Published in final edited form as:

AIChE J. 2020 October ; 66(10): . doi:10.1002/aic.16657.

A data-driven optimization algorithm for differential algebraic equations with numerical infeasibilities

Burcu Beykal^{a,b}, Melis Onel^{a,b}, Onur Onel^{a,b,c}, Efstratios N. Pistikopoulos^{*,a,b}

^aArtie McFerrin Department of Chemical Engineering, Texas A&M University, College Station, TX 77843, USA

^bTexas A&M Energy Institute, Texas A&M University, College Station, TX 77843, USA

^cDepartment of Chemical and Biological Engineering, Princeton University, New Jersey, NJ 08544, USA

Abstract

Support Vector Machines (SVMs) based optimization framework is presented for the data-driven optimization of numerically infeasible Differential Algebraic Equations (DAEs) without the full discretization of the underlying first-principles model. By formulating the stability constraint of the numerical integration of a DAE system as a supervised classification problem, we are able to demonstrate that SVMs can accurately map the boundary of numerical infeasibility. The necessity of this data-driven approach is demonstrated on a 2-dimensional motivating example, where highly accurate SVM models are trained, validated, and tested using the data collected from the numerical integration of DAEs. Furthermore, this methodology is extended and tested for a multi-dimensional case study from reaction engineering (i.e., thermal cracking of natural gas liquids). The data-driven optimization of this complex case study is explored through integrating the SVM models with a constrained global grey-box optimization algorithm, namely the ARGONAUT framework.

Keywords

Data-Driven Optimization; Differential Algebraic Equations; Support Vector Machines; Dynamic Optimization; Steam Cracking

Introduction

The system of differential algebraic equations (DAEs) is ubiquitous in mathematical modeling of chemical engineering systems, as many first-principles models include differential equations like mass, energy, momentum balances along with process constraints, such as physical properties and rate laws. DAE systems are commonly observed in the areas of process control, as well as chemical reactions and reactor design^{1,2}.

*Corresponding Author; Tel: +1 (979) 458-0259; stratos@tamu.edu.

The mathematical optimization of such systems is challenging since the direct implementation of deterministic optimization methods is prohibitive. Hence, many dynamic programming problems in the aforementioned application areas utilize commercial software like the gPROMS environment and Aspen Custom Modeler for first-principles modeling of DAE systems and their respective dynamic optimization via sequential approaches³⁻⁵. Alternatively, simultaneous approaches such as the full discretization of the DAE system and its incorporation into a nonlinear programming (NLP) formulation using orthogonal collocation on finite elements is also preferred for making DAEs amenable for optimization, specifically for unstable and ill-conditioned problems^{6;5}. For example, Caballero et al.⁷ investigated the optimization of ethylene production through a one-dimensional plug-flow model at steady-state conditions with heat flux along the reactor length to be the only decision variable. In the problem formulation, the equality constraints governing the rate, mass, energy, and momentum balance equations were expressed with stiff nonlinear DAEs, which inhibits the global optimization via direct deterministic methods. Hence, the authors implemented orthogonal collocation on finite elements to spatially discretize the DAEs into a set of nonlinear equality constraints, while solving the resulting large-scale NLP problem to local optimality. In another study by Onel⁸, the dynamic optimization of steam cracking of ethane, as well as the cracking of propane and butane with reactor coking considerations were investigated in detail. Similar to the aforementioned study, orthogonal collocation on finite elements was implemented to discretize the stiff DAEs and the resulting model was solved to local optimality using a multi-start approach to generate high-quality solutions. In addition, the reactor length was modeled using binary variables, and the optimal length that maximizes the ethylene yield was also investigated by Onel⁸.

A third alternative for dynamic optimization of the system of DAEs can be through the utilization of data-driven approaches and novel machine learning algorithms. The idea of representing highly complex engineering processes with simple tractable models using data (i.e., surrogate models) has gained accelerated attention in the last decade⁹. Although surrogate models were primarily used as means of replacing detailed unit operations in flowsheet synthesis¹⁰⁻¹², their application has also been expanding in different areas of process systems engineering including but not limited to dimensionality reduction in control¹³, black/grey-box optimization^{14;15}, multi-objective optimization¹⁶, and bi-level programming^{17;18}. In this work, we use surrogate models: (1) for modeling the numerical feasibility boundary in multi-dimensional DAEs such that the sampling points that may cause failures in the problem simulator are rapidly eliminated from the analysis with high accuracy; and, (2) for the model-based global constrained grey-box optimization of the DAE problem without the full discretization of the governing equations. Specifically, our framework addresses the sampling challenge under unknown interdependencies between decision variables (i.e., implicit constraints) using Support Vector Machines (SVMs). Furthermore, we demonstrate the applicability of our approach by integrating the SVM-based constraint handling scheme to a model-based data-driven global constrained optimization algorithm, namely the ARGONAUT framework^{19;20;14}, and perform a computational study of optimal reactor design and operation in thermal cracking of natural gas liquids.

This manuscript is organized as follows. First, we define the problem of interest and elaborate on the challenges with data-driven optimization in the presence of DAEs or Ordinary Differential Equations (ODEs) with numerical infeasibilities where the stated challenges are demonstrated on a motivating example. Next, we introduce the SVM-based filtering methodology and its implementation to a data-driven algorithm. Finally, the algorithm is tested on a steam cracking model for ethylene and propylene production, and the results for computational experiments are provided along with the concluding remarks.

Problem Definition

In this work, we are interested in solving a dynamic optimization problem of the following form using a data-driven methodology,

$$\begin{aligned}
 & \min_{\mathbf{u}, \mathbf{x}_0, t_f} \Phi(\mathbf{x}(t_f), \mathbf{y}(t_f), \mathbf{u}(t_f), \mathbf{p}) \\
 & \text{s.t. } \dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{p}) \\
 & \quad \mathbf{g}(t, \mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{p}) = \mathbf{0} \\
 & \quad \mathbf{h}(t_f, \mathbf{x}, \mathbf{y}, \mathbf{u}, \mathbf{p}) \leq \mathbf{0} \\
 & \quad \mathbf{x}(0) = \mathbf{x}_0, t \in [t_0, t_f], \mathbf{u} \in [\mathbf{u}^{lo}, \mathbf{u}^{up}], \mathbf{x}_0 \in [\mathbf{x}_0^{lo}, \mathbf{x}_0^{up}], t_f \in [t_f^{lo}, t_f^{up}]
 \end{aligned} \tag{1}$$

where $\mathbf{p}$ are the parameters, $\mathbf{x}$ are the differential variables with initial conditions $\mathbf{x}_0$, $\mathbf{y}$ are the algebraic variables, $\mathbf{u}$ are the manipulated variables, Φ is the objective function at final time t_f , $\mathbf{g}(\cdot)$ are the algebraic equations, $\mathbf{f}(\cdot)$ are the differential equations that define the dynamic system, and $\mathbf{h}(\cdot)$ are the process constraints that are required to be met at the final time point. The goal is to find the optimal manipulated actions $\mathbf{u}^*$ over the optimal time horizon t_f^* (or spatial horizon, if the problem is dynamic in spatial coordinates), as well as the optimal initial conditions $\mathbf{x}_0^*$ that will yield a successful solution of the differential algebraic equations, while minimizing the objective function value without path constraints. We assume that the problem given in Equation 1 is stable (i.e., the truncation error is satisfactorily small) but subject to one or more sources of numerical infeasibility (i.e., stiffness, ill-conditioned algebraic equations, undefined solutions) within the provided bounds, where the step size taken by the numerical integrator (either by an explicit or an implicit method) is forced to be extremely small to meet the desired solution accuracy²¹. Essentially, in the problems we are considering, if the initial conditions and the manipulated variables for the underlying DAE system are not chosen carefully, the solution could go beyond the stability region of the numerical method²¹ causing a failure in the problem simulator. In fact, the importance of handling numerical infeasibilities in process simulation and optimization has been historically noted by Allgor²² with implementations in ASCEND^{23;24} and ABACUSS²⁵. Hence, our goal is to utilize data-driven analysis to avoid these unstable regions and to guide data-driven optimization algorithms towards seeking viable and numerically feasible decision variables. Next, we elaborate on the challenges with solving such dynamic programming problems using a data-driven optimization methodology.

Challenges in Design of Experiments with Numerically Infeasible Ordinary Differential Equations

Data-driven modeling and derivative-free optimization rely on different sampling strategies that provide an initial plan for the controlled experiments on problem simulators, which is commonly known as the Design of Experiments (DoE). The goal of DoE is to provide possible candidate locations for the input variables within the pre-defined box-constraints such that these experiments capture a variety of system dynamics. There are many different ways of constructing this initial set of candidate points including Latin Hypercube (LHD) and full factorial designs. The details on different types of DoE and the current developments in DoE research are discussed in a recent review article by Garud et al.²⁶, as well as in a notable textbook by Cavazzuti²⁷.

It is important to note that the DoE is a statistical procedure and not guided by the physical information that entails an engineering process. These sampling plans assume that the decision variables are purely independent and box-constrained or any interdependency between them is explicitly known²⁰. However, in realistic cases and/or ill-conditioned problems of DAE and ODE systems, unknown interdependencies between the decision variables (i.e., the initial conditions of the DAEs/ODEs) define the feasibility of the final solution. Specifically, in these cases, a slight variation in the initial conditions causes the solution to change dramatically (i.e., stiff problems) or even tend to infinity (i.e., blow-up solution) at certain finite values of the independent variables in a bounded interval. Thus, a subset of candidate initial points generated by the DoE may result in unphysical and/or undesirable outcomes, such as an early termination of the problem simulator due to numerical failures where extremely small steps must be taken by a stable initial value problem solver to achieve satisfactory outputs.

This distinction between failed and successful simulations of candidate initial points indicates that the decision variables are not purely independent, and at least one constraint should exist where the explicit analytical formulation of this is unknown to the user. As a result, it is of critical importance for any data-driven optimization methodology to handle such implicit constraints effectively at the sampling stage, as the feasibility and numerical stability of the solution and the efficiency of the data-driven optimization algorithm strictly depend on this constraint. Consider the following initial value problem as a motivating example,

$$\dot{y} = y^2, \quad y(0) = y_o > 0 \quad (2)$$

The analytical solution of this separable ODE is given in the following form,

$$y(t) = \frac{y_o}{1 - y_o \cdot t}, \quad t < \frac{1}{y_o} \quad (3)$$

It is important to note that, the validity and the existence of the solution in Equation 3 strictly depend on the condition between the time and the initial condition value.

Specifically, at $t = 1/y_o$, the denominator will become zero and the solution will be undefined. If we wanted to explore the box-constrained space defined by t and y_o using DoE, the samples that violate this constraint are going to be removed *a priori* to sample collection, which will also prevent us from sampling in regions that won't yield a feasible or a numerically stable solution for the problem of interest. This is demonstrated in Figure 1.

Although in this motivating example it is rather easy to derive the constraint for a valid integration solution concerning the time horizon and the initial condition, in many complex engineering problems (i.e., reaction engineering), the analytical solution may not be trivial or may not even exist. Furthermore, in multi-dimensional problems where multiple variables are initialized for solving a system of ODEs, it is more challenging to postulate appropriate explicit constraints for the underlying relationships between the initial conditions. Hence, optimizing a black-box simulator with a system of multi-dimensional DAEs/ODEs that exert numerical infeasibility or contain implicit constraints (i.e., constraints that do not have an explicit mathematical or an analytical form) using a data-driven methodology is a challenging task.

Several approaches can be explored such as sampling in smaller regions or removing infeasible samples *a posteriori* to the simulator call. A smaller sampling region can be imposed such that all sampling points are feasible. For example, if we were to set the upper bound to be 0.5 for both t and y_o in the motivating example, all points collected within this new box will be feasible (Figure 1). However, in an optimization context, tighter bounds will yield a conservative decision variable space, where the global solution may lie outside these newly imposed bounds. Furthermore, *a posteriori* analysis on the input-output data can be computationally demanding since this will require all candidate points to be evaluated through the problem simulator. This is undesirable in many high-fidelity problems given that as the number of ODEs and the problem complexity increases, the numerical integration will become more time-consuming. Especially, evaluation of the candidate initial points that cause numerical failures can take more than a couple of hundred seconds before a stable initial value problem solver quits without meeting the integration precision. For example, assuming a failed simulation takes 200 s per sample to evaluate, collecting the output of 150 infeasible candidate points will take 30,000 s, where these points will not be viable for identifying the optimal solution.

Hence, this sampling challenge with numerically infeasible DAEs requires a systematic approach, where a wider range of system dynamics should be captured in a computationally efficient way. To this end, we are using a supervised machine learning methodology, namely the Support Vector Machines (SVMs), to assess the feasibility of a given combination of initial conditions with respect to the numerical integration at the several stages of a data-driven optimization process (i.e., initial sampling, and re-sampling) *a priori* to the simulator call. Previously, the idea of using SVMs to approximate the feasible region of optimization problems was explored in bi-level and mixed-integer programming problems^{28;29}. The study by Dias and Ierapetritou²⁸ showed that among different machine learning approaches for approximating the feasible region of a bi-level integrated planning and scheduling problem, neural nets had the lowest prediction errors followed by SVMs and tree-based modeling. For our applications, the computational expense of a single simulation can be costly. As the

computational demand for training a neural net is much higher in terms of sampling and parameter tuning (i.e., deciding on the optimal neural net structure) compared to an SVM approach, we are using SVMs to handle the numerical infeasibilities in a multi-dimensional system of DAEs such that they are amenable for data-driven modeling and optimization without the full discretization of the underlying first-principles model. Specifically, the SVMs are used to derive an implicit function that mimics the stability constraint for the solution of numerically infeasible DAEs in multi-dimensional space, rather than approximating the full feasible space of the problem as done in the aforementioned studies. Through the use of this supervised machine learning approach, the nonlinear dependencies between the initial conditions, manipulated actions, and the operating horizon that strictly define the feasibility and the existence of the DAE solution are captured with high accuracy. The details of our approach are further explained in the following section.

Modeling Implicit Constraints with Support Vector Machines

In machine learning, SVMs are extensively used for classification and regression-type of analyses, spanning over several different application areas including but not limited to fault detection and diagnosis³⁰⁻³², improvement of process operations³³, and predictive modeling of complex substances^{34;35}. In this work, we use an SVM model to mimic the implicit constraint that defines the feasibility of the solution of a DAE system. Specifically, we build an SVM-based classification model in the offline phase by using a dataset of simulated samples with their binary outcome (feasible/infeasible). The obtained classification model acts as a filter and guides the sampling strategy of a data-driven optimizer such that the infeasible combinations of decision variables that lead to numerical failures are eliminated *a priori* to sample collection.

If we now consider the initial value problem in Equation 2 from a data-driven perspective while assuming no knowledge of the constraint in Equation 3 that characterizes the validity of the solution, we can numerically integrate Equation 2 for every combination of t and y_o values provided by the DoE. At the end of each simulation, we can check whether the integration has failed or not and assign a label, “0” for “feasible” or a valid integration solution and “1” for “infeasible” or a failed integration solution. The resulting continuous input and the discrete output information can be used to formulate a nonlinear two-class classification problem using SVM, where this model will provide a decision boundary between feasible and infeasible combinations of t and y_o .

Essentially, as shown in Figure 2, if the SVM classifier is properly trained, validated, and tested, this separating nonlinear boundary will be the same as the constraint imposed on the input variables shown in Figure 1. As new samples are desired to be collected in this decision space, the SVM model can now be used to classify and filter the incoming combinations of input variables based on their probabilistic feasibility information provided by this model. This filtering step is essentially a function call that has minimal computational expense to execute and will allow us to remove the infeasible combinations *a priori* to the problem simulator call, improving the stability and the computational speed of the data-driven optimization process. In the next section, we describe our generalized

framework for using and implementing SVM classifiers in data-driven optimization of numerically infeasible multi-dimensional DAE systems.

Data-Driven Optimization Framework for Numerically Infeasible Multi-Dimensional DAE Systems

The outline of our generalized framework for handling implicit constraints in data-driven optimization is provided in Figure 3. In phase 1, which is the offline phase of the framework, sampling is performed within the lower and upper bounds of the decision variables of a given optimization problem with a numerically infeasible DAE system. For each sampling point, a respective output class information (feasible/infeasible) is collected as described in the previous section. Using this continuous input and categorical output dataset, a nonlinear two-class classification problem is formulated and an SVM model is tuned for an accurate representation of the stability constraint of a multi-dimensional numerically infeasible DAE system. In phase 2, this trained, validated, and tested SVM model is implemented to a grey-box optimization solver. In this phase, the SVM model filters the combination of input variables that yield numerically unstable solutions online, as the grey-box optimization algorithm is executed. Here, we demonstrate a generic two-phase approach that can be used to integrate the SVM model to any grey-box optimization solver. For this purpose, we chose a model-based global constrained grey-box optimization algorithm, the ARGONAUT framework, to explore high-quality solutions with constraint handling capabilities and minimum sampling requirements^{20;19;14}. Further details of each phase are provided in the Supplementary Material.

Our methodology to handle implicit constraints through the use of SVM models is incorporated at every stage of the grey-box optimization solver, where procedures regarding sampling are taken care of, including when the initial design is created, the existing design is augmented and when new samples are adaptively collected from the optimization step at every iteration. In the next section, we describe a more complicated computational case study, namely the steam cracking of ethane and propane, where the data-driven modeling and optimization of these numerically infeasible DAE systems are explored using our SVM approach.

Data-Driven Dynamic Steam Cracking Optimization for Ethylene and Propylene Production

The rapid increase in shale gas production in the Appalachian and Permian Basins for the last decade has led to significant growth in natural gas liquids (NGLs) production, as well as a projected increase for these petrochemical feedstocks in the upcoming years³⁶⁻³⁸. Ethane and propane, being the major constituents of NGLs, are predominantly used for the production of ethylene and propylene, respectively, where ethylene consumption is expected to increase by 49% from the year 2017 to 2020³⁷. Naturally, the growing petrochemical feedstock supply and the rising demand for light olefins sparks an interest in converting NGLs to olefins via the non-catalytic steam cracking process. In this perspective, many existing ethylene crackers have expanded capacity and new crackers are coming online to

benefit from this unique opportunity³⁹. Hence, the mathematical optimization of this process emerges as a necessity to determine the optimal operating conditions for the steam cracker, in such a way that the profit from ethylene and propylene production is maximized.

To this end, we are using the integrated SVM-ARGONAUT framework to handle the numerical infeasibilities in the cracking model equations while exploring high-quality solutions for the reactor length, inlet ethane/propane and steam flowrates, inlet temperature, inlet pressure, and heat flux profile along the reactor length, through surrogate modeling and optimization. The steam cracker reactor model for ethylene and propylene production is adapted from Onel⁸ and modeled as a one-dimensional plug flow reactor with coking effects (Figure S2). The detailed reactor model equations and parameters are presented in the Supplementary Material.

The mathematical formulations presented in the Supplementary Material for the steam cracking case studies are modeled in MATLAB and this problem simulator is used for the data-driven modeling and optimization of the thermal cracking process with SVM-ARGONAUT. For both case studies, we consider 10 input variables (the decision variables, Table S7), 1 known constraint (Equation S23 for ethane cracking model, Equation S24 for propane cracking model), and 4 grey-box constraints (Equations S19, S20, 4, 5), where our objective is to maximize the profitability of the operation (Equation S21 for ethane, S22 for propane). The detailed analysis of the results is provided in the following section.

Results of Computational Studies

The computational studies for the data-driven steam cracking optimization are performed on a High-Performance Computing (HPC) machine at Texas A&M High-Performance Research Computing facility (Ada HPC Cluster operated with Linux CentOS 6: Intel Xeon E5-2670 v2 10-core processor (Ivy Bridge-EP)). The supercomputer is used in both stages of the framework: (1) In the offline phase, for data collection and SVM model building; (2) In the online phase, for executing the ARGONAUT algorithm as a parallel job, using 1 node (20 cores per node with 64 GB RAM) on the supercomputer. Likewise, the data collection and SVM model-building phases are performed as a parallel job, using 1 node (20 cores per node with 64 GB RAM and 1 node (4 cores per node with 64 GB RAM), respectively.

Offline Phase: Results of SVM Model Building

As a first step, the SVM model is built using the data generated from the steam cracking model which is subject to the known constraint (Equation S23 for ethane cracking model, Equation S24 for propane cracking model) and the provided bounds on the decision variables (Table S7), following the methodology described in the Supplementary Material. As the bounds on the decision variables are the same in the first session of ARGONAUT runs, only 1 SVM model is built per case study. The model evaluation metrics on the test data for both ethane and propane cracking case studies are provided in Table 1.

The testing results show that SVM models for these datasets can be generated to high accuracy (98.492 %). It is important to note that the SVM model cannot guarantee feasibility to the selected sampling points as the model may misclassify % 1.508 of the samples as

“feasible” based on the testing dataset. When this model is implemented in the online phase, it may cause ARGONAUT to converge to a numerically infeasible solution at the end of its iterations. To guarantee the feasibility of the solution, we add an extra grey-box constraint in the online phase that essentially confirms the assigned value for the length variable as a candidate point is equal to the simulation endpoint (i.e., simulation does not quit prematurely without reaching the endpoint of length). As ARGONAUT cannot directly handle equality constraints, this constraint is reformulated into two inequalities (Equations 4 and 5) with an added relaxation parameter.

$$L^{in} - L^{out} \leq 0.000001 \quad (4)$$

$$L^{in} - L^{out} \geq 0.000001 \quad (5)$$

Once the SVM model for the first session is established, the online phase of the SVM-ARGONAUT integrated approach is executed 20 times for each cracking model, each starting with a random LHD. After the convergence is reached, the first session is completed and the variable bounds are tightened, new SVM models are generated for each run such that they represent the numerical feasibility of the cracking models in the reduced space. The detailed testing performances of these SVM models under the tightened bounds are summarized in the Supplementary Material (Tables S8 and S9). The overall results show that these models also have very high accuracy, with more than 97% and 95% correct classification performance among all tested samples for the ethane and propane cracking case studies, respectively. The other performance metrics are also satisfactory, where the model precision is greater than or equal to 98% and 94%, the recall is greater than 98% and 95%, the AUC is greater than 99% and 99%, and the F_1 score is greater than 98% and 96% for ethane and propane case studies, respectively.

Online Phase: Results of the Grey-Box Optimization

In the online phase, our goal is to find the optimal solution to the steam cracking problem using our integrated approach for the implicit modeling of the DAE solution stability constraint and the constrained grey-box optimization of the problem of interest. For all case studies and their respective 20 repetitive runs, the number of initial sampling points is set to be $N_{sample} = 30 \times N_{dim} + 1$. Here, we collect a larger number of initial sampling points compared to the default value to capture a variety of system dynamics of a very challenging optimization problem while balancing out the computational expense required to execute these samples. The same rule-of-thumb is also used when performing sample reduction via OSCAR⁴⁰ and in the second session of the algorithm, when the LHD is augmented for exploring the most promising region for the optimal solution. For processes regarding the surrogate modeling and grey-box optimization, ARGONAUT is executed in the default mode, where the algorithm decides on the surrogate model form for the objective function and the grey-box constraints.

The thermal cracking models provided in the Supplementary Material are used as grey-box problem simulators, where different combinations of decision variables are input to each

simulator, and the corresponding objective function value and the constraint violations are collected. The input combinations to the problem simulators are first evaluated to satisfy the known constraint and then evaluated by the SVM-feasibility checkpoint to ensure that this combination will yield a numerically feasible solution. If the sampling point passes these two feasibility checks, then that sample is evaluated in the process models and its corresponding outputs are collected and further processed in the parameter estimation and data-driven optimization stages of the algorithm. It is important to state the results reported in this phase are guaranteed feasible, but not guaranteed optimal for the original optimization problem. Although ARGONAUT uses global optimization at two key stages of the algorithm, the approach is based on functional approximations constructed without any prior knowledge on the continuity and/or the smoothness of the original problem. Hence, we report the best-found solution out of the 20 random runs for both case studies, where for the ethane cracking case study this is summarized in Figure 4 and Table 2.

Figure 4A shows that the molar flowrate of the main products is increasing along the reactor length as the desirable reaction is taking place. Clearly, the ethane cracking shows a single-feed-single-product trend where ethylene is produced through the favorable reaction alongside with H_2 . In addition, the molar flowrate of the byproducts is significantly limited compared to the desired products. Figure 4B shows the molar flowrate of the C_4^+ species which were previously identified as the coke precursors. We observe that among these 4 species, 1-butene has the highest flowrate along the reactor length, hence contributing the most to the reactor coking. This observation is also consistent with the findings of Onel⁸, demonstrating the validity of our data-driven approach for finding high-quality feasible solutions for the optimization of numerically infeasible DAEs.

Furthermore, Table 2 summarizes the results pertaining to the optimal decision variables achieved in the ethane cracking reactor using the SVM-ARGONAUT integrated framework. For the thermal cracking of ethane, the SVM-ARGONAUT framework identifies the maximum profit as \$0.3359/s which corresponds to an overall annual profit of \$10.6M. The best-found decision variables show that a greater heat flux supply is required at the first 3/5 portion of the reactor where it later decreases gradually towards the exit of the reactor. This is consistent with the inlet temperature value as the reactor entrance temperature is low, a greater heat flux needs to be supplied to ensure endothermic cracking reactions take place. In addition, a higher ethane flowrate is established where the steam flowrate is relatively lower. This is an expected result; as ethane cracking being a single-feed-single-product system, the only positive contribution to the profit comes from ethylene production. We also observe that the best-found value of the inlet temperature for the ethane cracker is lower than expected, as higher temperatures will increase the reaction rates. However, as high temperature favors faster reactions, the reactor coking will also be enhanced, hence resulting in loss of profit. As a result, the best-found inlet temperature value prevents early reactor coking and promotes a higher profit of operation. This is also supported by a lower steam flowrate for the ethane cracker where a minimal amount of steam will be required at minimal amounts of coking on the reactor wall. Moreover, a shorter reactor length is identified in the optimal configuration compared to a typical longer reactor length that is commonly reported in the literature. This is a key result showing that fine-tuning the reactor length will generate valid high-quality solutions with high-profit values by decreasing the

investment, heating, and decoking costs of bigger reactors. In addition, we observe that the reported best-found solution in Table 2 favors good ethane conversion and ethylene yield with high selectivity for ethylene.

Similarly, the optimal results of the thermal cracking of propane found using SVM-ARGONAUT integration are reported in Figure 5 and Table 3. The results show that the profit obtained from thermal cracking of propane is \$0.0845/s which corresponds to an annual profit of \$2.67M. The reactor molar flowrate profiles in Figure 5A show that the favorable reaction starts taking place early at the reactor entrance, where propane flowrate depletes, and products are produced along the reactor length. Figure 5A also shows that the thermal cracking of propane enables the production of two main products, namely ethylene and propylene. As reported in Table 3, this reactor configuration leads to a high propane conversion value with a larger yield and selectivity favored for ethylene. Hence, the best-found configuration of the propane cracker at maximum profit pushes for a greater ethylene production than propylene. Furthermore, Figure 5B shows the temperature and pressure profiles for the best-found solution of the propane cracker. Although higher inlet temperature and pressure are required for this case study where a slightly longer reactor length is also preferred for maximizing the profit, the solution is feasible with regards to the limits provided for outlet temperature and pressure.

Moreover, the overall profit breakdown in Table 4 shows that for both case studies the petrochemical feedstock costs, reactor heating, and decoking costs take the most out of the profit, whereas the reactor investment cost is relatively small due to the optimized reactor length. In the thermal cracking of propane, we also observe that ethylene production contributes more to the profit than the propylene, as the coking mechanism for this case study utilizes C_3H_6 as the coking precursor. Hence, this limits the propylene production while favoring ethylene production for profit and minimum coke generation on the reactor wall. It is important to note that the reaction mechanism for propane cracking allows flexibility in the mode of operation depending on the market demand or prices (i.e., maximizing ethylene or maximizing propylene production). However, it is important to note that exhaustive exploration of the full Pareto solution between maximizing ethylene versus maximizing propylene is possible, but out of the scope of this work.

Finally, the total elapsed computational time of the online phase with and without the SVM approach is compared. Figure 6 shows that the integrated SVM-based data-driven optimization algorithm is more computationally efficient than in the absence of this approach for both ethane and propane cracking case studies in the online phase. Furthermore, lower profit values are observed for both cracking case studies when the SVM approach is not implemented. For ethane cracking, the best-found profit over 20 random runs without the SVM approach is \$0.2970/s, whereas for propane cracking the best profit is \$0.0834/s. We observe that the integrated approach can locate better solutions where the profit is improved by 13.1% for ethane cracking and 1.3% for the propane cracking problem. In the future, the effect of DAE solver options on numerically infeasible problems can be investigated and the results may further be improved by considering the solver options as additional decision variables. By fine-tuning the DAE solver options using data-driven optimization, an optimal DAE solver setting can be identified for a given problem alongside

the optimal process variables. In fact, in deterministic optimization, Liu et al.⁴¹ have shown that fine-tuning BARON's options using derivative-free optimization did improve the solver's default performance in benchmark problems. Nonetheless, our overall results show that the SVM-based optimization algorithm can find superior feasible solutions to multi-dimensional DAEs with numerical infeasibilities in a computationally efficient way.

Conclusion

In this work, a data-driven optimization algorithm is proposed using Support Vector Machines (SVMs) for systems with numerically infeasible Differential Algebraic Equations (DAEs). The stability constraint of the numerical integration solution for a DAE system is formulated as a nonlinear two-class classification problem, where the boundary of numerical feasibility is implicitly modeled using an SVM model. Later by incorporating SVM models to a global constrained grey-box optimization solver, namely the ARGONAUT framework, we were able to filter and remove any numerically infeasible sampling points *a priori* to simulator call and further explore the optimal solution of the complex process model using a data-driven approach. The fundamental idea behind this integrated approach is demonstrated on a 2-dimensional motivating example where the SVM approximation of the numerical integration solution stability constraint is shown to achieve high testing accuracy. Further, this approach is extended and tested on more challenging case studies, namely the thermal cracking of natural gas liquids. The results from thermal cracking case studies show that an SVM-based approach enables feasible and high-quality solutions for the data-driven optimization of systems with numerically infeasible DAEs without the full discretization of the underlying first-principles process model.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to acknowledge the funding and support provided by the U.S. National Institutes of Health Superfund Research Program (NIH P42-ES027704). Portions of this research were conducted with the advanced computing resources provided by Texas A&M High-Performance Research Computing. The manuscript contents are solely the responsibility of the grantee and do not necessarily represent the official views of the NIH. Further, NIH does not endorse the purchase of any commercial products or services mentioned in the publication.

References

1. Biegler LT. Optimization of differential-algebraic equation systems. <https://www.lehigh.edu/~wes1/apci/26may00.pdf>. 2000.
2. Daoutidis P DAEs in model reduction of chemical processes: An overview In: Ilchmann A, Reis T. Surveys in Differential-Algebraic Equations II. Switzerland: Springer International Publishing, 2015:69–102.
3. Dangelakis NA, Burnak B, Katz J, Pistikopoulos EN. Process design and control optimization: A simultaneous approach by multi-parametric programming. *AICHE Journal*. 2017;63(11):4827–4846.
4. Berreni M, Wang M. Modelling and dynamic optimization of thermal cracking of propane for ethylene manufacturing. *Computers & Chemical Engineering*. 2011;35(12):2876–2885.
5. Lang YD, Biegler LT. A software environment for simultaneous dynamic optimization. *Computers & Chemical Engineering*. 2007;31(8):931–942.

6. Biegler LT. Nonlinear programming: concepts, algorithms, and applications to chemical processes. Philadelphia: Society of Industrial and Applied Mathematics, 2010.
7. Caballero DY, Biegler LT, Guirardello R. Simulation and optimization of the ethane cracking process to produce ethylene. *Computer Aided Chemical Engineering*. 2015;37:917–922.
8. Onel O. Advances in modeling, synthesis, and global optimization of hybrid energy systems toward the production of liquid fuels and olefins. Ph.D. thesis. Princeton University, New Jersey, 2017.
9. Bhosekar A, Ierapetritou M. Advances in surrogate based modeling, feasibility analysis, and optimization: A review. *Computers & Chemical Engineering*. 2018;108:250–267.
10. Caballero JA, Grossmann IE. An algorithm for the use of surrogate models in modular flowsheet optimization. *AIChE Journal*. 2008;54(10):2633–2650.
11. Henao CA, Maravelias CT. Surrogate-based superstructure optimization framework. *AIChE Journal*. 2011;57(5):1216–1232.
12. Fahmi I, Cremaschi S. Process synthesis of biodiesel production plant using artificial neural networks as the surrogate models. *Computers & Chemical Engineering*. 2012;46:105–123.
13. Sorek N, Gildin E, Boukouvala F, Beykal B, Floudas CA. Dimensionality reduction for production optimization using polynomial approximations. *Computational Geosciences*. 2017;21(2):247–266.
14. Beykal B, Boukouvala F, Floudas CA, Sorek N, Zalavadia H, Gildin E. Global optimization of grey-box computational systems using surrogate functions and application to highly constrained oil-field operations. *Computers & Chemical Engineering*. 2018;114:99–110.
15. Cozad A, Sahinidis NV, Miller DC. Learning surrogate models for simulation-based optimization. *AIChE Journal*. 2014;60(6):2211–2227.
16. Beykal B, Boukouvala F, Floudas CA, Pistikopoulos EN. Optimal design of energy systems using constrained grey-box multi-objective optimization. *Computers & Chemical Engineering*. 2018;116:488–502. [PubMed: 30546167]
17. Avraamidou S, Beykal B, Pistikopoulos IPE, Pistikopoulos EN. A hierarchical food-energy-water nexus (FEW-N) decision-making approach for land use optimization. *Computer Aided Chemical Engineering*. 2018;44:1885–1890.
18. Beykal B, Avraamidou S, Pistikopoulos IPE, Onel M, Pistikopoulos EN. DOMINO: Data-driven optimization of bi-level mixed-integer nonlinear problems. *Journal of Global Optimization*. 2020; DOI: 10.1007/S10898-020-00890-3.
19. Boukouvala F, Floudas CA. ARGONAUT: Algorithms for Global Optimization of coNstrained grey-box compUTational problems. *Optimization Letters*. 2017;11:895–913.
20. Boukouvala F, Hasan MMF, Floudas CA. Global optimization of general constrained grey-box models: new method and its application to constrained PDEs for pressure swing adsorption. *Journal of Global Optimization*. 2017;67:3–42.
21. Chachuat B. Nonlinear and dynamic optimization: From theory to practice. [https://infoscience.epfl.ch/record/111939/files/Chachuat_07\(IC32\).pdf](https://infoscience.epfl.ch/record/111939/files/Chachuat_07(IC32).pdf). 2007.
22. Allgor RJ. Modeling and computational issues in the development of batch processes. Ph.D. thesis. Massachusetts Institute of Technology, Boston, 1997.
23. Piela PC, Epperly TG, Westerberg KM, Westerberg AW. ASCEND: An object-oriented computer environment for modeling and analysis: The modeling language. *Computers & Chemical Engineering*. 1991;15(1):53–72.
24. Abbott KA. Very large scale modeling. Ph.D. thesis. Carnegie Mellon University, Pittsburgh, 1996.
25. Barton PI. The modelling and simulation of combined discrete/continuous processes. Ph.D. thesis. Imperial College London, London, 1992.
26. Garud SS, Karimi IA, Kraft M. Design of computer experiments: A review. *Computers & Chemical Engineering*. 2017;106:71–95.
27. Cavazzuti M. Optimization methods: from theory to design scientific and technological aspects in mechanics. New York: Springer Science & Business Media, 2012.
28. Dias LS, Ierapetritou MG. Data-driven feasibility analysis for the integration of planning and scheduling problems. *Optimization and Engineering*. 2019;20(4):1029–1066.
29. Basudhar A, Dribusch C, Lacaze S, Missoum S. Constrained efficient global optimization with support vector machines. *Structural and Multidisciplinary Optimization*. 2012;46(2):201–221.

30. Chiang LH, Kotanchek ME, Kordon AK. Fault diagnosis based on fisher discriminant analysis and support vector machines. *Computers & Chemical Engineering*. 2004;28(8):1389–1401.
31. Onel M, Kieslich CA, Pistikopoulos EN. A nonlinear support vector machine-based feature selection approach for fault detection and diagnosis: Application to the tennessee eastman process. *AIChE Journal*. 2019;65(3):992–1005. [PubMed: 32377021]
32. Onel M, Kieslich CA, Guzman YA, Floudas CA, Pistikopoulos EN. Big data approach to batch process monitoring: Simultaneous fault detection and diagnosis using nonlinear support vector machine-based feature selection. *Computers & Chemical Engineering*. 2018;115:46–63. [PubMed: 30386002]
33. Jemwa GT, Aldrich C. Improving process operations using support vector machines and decision trees. *AIChE Journal*. 2005;51(2):526–543.
34. Onel M, Beykal B, Ferguson K, Chiu WA, McDonald TJ, Zhou L, House JS, Wright FA, Sheen DA, Rusyn I, Pistikopoulos EN. Grouping of complex substances using analytical chemistry data: A framework for quantitative evaluation and visualization. *Plos One*. 2019;14(10).
35. Onel M, Beykal B, Wang M, Grimm FA, Zhou L, Wright FA, Phillips TD, Rusyn I, Pistikopoulos EN. Optimal chemical grouping and sorbent material design by data analysis, modeling and dimensionality reduction techniques. *Computer Aided Chemical Engineering*. 2018;43:421–426.
36. EIA. Annual energy outlook 2019 with projections to 2050. <https://www.eia.gov/outlooks/aeo/pdf/aeo2019.pdf>. 2019.
37. EIA. Short-term energy outlook. U.S. hydrocarbon gas liquids (HGL) and petroleum refinery balances. <https://www.eia.gov/outlooks/steo/data/browser/>. 2019.
38. EIA. Petroleum & other liquids natural gas plant field production. https://www.eia.gov/dnav/pet/pet_pnp_gp_dc_nus_mbbldpd_a.htm. 2019.
39. EIA. U.S. ethane consumption, exports to increase as new petrochemical plants come online. <https://www.eia.gov/todayinenergy/detail.php?id=35012>. 2018.
40. Li Z, Floudas CA. Optimal scenario reduction framework based on distance of uncertainty distribution and output performance: I. single reduction via mixed integer linear optimization. *Computers & Chemical Engineering*. 2014;70:50–66.
41. Liu J, Ploskas N, Sahinidis NV. Tuning BARON using derivative-free optimization algorithms. *Journal of Global Optimization*. 2019;74(4):611–637.

Figure 1:

Design of experiments for the motivating example with $t \in [0, 2]$ and $y_o \in [10^{-6}, 2]$. The shaded area represents the feasible region defined by the constraint in Equation 3, $t < 1/y_o$ within the provided variable bounds. The sampling points that satisfy this constraint are represented with filled circles. Candidate points that violate this constraint are removed before calling the problem simulator. Removed samples are represented with hollow circles in the infeasible region.

Figure 2:

Nonlinear SVM model is trained to mimic the constraint, $t < 1/y_o$, by only using the input-output data from the numerical integration of the initial value problem given in Equation 2. SVM classifier can model the boundary of the stability constraint with high accuracy, where the green area corresponds to the feasible, and the red area corresponds to the infeasible class, respectively. Testing scores: Accuracy = 100%, Precision = 100%, Recall = 100%, AUC = 100%, F₁ score = 100%.

Implicit Constraint Handling in Data-Driven Optimization

Phase 1:

Offline Implementation

- Sample Collection for model training, validation and testing
- Two-class classification using Support Vector Machines (SVM)

Phase 2:

Online Implementation

- Implementation of the SVM model to a grey-box optimization solver
- The trained model will filter “infeasible” samples *a priori* to simulation call

Figure 3:

Outline of the SVM-based constraint handling framework for data-driven optimization with numerically infeasible DAEs.

Figure 4:

(A) The molar flowrate of species for the best-found configuration of an ethane thermal cracker. (B) The molar flowrate of C_4^+ species which lead to reactor coking in the best-found configuration.

Figure 5:

(A) The molar flowrate of the main products for the best-found configuration of a propane thermal cracker. (B) The reactor temperature and pressure profiles at the best-found configuration for the propane cracker. The dashed line represents the atmospheric pressure.

Figure 6: Boxplots for the total elapsed time in the online phase for the data-driven optimization of (A) ethane and (B) propane cracking case studies in the presence and absence of the SVM approach.

Table 1:

SVM model testing performance for the first session of runs with ARGONAUT.

Cracking Model	Accuracy (%)	Precision (%)	Recall (%)	AUC (%)	F ₁ score (%)
Ethane	98.492	96.591	100	99.948	98.266
Propane	98.492	98.780	97.590	99.875	98.182

Table 2:

The results of the best solution found with SVM-ARGONAUT integration for the ethane cracking case study.

Decision Variables	Optimal Value	Decision Variables	Optimal Value	Results	Value
Q_1 (kW/m ²)	507.393	$F_{C_2H_6}^0$ (kmol/s)	0.04272	Ethane Conversion	0.7247
Q_2 (kW/m ²)	741.484	$F_{H_2O}^0$ (kmol/s)	0.00377	Ethylene Yield	0.6161
Q_3 (kW/m ²)	954.021	T^o (K)	727.027	Ethylene Selectivity	0.8501
Q_4 (kW/m ²)	462.987	P^o (kPa)	303.741	T^{out} (K)	1170.207
Q_5 (kW/m ²)	260.125	L (m)	32.117	P^{out} (kPa)	131.289

Table 3:

The results of the best solution found with SVM-ARGONAUT integration for the propane cracking case study.

Decision Variables	Optimal Value	Decision Variables	Optimal Value	Results	Value
Q_1 (kW/m ²)	910.298	$F_{C_3H_8}^0$ (kmol/s)	0.02883	Propane Conversion	0.8708
Q_2 (kW/m ²)	282.332	$F_{H_2O}^0$ (kmol/s)	0.01004	Propylene Yield	0.1745
Q_3 (kW/m ²)	55.025	T^o (K)	916.716	Ethylene Yield	0.5377
Q_4 (kW/m ²)	177.221	P^o (kPa)	325.276	Propylene Selectivity	0.2004
Q_5 (kW/m ²)	21.516	L (m)	34.957	Ethylene Selectivity	0.6175

Table 4:

The profit breakdown for the best-found solution of ethane and propane cracking case studies.

Ethane Cracking		Propane Cracking	
Objective Variable	Value (\$/s)	Objective Variable	Value (\$/s)
Ethane Feed Cost	− 0.2991	Propane Feed Cost	− 0.6361
Steam Feed Cost	− 0.0006	Steam Feed Cost	− 0.0022
Heating Cost	− 0.0621	Heating Cost	− 0.0381
Investment Cost	− 0.0088	Investment Cost	− 0.0095
Decoking Cost	− 0.0855	Decoking Cost	− 0.0345
Ethylene Production	+ 0.7920	Propylene Production	+ 0.2581
-		Ethylene Production	+ 0.5468
Total	+ 0.3359	Total	+ 0.0845