shown in Figure 1.



Figure 1. Flowchart of the proposed framework

* + 1. Surrogate construction

This phase involves sampling the input space, conducting simulations on these samples to create a database, and developing a surrogate model for parameter estimation.

Variable ranges of concern in industrial production are identified using production datasets and furnace design specifications. Given the vastness of the resulting input space, simple random or grid sampling may introduce bias and patterns. To address this, Latin hypercube sampling (LHS), a traditional technique for space-filling experimental designs, is used for sampling. LHS ensures adequate coverage of both the input space and subspaces (Liu et al., 2018), producing representative samples. This leads to improved performance in industrial settings where few variables change simultaneously. The method is also sample-efficient, making it ideal for use with costly simulators.

Figure 2 illustrates the process of collecting and modeling simulated data.



Figure 2. A schematic diagram of the simulator and the regression model

Initial simulations begin using random parameters. The diagram above displays the scenario with 400 samples undergoing 16 parallel stochastic parameter searches. These results are the basis of the simulation database. To ensure the database captures scenarios near convergence, an extra simulation is conducted for each sample using parameters derived from a simple linear interpolation of losses from the previous searches.

The regression model uses the Nystroem method for kernel map approximation combined with Bayesian ridge regression to balance accuracy and efficiency. Since linear models are inadequate for this complex system, kernel tricks are used to capture the nonlinear traits of the data, where the idea is to map data into a high-dimensional feature space to make it easier to perform linear classification or regression. However, working in such high-dimensional spaces can be computationally intensive, especially for large datasets. Gaussian processes (GPs), which use covariance functions as kernels, do not scale well to large datasets. Training GPs also require hyperparameter tuning, which complicates real-time model updates.

The Nystroem method produces a low-rank approximation of the original kernel matrix using a strategically selected subset of the training data, typically using a radial basis function (RBF) kernel (Martín et al., 2019). It is essential to tune the number of components in the feature map. Meanwhile, Bayesian ridge regression enhances the conventional ridge regression by integrating it into a Bayesian framework and introducing Gaussian priors on the coefficients for improved regularization. This method provides point estimates similar to standard linear regression and measures uncertainty about these estimates via the posterior distributions of random variables (Shi et al., 2016).

The surrogate model estimates parameters based on simulation inputs and convergence losses. It is trained using data that includes actual loss values, but it treats convergence losses as zero during prediction. This approach enables model training with data from simulations that have not fully converged.

* + 1. Surrogate-assisted simulation

After constructing the surrogate for parameter estimation, simulations can be assisted by its prediction. When the end user enters simulation settings, the surrogate model predicts average values and standard deviations ($σ$) for those parameters.

A surrogate model predicts low $σ$ values for inputs from well-sampled areas, typically requiring a single simulation to converge. Should a stricter loss tolerance be required, the software may employ iterative or parallel refinement to enhance the parameter further.

A commonly used threshold for high variance is the $3σ$ rule, which compares $3σ$ with physical quantities with the same dimension, like the error tolerance for COT. When high $σ$ values are encountered, random searches for parameters are required, or users may broaden resampling by defining new ranges for input variables. Newly simulated samples are appended to the database and utilized in regression analysis to update the surrogate model, resulting in improved parameter estimations. This approach allows computational efforts to be efficiently applied across different simulation scenarios.

* 1. Results and discussion
		1. Surrogate construction

Input variables were sampled within the bounds specified in Table 1.

Table 1. Sampled input variables

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Description | Lower bound | Upper bound |
| Propane | Propane mass fraction | 0.1 | 0.2 |
| n-Butane | n-Butane mass fraction | 0.1 | 0.2 |
| (Ethane) | Ethane mass fraction | 0.6 | 0.8 |
| FFR | Feed flow rate (t/h) | 9 | 13 |
| COT | Coil output temperature (°C) | 860 | 875 |
| CIP | Coil input pressure (MPa) | 0.23 | 0.37 |

The decisions and variations in industrial production determined the boundaries shown above. Ethane mass fractions were calculated from the remaining feed compositions. The LHS method generated 400 input samples, and each sample led to 16 random parameter searches in parallel, where dTg ranges from -25 K to 25 K, and VPR values are between 0.55 and 0.85. Subsequently, an additional 400 simulations were conducted using linearly interpolated parameter values. The training set gathered 6,800 instances in total.

The training data were first normalized to reduce numerical artifacts. Subsequently, the Nystroem method was used to project the inputs to 1,000 components. The regression model then correlated these features to two targets, dTg and VPR. Table 2 displays the model metrics, including the root-mean-square error (RMSE) and coefficient of determination ($R^{2}$) for both the training and test sets.

Table 2. Regression model test metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Target | RMSE (train) | $R^{2}$ (train) | RMSE (test) | $R^{2}$ (test) |
| dTg (K) | 0.291 | 0.9997 | 0.521 | 0.9990 |
| VPR | 0.00142 | 0.9998 | 0.00280 | 0.9992 |

The model fits the simulated data well without obvious overfitting. Considering the simulator’s convergence tolerance, the estimated parameters are satisfactory, with predicted temperature deviations near 0.5K and pressure deviations under 1 kPa.

* + 1. Surrogate-assisted simulation

An additional 100 samples were LHS-generated as the test set, mimicking end-user inputs in the same region of furnace operation. The surrogate model predicted parameters corresponding to the converged simulations for these samples, and simulations are carried out using these parameters. Meanwhile, stochastic parameter searches were also tested for comparison. Figure 3 illustrates the predictive capabilities of the surrogate-assisted simulation using two evaluation methods: 1) reduced loss compared with random searches and 2) reduced iterations compared with a sequential iterative algorithm.



Figure 3. Surrogate model evaluation

Figures 3(a) and 3(b) illustrate the improvements made in parameter estimation of temperature and pressure. The surrogate model generated predictions for each sample, verified using the original simulator (shown on the x-axis). The results were compared to a random parameter search (shown on the y-axis), where the best outcome (minimum loss) was selected for each sample. Random search losses demonstrated a mean shift and high variance, and pressure data also took a very skewed distribution. On the contrary, surrogate-assisted simulations contributed to average losses that were an order of magnitude lower than the best outcomes achievable through random searches. The results were still slightly skewed, probably due to biased training datasets generated by stochastic searches. Increasing the sample size may reduce bias and the long tail in distribution. The surrogate model consistently outperformed the random search by reducing computational resource usage while maintaining accuracy.

Regarding uncertainties, the average simulated COT loss was consistent with the model’s predicted standard deviations of $dTg$. In the 100 surrogate-assisted simulations, the mean COT loss ($μ\_{loss\\_T}=0.96$) was almost identical to the mean predicted standard deviation ($\overbar{σ}\_{dTg}=0.97$). The rationale remains the same for both parameters, but VPR must be multiplied by CIP to restore the unit during pressure variance calculations.

Figure 3(c) demonstrates the number of iterations required to reach the same or smaller absolute loss for dTg and VPR as the simulator starts from a fixed parameter set (dTg=0 and VPR=0.7). Most samples needed four or more iterations, and some even reached 8~10. The mean iteration times were 5.33, indicating that the surrogate model can reduce simulation time by more than 4 iterations on average. This is valuable when a single iteration takes several minutes.

* 1. Conclusions

This study presents a novel approach that uses surrogate assistance for parameter estimation to expedite steam cracking simulations, primarily to address the computational demands of first-principles simulators. The proposed method constructs a surrogate model by exploring the input space with Latin hypercube sampling, followed by the Nystroem approximation for dealing with nonlinearity and Bayesian ridge regression for parameter estimation with uncertainty measurements.

By integrating Bayesian regression into the model, precise parameter estimation with quantified uncertainties can be achieved. The results demonstrated the surrogate model’s capability to outperform traditional random parameter searches. Convergence losses were an order lower than multiple random searches, and more than 4 iterations were saved when achieving the same error tolerance. As evidenced in surrogate-assisted simulations, the model’s ability to accurately predict parameters with minimal loss confirms its potential as a powerful tool in steam cracking process optimization.

A potential next step involves developing a systematic and flexible method for users to refine simulation data sampling and surrogate-assisted parameter estimation. For instance, the surrogate model could propose various candidate parameters for simultaneous simulations.

The surrogate-based technique offers a tangible solution to enhance simulation efficiency and scalability, which can be extended to other simulations where accurate parameters are crucial for calculation efficiency and outcome reliability.

References

M. Fakhroleslam, & S. M. Sadrameli, 2020, Thermal Cracking of Hydrocarbons for the Production of Light Olefins; A Review on Optimal Process Design, Operation, and Control, Industrial & Engineering Chemistry Research, 59, 27, 12288–12303.

H. Liu, Y. Ong, & J. Cai, 2018, A survey of adaptive sampling for global metamodeling in support of simulation-based complex engineering design, Structural and Multidisciplinary Optimization, 57, 1, 393–416.

M.L. Martín, B. Carro, A.J. ánchez-Esguevillas, & J. Lloret, 2019, Shallow neural network with kernel approximation for prediction problems in highly demanding data networks, Expert Systems with Applications, 124, 196-208.

Q. Shi, M.A. Abdel-Aty, & J. Lee, 2016, A Bayesian ridge regression analysis of congestion’s impact on urban expressway safety, Accident Analysis & Prevention, 88, 124–137.

Q. Zhao, K. Bi, & T. Qiu, 2023, Data-driven intelligent modeling framework for the steam cracking process, Chinese Journal of Chemical Engineering, 61, 237–247.