A collection of selected scientific papers

1. Gottu Mukkula, A. R. [35%] – Paulen, R. [65%] Model-based design of optimal experiments for nonlinear systems in the context of guaranteed parameter estimation, Computers & Chemical Engineering, vol. 99, pp. 198-213, 2017

2. Gottu Mukkula, A. R. [50%] – Paulen, R. [50%] Optimal experiment design in nonlinear parameter estimation with exact confidence regions, Journal of Process Control, vol. 83, pp. 187-195, 2019

3. Kusumo, K. [30%] – Gomoescu, L. [25%] – Paulen, R. [15%] – García Muñoz, S. [5%] – Pantelides, C. C. [5%] – Shah, N. [5%] – Chachuat, B. [15%] Bayesian Approach to Probabilistic Design Space Characterization: A Nested Sampling Strategy, Industrial & Engineering Chemistry Research, vol. 59, np. 6, pp. 2396-2408, 2020

4. Mojto, M. [40%] – Ľubušký, K. [10%] – Fikar, M. [20%] – Paulen, R. [30%] Data-based design of inferential sensors for petrochemical industry, Computers & Chemical Engineering, vol. 153, pp. 107437, 2021

5. Paulen, R. [50%] – Villanueva, M. [10%] – Chachuat, B. [40%]: Guaranteed parameter estimation of non-linear dynamic systems using high-order bounding techniques with domain and CPU-time reduction strategies. IMA Journal of Mathematical Control and Information, pp. 563–587, 2016.

6. Peric, N. [20%] – Paulen, R. [30%] – Villanueva, M. [10%] – Chachuat, B. [40%] Set-membership nonlinear regression approach to parameter estimation, Journal of Process Control, vol. 70, pp. 80-95, 2018

7. Martí, R. [35%] – Lucia, S. [25%] – Sarabia, D. [10%] – Paulen, R. [20%] – Engell, S. [5%] – de Prada, C. [5%] Improving scenario decomposition algorithms for robust nonlinear model predictive control, Computers & Chemical Engineering, vol. 79, pp. 30-45, 2015

8. Thangavel, S. [30%] – Lucia, S. [30%] – Paulen, R. [30%] – Engell, S. [10%]: Dual robust nonlinear model predictive control: A multi-stage approach. Journal of Process Control, vol. 72, pp. 39–51, 2018.

9. Subramanian, S. [40%] – Lucia, S. [30%] – Paulen, R. [20%] – Engell, S. [10%] Tube-enhanced multi-stage model predictive control for flexible robust control of constrained linear systems with additive and parametric uncertainties, International Journal of Robust and Nonlinear Control, vol. 31, no. 9, pp. 4458-4487, 2021

10. Wenzel, S. [55%] – Paulen, R. [25%] – Beisheim, B. [5%] – Krämer, S. [5%] – Engell, S. [10%]: Market-Based Coordination of Shared Resources in Cyber-physical Production Sites. Chemie Ingenieur Technik, no. 5, vol. 89, pp. 636–644, 2017.

Computers and Chemical Engineering 99 (2017) 198–213

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Model-based design of optimal experiments for nonlinear systems in the context of guaranteed parameter estimation

Anwesh Reddy Gottu Mukkula, Radoslav Paulen*

Process Dynamics and Operations Group, Technische Universität Dortmund, Emil-Figge-Strasse 70, Dortmund 44227, Germany

ARTICLE INFO

Article history:
Received 4 October 2016
Received in revised form 7 January 2017
Accepted 16 January 2017
Available online 21 January 2017

Keywords:
Optimal experiment design
Estimation algorithms
Parameter estimation
Bounded noise
Bounded-error estimation

ABSTRACT

An approach to the design of experiments is presented in the framework of bounded-error (guaranteed) parameter estimation for nonlinear static and dynamic systems. The guaranteed parameter estimation determines non-asymptotic confidence limits on the unknown parameters of a mathematical model. An essential part of the solution procedure is the approximation of the joint-confidence region. In this contribution, we develop and analyze the procedure and different ways of achieving a tight over-approximation of the solution set of guaranteed parameter estimation based on the expected values of parameters. Finally we propose to solve the problem of the design of experiments as a bilevel program. We demonstrate our approach and analyze the nature of the problem in the static and dynamic case studies. The proposed approach is also compared to the experiment design in the context of least-squares estimation and to the linearization-based techniques for optimal experiment design proposed in the literature earlier.

1. Introduction

A mathematical model is a (usually abstract) representation of a true system via sets of equations (algebraic equations, ordinary differential equations, partial differential equations, etc.) and inequality and logical conditions (e.g., a range of model validity). The mathematical models are widely used in all fields of science and engineering (e.g., physics, economics, climatology, aerospace, and many more) (Quarteroni, 2009). They have also, quite naturally, found their use and became an integral part of state-of-the-art methodologies in process systems engineering such as product and plant design, control system design and operations optimization (Pantelides and Renfro, 2013; Fung et al., 2016; Safdarnejad et al., 2016).

The procedure for model development is usually divided into three major steps: specification of the model structure, design and realization of the experiments, and estimation of the unknown parameters. The latter phase, often referred to as model fitting, normally proceeds by determining parameter values for which the model predictions closely match the available process measurements (Ogunnaike and Ray, 1994; Ljung, 1999), where the results are strongly dependent on the observation (measuring) capabilities and experimental conditions.

Many techniques have been developed for estimating the unknown parameters of mathematical models (Bates and Watts, 1988). Among them, least-squares estimation is an approach to estimate the most likely values of unknown parameters under the assumption that the measurement noise is normally distributed. Should the assumption of normal distribution of the measurement noise be violated, the results of least-squares estimation might become unreliable.

One possible way of addressing the problem of estimation bias, resulting from *local* violation of the normal-distribution assumption, is the use of set-membership estimation. The set-membership estimation (Milanese and Vicino, 1991; Bai et al., 1995), often referred to as guaranteed parameter estimation (GPE), has been developed as a technique that overcomes the problem of insufficient knowledge on the measurement error distribution assuming this to be arbitrary but bounded. The problem of GPE can be formulated as an identification of the set of all possible parameter values of the given model such that these are not falsified by the corresponding plant measurements within prescribed error bounds. A set-inversion algorithm (e.g., SIVIA by Jaulin and Walter (1993)) can be generally applied for GPE of a nonlinear system, whereby the parameter set is successively partitioned into smaller boxes and exclusion tests are performed to eliminate some of these boxes, until a given threshold on the approximation level is met. Since its advent, GPE found applications in many engineering areas (Marco et al., 2000; Jaulin et al., 2002; Lin and Stadtherr, 2007; Hast et al., 2015; Paulen et al., 2016). Naturally the result of the GPE greatly depends on the quality of experimental data.

* Corresponding author.
E-mail address: radoslav.paulen@bci.tu-dortmund.de (R. Paulen).

http://dx.doi.org/10.1016/j.compchemeng.2017.01.029

In order to mitigate the negative effects of measurement noise in identifying the unknown parameters, one has to design an ideal experiment to gain maximum knowledge about the process possibly at a minimum expense of time and resources. Optimal experiment design (OED) has been extensively studied in literature as an approach that provides the best available selection of experimental conditions of a system for collection of the most information-rich data (Fisher, 1935). The problem of the optimal experiment design can also be interpreted as identification of learning capabilities of the given model, given the quality of observation data. This study concentrates on model-based design of experiments so the OED problem boils down to an optimization problem whose nature is given by the nature and complexity of the chosen (given) model.

The model-based OED techniques are well-established in the framework of least-squares estimation (Franceschini and Macchietto, 2008). The situation is, however, different with respect to the techniques available to address the problem of optimal experiment design in the context of guaranteed parameter estimation. While the problem of OED for GPE of linear models (static, discrete-time and FIR models) has been considered to sufficient extent (Pronzato and Walter, 1989; Bai et al., 1995; Borchers et al., 2011; Tanaskovic et al., 2014), the problem of OED for GPE of nonlinear (and general dynamic) systems has received a little attention. The techniques proposed so far use either linearization (Pronzato and Walter, 1989, 1990) or state bounding (Telen et al., 2013), e.g., via ellipsoidal calculus (Kurzhanski et al., 1994), which was introduced in the context of OED for GPE of dynamic systems. Despite resulting in a simple optimization problem, the use of linearization-based techniques might yield substantially suboptimal results due to the approximation error. The approach based on state bounding mitigates the error of linearization. However, a certain amount of conservatism is introduced due to loose over-approximation of the GPE solution set.

The work presented here provides a methodology for performing the model-based optimal experiment design in the context of guaranteed parameter estimation for nonlinear static and dynamic systems. The GPE solution set is first *tightly* over-approximated by a box (orthotope). Several ways of setting up the problem of over-approximation of the GPE solution set are presented and analyzed. The OED problem is then formulated as a bilevel optimization of the selected design criterion over the nonlinear optimization problem that over-approximates the set of guaranteed parameter estimates. The arising bilevel program is then regularized such that the resulting nonlinear optimization problem with complementarity constraints is well-conditioned. The new methodology proposed here is demonstrated using two case studies, one in static and one in dynamic design of experiments, where the results of OED in the context of GPE are evaluated for different design criteria, different decisions on over-approximation nature of the GPE solution set and compared to the linearization-based OED as well as to the classical optimal experiment design. The preliminary results of the here proposed methodology, including some applications to the domain of chemical engineering, were presented in Gottu Mukkula and Paulen (2016a,b).

2. Problem formulation

This paper is concerned with the design of experiments for parameter estimation of nonlinear systems. We consider an identified system which can be schematically depicted as illustrated in Fig. 1. We study both static and dynamic systems that can be described using the mathematical models of the following forms

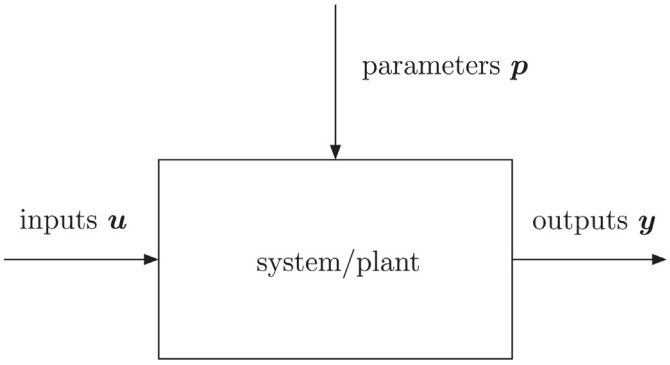


Fig. 1. Conceptual view on the identified system.

- Static system

$$\boldsymbol{x}_\tau = \boldsymbol{f}(\boldsymbol{u}_\tau, \boldsymbol{p}), \tag{1a}$$

$$\hat{\boldsymbol{y}}(\boldsymbol{p}, \tau) = \boldsymbol{h}(\boldsymbol{x}_\tau, \boldsymbol{u}_\tau, \boldsymbol{p}), \tag{1b}$$

- Dynamic system

$$\dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{p}), \quad \forall t \in [t_0, t_f], \tag{2a}$$

$$\boldsymbol{x}(t_0) = \boldsymbol{g}(\boldsymbol{u}(t_0), \boldsymbol{p}), \tag{2b}$$

$$\hat{\boldsymbol{y}}(\boldsymbol{p}, \tau) = \boldsymbol{h}(\boldsymbol{x}(\tau), \boldsymbol{u}(\tau), \boldsymbol{p}), \quad \tau \in [t_0, t_f]. \tag{2c}$$

Here $\boldsymbol{x} \in \mathbb{R}^{n_x}$ ($\boldsymbol{x}(t) \in \mathbb{R}^{n_x}$ in the dynamic case) denotes the vector of internal (state) variables, $\boldsymbol{u}_\tau \in \mathbb{R}^{n_u}$ ($\boldsymbol{u}(t) \in \mathbb{R}^{n_u}$ in the dynamic case) represents inputs (degrees of freedom) of the system, $\boldsymbol{p} \in \mathbb{R}^{n_p}$ is a vector of (a priori unknown) parameters, $\hat{\boldsymbol{y}} \in \mathbb{R}^{n_y}$ are predicted outputs (observations) of the system and $\tau \in \mathbb{R}$ is a degree of freedom concerned with the measurement (output) function. State equation functions $\boldsymbol{f}(\cdot)$, $\boldsymbol{g}(\cdot)$ and measurement function $\boldsymbol{h}(\cdot)$ are assumed to be twice continuously differentiable.

The variable τ denotes the serial number of an experiment in the static case and, respectively, the measurement instant in the dynamic case. An experiment is characterized, in the static case, by assigning the time-invariant degrees of freedom $\boldsymbol{u}_\tau$ and collecting the corresponding (steady-state) measurements in N experiments. Hence the measurements are obtained at $\boldsymbol{\tau} := [\tau_1, \tau_2, \ldots, \tau_N]$. The dynamic case is characterized by the selected time-varying input $\boldsymbol{u}(t)$ and by measuring the system outputs at $\boldsymbol{\tau} := [\tau_1, \tau_2, \ldots, \tau_N]$. Hence that we limit the scope of this study to the case where measurements are only taken at discrete time instants and do not treat the case of continuous measurements taken on some chosen intervals, as for example done in Sager (2013), explicitly. It is, however, possible to extend the here presented methodology to this case.

Note also that the form of the function $\boldsymbol{g}(\cdot)$ allows for a subset of initial conditions being the degrees of freedom of the experiment. Additionally, the independent variable t in (2a) can be interpreted as a space coordinate and, thus, the dynamic case could boil down to a static case of a parameter-distributed system. The dynamic case of parameter-distributed system is not treated explicitly in this work but the presented methodologies are extensible to this case straightforwardly.

In the remainder of the theoretical part of this paper, we will present our developments in the context of dynamic systems only. The application to static cases is straightforward and some of its specificities will be recalled at appropriate places.

3. Optimal experiment design in the context of least-squares estimation

Preliminary to main results regarding OED for guaranteed estimation, we review some classical results in the context of least-squares estimation.

3.1. Joint-confidence regions of least-squares estimates

Once output measurements, $\boldsymbol{y}(\tau_i), \forall i \in \{1, 2, \ldots, N\}$, are taken, the *classical* parameter estimation seeks for one particular instance $\boldsymbol{p}^E$ of the parameters for which the (possibly weighted) normed difference between the measurements and the corresponding model outputs $\hat{\boldsymbol{y}}$ is minimized e.g., in the least-squares sense

$$\begin{aligned} \min_{\boldsymbol{p}\in\mathcal{P}_0} \quad & \sum_{i=1}^{N} \left(\boldsymbol{y}(\tau_i) - \hat{\boldsymbol{y}}(\boldsymbol{p}, \tau_i)\right)^T \mathbf{Q} \left(\boldsymbol{y}(\tau_i) - \hat{\boldsymbol{y}}(\boldsymbol{p}, \tau_i)\right) \\ \text{s.t.} \quad & \forall i \in \{1, 2, \ldots, N\}: \\ & \hat{\boldsymbol{y}}(\boldsymbol{p}, \tau_i) = \boldsymbol{h}(\boldsymbol{x}(\tau_i), \boldsymbol{u}(\tau_i), \boldsymbol{p}), \\ & \dot{\boldsymbol{x}}(t) = \quad \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{p}), \quad \forall t \in [t_0, t_f], \\ & \boldsymbol{x}(t_0) = \boldsymbol{g}(\boldsymbol{u}(t_0), \boldsymbol{p}), \end{aligned} \tag{3}$$

for given $\boldsymbol{u}(t), \forall t \in [t_0, t_f]$ and $\boldsymbol{y}(\tau_i)\ \forall \tau_i \in [t_0, t_f], \forall i \in \{1, 2, \ldots, N\}$. Here $\mathbf{Q}$ is the weighting function for different outputs, usually taken as the inverse of the covariance matrix of the measurement noise. The purpose of the set $\mathcal{P}_0$ is to restrict the parameters to physically meaningful values.

Due to the presence of measurement noise, the true plant outputs $\boldsymbol{y}_p$ cannot be measured precisely and are known to be biased with measurement errors $\boldsymbol{e}$ such that

$$\boldsymbol{y}(\tau_i) = \boldsymbol{y}_p(\tau_i) + \boldsymbol{e}(\tau_i). \tag{4}$$

The least-squares estimation coincides with maximum-likelihood approach when the assumption of normal statistical distribution of the error is made, i.e., $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, \mathbf{Q}^{-1})$. In order to account for the propagation of the measurement error in the obtained parameter estimates, a joint-confidence ellipsoid can be constructed. This takes the form

$$(\boldsymbol{p} - \boldsymbol{p}^E)^T \mathbf{FIM} (\boldsymbol{p} - \boldsymbol{p}^E) \leq n_p\, F_{dist}(n_p, N - n_p, \alpha), \tag{5}$$

where F_{dist} is a quantile of the Fisher statistical distribution, α stands for the desired confidence level (normally 95% or 99%) and **FIM** is the Fisher information matrix (Franceschini and Macchietto, 2008; Kitsos, 2013)

$$\mathbf{FIM} := \sum_{i=1}^{N} \left(\frac{\partial \hat{\boldsymbol{y}}(\boldsymbol{p}, \tau_i)}{\partial \boldsymbol{p}}\Big|_{\boldsymbol{p}^E} \right)^T \mathbf{Q} \left(\frac{\partial \hat{\boldsymbol{y}}(\boldsymbol{p}, \tau_i)}{\partial \boldsymbol{p}}\Big|_{\boldsymbol{p}^E} \right). \tag{6}$$

Note that here the employment of $\boldsymbol{p}^E$ implies that a valid confidence region can only be obtained once the least-squares estimate converges to the true values of parameters (Galvanin et al., 2009).

3.2. Optimal experiment design for least-squares estimation

An optimization can be performed to shape the **FIM** such as to improve the quality of estimates implied by the shape and size of the resulting joint-confidence region due to the possibility of exploration of the system under different input conditions. In the dynamic case, this is referred to as excitation of the system implied by time-varying control input sequence. We refer to the optimal experiment design in the context of least-squares estimation as a *classical* design of experiments.

Assuming a piece-wise constant parametrization of the input, the OED problem can be formulated as an optimization problem with $N(n_u + 1)$ decision variables that minimizes a certain measure of **FIM**, $\phi_{\text{LS}}(\mathbf{FIM})$, over admissible values of inputs $\boldsymbol{u}(t)$ and a prespecified number of measurements (N).

$$\begin{aligned} \min \quad & \phi_{\text{LS}}(\mathbf{FIM}) \\ \boldsymbol{u}(t) \in & \left[\boldsymbol{u}^L(t), \boldsymbol{u}^U(t)\right], \quad \forall t \in \left[t_0, t_f\right] \\ \{\tau_1, \tau_2, \ldots, \tau_N\} \in & \left[t_0, t_f\right] \\ \text{s.t.} \quad & \forall i \in \{1, 2, \ldots, N\}: \\ & \hat{\boldsymbol{y}}(\boldsymbol{p}, \tau_1) = \boldsymbol{h}(\boldsymbol{x}(\tau_1), \boldsymbol{u}(\tau_1), \boldsymbol{p}), \\ & \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{p}), \quad \forall t \in \left[t_0, t_f\right], \\ & \boldsymbol{x}(t_0) = \boldsymbol{g}(\boldsymbol{u}(t_0), \boldsymbol{p}). \end{aligned} \tag{7}$$

Several experiment designs are then possible based on the functional form of $\phi_{\text{LS}}(\cdot)$. We devote our attention in this study to designs that are completely presented below (Table 1). For instance, the A-optimal design (in short, A design) minimizes the trace of $\mathbf{FIM}^{-1}$. This results in the minimum circumference of the box enclosing the confidence ellipsoid which can be also formulated as $\sum_{j=1}^{n_p} p_j^U - p_j^L$ where the parametric bounds are derived from the projections of the joint-confidence ellipsoid on parametric axes. The D-optimal design (in short, D design) minimizes the determinant of $\mathbf{FIM}^{-1}$. This results in the minimum volume of the joint-confidence ellipsoid.

Table 1
Objective function of optimal experiment design problem for least-squares (LS) and guaranteed parameter estimation (GPE).

Criterion	LS formulation ($\phi_{\text{LS}}(\mathbf{FIM})$)	GPE formulation ($\phi_{\text{GPE}}(\boldsymbol{\pi})$)
A	$\text{trace}(\mathbf{FIM}^{-1})$	$\sum_{j=1}^{n_p}(p_j^U - p_j^L)$
D	$\det(\mathbf{FIM}^{-1})$	$\prod_{j=1}^{n_p}(p_j^U - p_j^L)$
E	$\max_i \lambda_i(\mathbf{FIM})$	$\max_j(p_j^U - p_j^L)$
Modified E	$\max_i \lambda_i(\mathbf{FIM})/\min_i \lambda_i(\mathbf{FIM})$	$\max_j(p_j^U - p_j^L)/\min_j(p_j^U - p_j^L)$

4. Optimal experiment design in the context of guaranteed parameter estimation

This section contains the main theoretical results of this contribution where we formulate a methodology for finding optimal experiment designs in the context of guaranteed parameter estimation. As in the case of the experiment design for least-squares estimation, there are two steps involved in this procedure: (a) description of the joint-confidence region and (b) design of input sequence that optimizes the geometry of this region. We proceed by introducing both these ingredients in the framework of guaranteed parameter estimation.

4.1. Joint-confidence regions of guaranteed parameter estimation

The guaranteed parameter estimation (GPE), referred to as bounded-error parameter estimation or set-membership estimation was formulated in Schweppe (1968) and Bertsekas and Rhodes (1971) as an approach that identifies a subset of the a priori admissible parametric space $\mathcal{P}_0$ of all values of the parameters that give the model predictions consistent with the obtained measurements from N experiments within the tolerance interval of the, assumed, bounded measurement noise. In other words, guaranteed parameter estimation assumes that $\boldsymbol{e}(\tau_i)$ is bounded by a known interval $[\boldsymbol{e}^L(\tau_i), \boldsymbol{e}^U(\tau_i)]$, where the superscripts L and U represent lower and upper bounds of an interval box (understood component-wise), and that the error distribution inside this interval is arbitrary. The lower and upper bounds are usually obtained from the data-sheets of the sensor, which are usually provided by the vendor (Hast et al., 2015). Note that we study here the case of absolute measurement

error bounds. The extension of our study, which would consider relative bounds, is not significantly difficult and a direction towards it is shown in Pronzato and Walter (1990) for linear systems.

The objective of the guaranteed parameter estimation is to characterize, usually over-approximate as closely as possible (Jaulin and Walter, 1993), the set

$$\mathcal{P}_{\mathrm{e}} := \left\{ \boldsymbol{p} \in \mathcal{P}_0 \;\middle|\; \begin{array}{l} \exists \hat{\boldsymbol{y}}_i(\boldsymbol{p},\tau_i), \quad \forall i \in \{1,2,\ldots,N\}: \\ \hat{\boldsymbol{y}}_i(\boldsymbol{p},\tau_i) \in \boldsymbol{y}(\tau_i) + [\boldsymbol{e}^L(\tau_i), \boldsymbol{e}^U(\tau_i)], \\ \hat{\boldsymbol{y}}_i(\boldsymbol{p},\tau_i) = \boldsymbol{h}(\boldsymbol{x}(\tau_i),\boldsymbol{u}(\tau_i),\boldsymbol{p}), \\ \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t),\boldsymbol{u}(t),\boldsymbol{p}), \quad \forall t \in [t_0,t_f], \\ \boldsymbol{x}(t_0) = \boldsymbol{g}(\boldsymbol{u}(t_0),\boldsymbol{p}). \end{array} \right\}, \tag{8}$$

for given $\boldsymbol{u}(t), \forall\, t \in [t_0, t_f]$ and $\boldsymbol{y}(\tau_i), \forall \tau_i \in [t_0, t_f], \forall\, i \in \{1, 2, \ldots, N\}$. This set is a counterpart of the joint-confidence ellipsoid (5) within the context of the bounded-error (guaranteed) estimation.

The problem of GPE is essentially a problem of constraint propagation and an implied feasibility problem and can be formulated as

$$\begin{aligned} \mathcal{P}_{\mathrm{e}} := \underset{\boldsymbol{p}\in\mathcal{P}_0}{\arg\,\mathrm{find}} \quad & \text{all } \boldsymbol{p} \\ \text{s.t.} \quad & \tau_i \in [t_0,t_f], \quad \forall i \in \{1,2,\ldots,N\}: \\ & \boldsymbol{e}^L(\tau_i) \le \boldsymbol{y}(\tau_i) - \hat{\boldsymbol{y}}(\boldsymbol{p},\tau_i) \le \boldsymbol{e}^U(\tau_i), \\ & \hat{\boldsymbol{y}}(\boldsymbol{p},\tau_i) = \boldsymbol{h}(\boldsymbol{x}(\tau_i),\boldsymbol{u}(\tau_i),\boldsymbol{p}), \\ & \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t),\boldsymbol{u}(t),\boldsymbol{p}), \quad \forall t \in [t_0,t_f], \\ & \boldsymbol{x}(t_0) = \boldsymbol{g}(\boldsymbol{u}(t_0),\boldsymbol{p}), \end{aligned} \tag{9}$$

for given $\boldsymbol{u}(t), \forall\, t \in [t_0, t_f]$ and $\boldsymbol{y}(\tau_i), \forall \tau_i \in [t_0, t_f], \forall\, i \in \{1, 2, \ldots, N\}$.

Note that the estimation is not carried out in the least-squares sense and the resulting set $\mathcal{P}_{\mathrm{e}}$ is generally not consistent with the joint-confidence region inferred from (5). This means that the experiment design obtained from (7) is generally not optimal for the guaranteed parameter estimation since the results of the least-squares estimation robustified with the a posteriori confidence analysis and the results of guaranteed estimation are not the same, in general, and since the expectation of the realization of measurement error differs in the least-squares (normal distribution, unbounded error) and guaranteed (arbitrary distribution, bounded error) estimation. For the sake of completeness, note that a special case exists here, i.e., for linear systems the D-optimal experiment design coincides with the design that minimizes the volume of GPE solution set (Pronzato and Walter, 1989).

4.2. Over-approximation of the set of guaranteed parameter estimates

We describe the joint-confidence parametric region by the orthotopic projection of the set $\mathcal{P}_{\mathrm{e}}$. Our assumption is that the set $\mathcal{P}_{\mathrm{e}}$ is generally nonconvex but connected. The factors that cause the non-connectedness of the set $\mathcal{P}_{\mathrm{e}}$ are discussed in Pronzato and Walter (1990) and therein the ways to mitigate this phenomenon are proposed.

The concept of the tight over-approximation of the set $\mathcal{P}_{\mathrm{e}}$ using an orthotope (box) is illustrated in Fig. 2. We characterize each face of the orthotope by a point in the parametric space, which is naturally a combination of n_p variables. For instance, the point that bears the information about the lower bound of the parameter p_1, i.e., p_1^L, is characterized as $(p_1^L, p_2^{1,L}, \ldots, p_{n_p}^{1,L})^T$ where values of $p_2^{1,L}, \ldots, p_{n_p}^{1,L}$ must be such that the point is a member of the set $\mathcal{P}_e$. The resulting orthotope is then given by the set of $2n_p$ points that

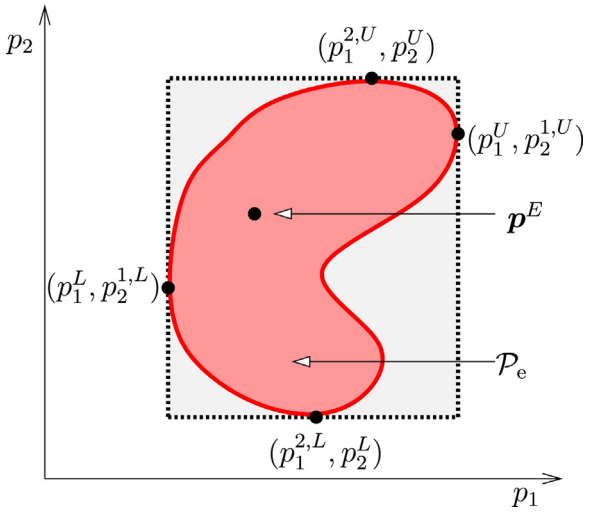


Fig. 2. Illustration of the over-approximation of the GPE solution set.

are parametrized by $2n_p^2$ variables

$$\boldsymbol{\pi} := \left\{ \begin{pmatrix} p_1^L \\ p_2^{1,L} \\ \vdots \\ p_{n_p}^{1,L} \end{pmatrix}, \begin{pmatrix} p_1^U \\ p_2^{1,U} \\ \vdots \\ p_{n_p}^{1,U} \end{pmatrix}, \begin{pmatrix} p_1^{2,L} \\ p_2^{L} \\ \vdots \\ p_{n_p}^{2,L} \end{pmatrix}, \begin{pmatrix} p_1^{2,U} \\ p_2^{U} \\ \vdots \\ p_{n_p}^{2,U} \end{pmatrix}, \ldots, \begin{pmatrix} p_1^{n_p,L} \\ p_2^{n_p,L} \\ \vdots \\ p_{n_p}^{L} \end{pmatrix}, \begin{pmatrix} p_1^{n_p,U} \\ p_2^{n_p,U} \\ \vdots \\ p_{n_p}^{U} \end{pmatrix} \right\}. \tag{10}$$

The choice of this parametrization goes in line with the partial motivation of our study that lies in the recent developments in robust nonlinear model predictive control (Lucia et al., 2014) that use scenario representation for the possible realization of uncertain parameters for which the over-approximation based on a box gives practically relevant representation of uncertainty, for example w.r.t. constraints.

We will assume, throughout the paper, that the set $\mathcal{P}_{\mathrm{e}}$ is not constrained by the set $\mathcal{P}_0$, i.e., $\partial\mathcal{P}_0 \cap \boldsymbol{\pi} = \emptyset$. This is influenced by the choice of $\mathcal{P}_0$, which might include some a priori knowledge about the identified parameters, and also by the well-posedness of the estimation problem (identifiability, parameter correlations, etc.). We note that the study of such phenomena is beyond the scope of the paper, but the presented developments are readily applicable and extensible to encompass these cases. We will assume, in line with the classical optimal design of experiments, that the a priori estimate $\boldsymbol{p}^E$ represents true values of parameters and is, thus, contained in the set $\mathcal{P}_{\mathrm{e}}$. This essentially means that we can predict the (quality of) future measurements if we know the realization of the future realization of measurement error. As the basic assumption of guaranteed parameter estimation is that the distribution of measurement error is unknown, one has to account for this in some way. We propose here three distinct choices for the expectation of the realization of the measurement error as shown in Fig. 3, for the case of dynamic system:

1. *Nominal over-approximation.* Here we assume that the realization of the measurement error is zero. This means that the procedure of guaranteed estimation provides a general shape of the set $\mathcal{P}_{\mathrm{e}}$, assuming that the set of active constraints of (9)

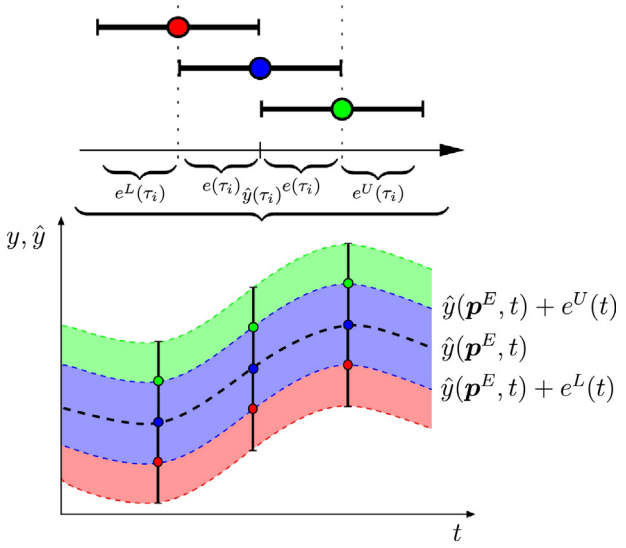


Fig. 3. Illustration of the over-approximation of the solution set of guaranteed parameter estimation in the space of output trajectories for the case of dynamic system.

remains constant under the particular (true) realization of the measurement error.

2. *Worst-case over-approximation*. Here the realization of error is chosen such that the maximal perimeter/volume (of the over-approximation) of the set $\mathcal{P}_e$ is reached.
3. *Global over-approximation*. All possible realizations of the measurement error are taken into account here, so all the possible guaranteed parameter estimation sets are contained in the over-approximating box.

The employment of a concrete over-approximation scheme is a tuning parameter of the procedure for optimal experiment design and, as particularly shown in the case studies section of this paper, the appropriateness of this choice, as might be expected, depends on the underlying unknown measurement error distribution, whose qualitative a priori guess might be available for the problem at hand.

4.2.1. Nominal over-approximation of the set of guaranteed parameter estimates

Here we assume that the realization of the measurement error is zero, i.e.,

$$\boldsymbol{y}(\tau_i) = \boldsymbol{y}_p(\tau_i). \tag{11}$$

This is illustrated in Fig. 3 where the blue dots represent noise-free measurements. The blue tube illustrates the output trajectories that are not falsified by guaranteed parameter estimation under the assumption of noise-free measurements. Using this assumption the orthotopic projection of the set $\mathcal{P}_e$ can be identified as

$$\begin{aligned}
\max_{\boldsymbol{\pi}} \quad & \sum_{j=1}^{n_p} p_j^U - p_j^L \\
\text{s.t.} \quad & \forall i \in \{1,2,\ldots,N\}, \quad \forall j \in \{1,\ldots,2n_p\}, \quad \forall \boldsymbol{p} \in \{\boldsymbol{\pi}, \boldsymbol{p}^E\}: \\
& \boldsymbol{e}^L(\tau_i) \le \hat{\boldsymbol{y}}(\boldsymbol{p}^E, \tau_i) - \hat{\boldsymbol{y}}(\boldsymbol{\pi}_j, \tau_i) \le \boldsymbol{e}^U(\tau_i), \\
& \hat{\boldsymbol{y}}(\boldsymbol{p}, \tau_i) = \boldsymbol{h}(\boldsymbol{x}(\tau_i), \boldsymbol{u}(\tau_i), \boldsymbol{p}), \\
& \dot{\boldsymbol{x}}(t) = \quad \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{p}), \quad \forall t \in [t_0, t_f], \\
& \boldsymbol{x}(t_0) = \boldsymbol{g}(\boldsymbol{u}(t_0), \boldsymbol{p}),
\end{aligned} \tag{12}$$

for given $\boldsymbol{u}(t), \forall\, t \in [t_0, t_f]$ and $\boldsymbol{y}(\tau_i), \forall \tau_i \in [t_0, t_f], \forall\, i \in \{1,2,\ldots,N\}$. The optimization problem (12) involves $2n_p^2$ decision variables and $4Nn_yn_p$ inequality constraints. We note that the problem (12) is separable and can be solved in parallel as a series of optimization problems, of the form $\min_{\boldsymbol{\pi}_j}/\max_{\boldsymbol{\pi}_j} p_j^{L/U}$, $\forall j \in \{1,\ldots,2n_p\}$, with common knowledge about the evolution of the outputs for expected values of the parameters. We also note that an equivalent yet nonseparable formulation of the objective function in (12) is $\prod_{j=1}^{n_p} p_j^U - p_j^L$.

4.2.2. Worst-case over-approximation of the set of guaranteed parameter estimates

The downside of the nominal formulation is that the resulting over-approximation might not cover the set $\mathcal{P}_e$ that would result from the true realization of the measurement error. A possible (partial) remedy is to reformulate the problem (12) for the worst-case realization of the error, i.e., the realization of error that results in the maximal perimeter/volume (of the over-approximation) of the set $\mathcal{P}_e$. This can be achieved by introducing an additional Nn_y-dimensional vector of degrees of freedom, $\boldsymbol{e}(\tau_i), \forall\, i \in \{1,\ldots,N\}$, to the problem (12) such that the worst case of the measurement

$$\boldsymbol{y}(\tau_i) = \boldsymbol{y}_p(\tau_i) + \boldsymbol{e}(\tau_i), \tag{13}$$

is identified with $\boldsymbol{e}(\tau_i) \in [\boldsymbol{e}^L(\tau_i), \boldsymbol{e}^U(\tau_i)]$. This concept is illustrated in Fig. 3 where the red and green dots represent measurements with minimal and maximal measurement error, respectively. The idea of the underlying optimization problem is to find the realization of the measurement error that results in the largest possible over-approximation of the set $\mathcal{P}_e$. This can be visualized, as in the bottom plot of Fig. 3, as an unfalsification of the trajectories in blue tube and the trajectories from either red or green tube. The resulting formulation reads as

$$\begin{aligned}
\max_{\boldsymbol{\pi},\ \boldsymbol{e}(\tau_i) \in [\boldsymbol{e}^L(\tau_i), \boldsymbol{e}^U(\tau_i)]} & \sum_{j=1}^{n_p} p_j^U - p_j^L \\
\forall i \in \{1,2,\ldots,N\} & \\
\text{s.t.} \quad \forall i \in \{1,2,\ldots,N\}, \quad & \forall j \in \{1,\ldots,2n_p\}, \quad \forall \boldsymbol{p} \in \{\boldsymbol{\pi}, \boldsymbol{p}^E\}: \\
& \boldsymbol{e}^L(\tau_i) \le \hat{\boldsymbol{y}}(\boldsymbol{p}^E, \tau_i) + \boldsymbol{e}(\tau_i) - \hat{\boldsymbol{y}}(\boldsymbol{\pi}_j, \tau_i) \le \boldsymbol{e}^U(\tau_i), \\
& \hat{\boldsymbol{y}}(\boldsymbol{p}, \tau_i) = \boldsymbol{h}(\boldsymbol{x}(\tau_i), \boldsymbol{u}(\tau_i), \boldsymbol{p}), \\
& \dot{\boldsymbol{x}}(t) = \quad \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{p}), \quad \forall t \in [t_0, t_f], \\
& \boldsymbol{x}(t_0) = \boldsymbol{g}(\boldsymbol{u}(t_0), \boldsymbol{p}),
\end{aligned} \tag{14}$$

for given $\boldsymbol{u}(t), \forall\, t \in [t_0, t_f]$ and $\boldsymbol{y}(\tau_i), \forall \tau_i \in [t_0, t_f], \forall\, i \in \{1,2,\ldots,N\}$. Note that the separable nature of the problem (12) is lost under this transformation as there exists a coupling of the variables through the constraints.

4.2.3. Global over-approximation of the set of guaranteed parameter estimates

Another possibility for over-approximation of the set $\mathcal{P}_e$ that does not depend on the actual realization of the measurement error can be achieved by considering all the possible output trajectories

$$\boldsymbol{y}(\tau_i) = \boldsymbol{y}_p(\tau_i) + [\boldsymbol{e}^L(\tau_i), \boldsymbol{e}^U(\tau_i)], \tag{15}$$

not only the worst-case one (as discussed in the previous section). This can be illustrated via Fig. 3 as an inclusion of all the output trajectories in green, blue and red tubes and an identification of the corresponding parameters. As a result the proposed over-approximation envelopes all the possible results of the guaranteed parameter estimation under each possible realization of the measurement error. Note that such finite robust solution to parameter estimation is inadmissible under the assumption of unbounded measurement noise, e.g., under normally distributed measurement noise.

The *global* over-approximation of the set $\mathcal{P}_e$ can be identified by solving

$$\begin{aligned}
\max_{\boldsymbol{\pi}} \quad & \sum_{j=1}^{n_p} p_j^U - p_j^L \\
\text{s.t.} \quad & \forall i \in \{1,2,\ldots,N\}, \quad \forall j \in \{1,\ldots,2n_p\}, \quad \forall \boldsymbol{p} \in \{\boldsymbol{\pi}, \boldsymbol{p}^E\}: \\
& \boldsymbol{e}^L(\tau_i) - \boldsymbol{e}^U(\tau_i) \le \hat{\boldsymbol{y}}(\boldsymbol{p}^E, \tau_i) - \hat{\boldsymbol{y}}(\boldsymbol{\pi}_j, \tau_i), \\
& \hat{\boldsymbol{y}}(\boldsymbol{p}^E, \tau_i) - \hat{\boldsymbol{y}}(\boldsymbol{\pi}_j, \tau_i) \le \boldsymbol{e}^U(\tau_i) - \boldsymbol{e}^L(\tau_i), \\
& \hat{\boldsymbol{y}}(\boldsymbol{p}, \tau_i) = \boldsymbol{h}(\boldsymbol{x}(\tau_i), \boldsymbol{u}(\tau_i), \boldsymbol{p}), \\
& \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{p}), \quad \forall t \in [t_0, t_f], \\
& \boldsymbol{x}(t_0) = \boldsymbol{g}(\boldsymbol{u}(t_0), \boldsymbol{p}),
\end{aligned} \tag{16}$$

for given $\boldsymbol{u}(t)$, $\forall t \in [t_0, t_f]$ and $\boldsymbol{y}(\tau_i)$, $\forall \tau_i \in [t_0, t_f]$, $\forall i \in \{1, 2, \ldots, N\}$. This modification of the problem (12) does not require an addition of any new optimization variable and achieves a global over-approximation of the set $\mathcal{P}_e$ as it does not exclude any possible realization of the measurement error. It also provides certain robustification against the choice of $\boldsymbol{p}^E$. Note that the full robustification of the scheme w.r.t. the choice of $\boldsymbol{p}^E$, which is not considered in this study, requires some additional considerations. The expected values of parameters can, for example, be required to lie in some a priori chosen interval. Another possibility would be a combination with the sigma-point approach of Telen et al. (2014) or polynomial chaos expansion techniques (Streif et al., 2016) should some a priori assumptions be possible to make on the probability distribution of $\boldsymbol{p}^E$.

4.3. Identification of optimal experiments

Pronzato and Walter (1989) suggested V-optimal design in the context of guaranteed parameter estimation as a design that minimizes the volume vol($\mathcal{P}_e$) and showed that this is equivalent to D design for least-squares estimation in case the model is linear in parameters and $N = n_p$. The same authors suggest a $\hat{\text{V}}$-optimal design (Pronzato and Walter, 1990) for nonlinear problems where the equivalent, general reformulation in terms of classical design criteria cannot be established. These authors generally suggest linearization-based approach to the experiment design of nonlinear systems. Telen et al. (2013) proposed generalized A-, D-, and E-optimal design criteria based on the employed parametrization of the over-approximation of the GPE solution set. We use the generalized criteria, i.e., counterparts of the classical OED criteria described above, adapted to the orthotopic shape of the over-approximation of the set $\mathcal{P}_e$.

We propose solving an optimization problem to determine the optimal experiment in the context of guaranteed parameter estimation. We formulate this, using the nominal over-approximation of the GPE solution set, as a bilevel program of the form

$$\begin{aligned}
& \min_{\boldsymbol{u}(t) \in \left[\boldsymbol{u}^L(t), \boldsymbol{u}^U(t)\right]} \quad \phi_{\text{GPE}}(\boldsymbol{\pi}) \\
& \{\tau_1, \tau_2, \ldots \tau_N\} \in \left[t_0, t_f\right] \\
& \text{s.t.} \quad \max_{\pi} \sum_{j=1}^{n_p} p_j^U - p_j^L \\
& \quad \forall i \in \{1,2,\ldots,N\}, \forall j \in \{1,\ldots,2n_p\}, \forall \boldsymbol{p} \in \left\{\boldsymbol{\pi}, \boldsymbol{p}^E\right\}: \\
& \quad \boldsymbol{e}^L(\tau_i) \le \hat{\boldsymbol{y}}\left(\boldsymbol{p}^E, \tau_i\right) - \hat{\boldsymbol{y}}\left(\boldsymbol{\pi}_j, \tau_i\right) \le \boldsymbol{e}^U(\tau_i), \\
& \quad \hat{\boldsymbol{y}}(\boldsymbol{p}, \tau_i) = \boldsymbol{h}(\boldsymbol{x}(\tau_i), \boldsymbol{u}(\tau_i), \boldsymbol{p}), \\
& \quad \dot{\boldsymbol{x}}(t) = \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{p}), \quad \forall t \in \left[t_0, t_f\right], \\
& \quad \boldsymbol{x}(t_0) = \boldsymbol{g}(\boldsymbol{u}(t_0), \boldsymbol{p}).
\end{aligned} \tag{17}$$

The proposed optimization problem identifies, in the lower level, the joint-confidence region and optimizes, in the upper level, its geometry defined by the function $\phi_{\text{GPE}}(\cdot)$. In accord with the preceding discussion, the inequality constraints in (17), could be replaced with the worst-case of global formulation (see Section 4.2.2, respectively, Section 4.2.3).

The parametrization of the over-approximation of the set $\mathcal{P}_e$ makes it possible to intuitively formulate the function $\phi_{\text{GPE}}(\cdot)$ in (17) as a variant of some (classical) design of experiments. This follows from the geometrical interpretations of the classical design criteria (Franceschini and Macchietto, 2008) and is summarized and compared to experiment design for least-squares estimation in Table 1. Note that, due to the interchangeability of $\max_{\boldsymbol{\pi}} \sum_{j=1}^{n_p} (p_j^U - p_j^L)$ with $\max_{\boldsymbol{\pi}} \prod_{j=1}^{n_p} (p_j^U - p_j^L)$ in the lower-level problem, the A and D designs represent special cases of (17) as they boil down to min–max problems. Further we note that the orthotopic over-approximation bears only little to no information about the orientation of the set $\mathcal{P}_e$. Thus the standard consideration of E and modified E designs as decorrelation criteria is not truly valid here. Also, as the modified E design tries to make the orthotopic over-approximation as square (hypercubic) as possible, it might yield a GPE set with very large volume. A similar deficiency of modified E design in the context of least-squares estimation was observed by several authors (Franceschini and Macchietto, 2008). We finally note that the presented form of OED problem gives a possibility of formulating more advanced design criteria (e.g., economic-oriented design (Houska et al., 2015; Lucia and Paulen, 2014; Recker et al., 2013; Telen et al., 2016)) compared to classical OED directly, and in many cases more conveniently, in the parametric space although that, due to the nature of the proposed over-approximation (a box), some restrictions are present w.r.t. approaches that require joint-confidence region to capture the (linear) correlation between the parameters, as this information is lost in the box-shaped region.

5. Numerical implementation

As the inner maximization problem is in general nonconvex, the resulting bilevel optimization problem (17) is of very complex nature and it is, consequently, hard to solve. The special cases with min–max formulation (A and D design) are also generally not easier either. The linearization of the nonlinear constraints of problem (17) does not yield, generally, a better posed optimization problem and, moreover, might lead to deficiencies of the obtained solution.

The lower-level optimization problem can be implicitly solved by expressing the necessary (KKT) conditions for optimality as constraints. It must be, however, assured that the KKT conditions result in a globally optimal solution to the lower-level problem (Mitsos et al., 2007). This yields a reformulation into a nonconvex, in this case dynamic, optimization problem, a so-called mathematical program with equilibrium constraints (MPEC). It can often be found, however, that the reformulated problem fails to satisfy common constraint qualifications. Possible remedies to this issue were proposed in Gümüs and Floudas (2001), Vicente and Calamai (1994) and Mitsos et al. (2009), in the framework of global optimization, and in Fletcher and Leyffer (2002), Fletcher et al. (2006) and Hatz et al. (2012, 2013) where the regularization and lifting strategies were studied for well-posed reformulations of MPECs.

Beside the appealing features of the global optimization techniques for solving (17), issues arise w.r.t. computational tractability of such schemes. Note that the approach by Gümüs and Floudas (2001) requires an addition of one binary variable per each

inequality constraint in (17). This makes the resulting problem to be challenging for the state-of-the-art mixed-integer nonlinear programming solvers. We therefore restrain ourselves to the regularization approach by lifting as proposed by Hatz et al. (2013).

The problem of optimal design of experiments is then reformulated as

$$
\begin{aligned}
\min_{\boldsymbol{\pi},\, \boldsymbol{u}(t)\, \in\, [\boldsymbol{u}^L(t),\, \boldsymbol{u}^U(t)]} \ & \phi_{\mathrm{GPE}}(\boldsymbol{\pi}) + \rho\|\mathbf{1}^T\boldsymbol{w}^L\|_1 + \rho\|\mathbf{1}^T\boldsymbol{w}^U\|_1 \\
\{\tau_1, \tau_2, \ldots, \tau_N\} \,\in\, & [t_0, t_f] \\
\boldsymbol{\lambda}^L, \boldsymbol{\lambda}^U, \boldsymbol{w}^L, \boldsymbol{w}^U \geq\, & 0 \\
\text{s.t.} \quad \forall i \in \{1,2,\ldots,N\}, \quad & \forall j \in \{1,\ldots,2n_p\}, \quad \forall \boldsymbol{p} \in \{\boldsymbol{\pi}, \boldsymbol{p}^E\}: \\
0 = \ & \nabla_{\pi_j} L(\hat{\boldsymbol{y}}(\boldsymbol{\pi}_j, \boldsymbol{\tau}), \boldsymbol{u}(t), \boldsymbol{\pi}_j, \boldsymbol{\lambda}_j^L, \boldsymbol{\lambda}_j^U, \boldsymbol{w}^L, \boldsymbol{w}^U), \\
0 = \ & \mathrm{diag}\left(\boldsymbol{\lambda}_{ij}^L\right)\left(\hat{\boldsymbol{y}}(\boldsymbol{p}^E, \tau_i) - \hat{\boldsymbol{y}}(\boldsymbol{\pi}_j, \tau_i) - \boldsymbol{e}_i^L + \boldsymbol{w}_{ij}^L\right), \\
0 = \ & \mathrm{diag}\left(\boldsymbol{\lambda}_{ij}^U\right)\left(\hat{\boldsymbol{y}}(\boldsymbol{\pi}_j, \tau_i) - \hat{\boldsymbol{y}}(\boldsymbol{p}^E, \tau_i) + \boldsymbol{e}_i^U + \boldsymbol{w}_{ij}^U\right), \\
0 \leq \ & \hat{\boldsymbol{y}}(\boldsymbol{p}^E, \tau_i) - \hat{\boldsymbol{y}}(\boldsymbol{\pi}_j, \tau_i) - \boldsymbol{e}^L(\tau_i) + \boldsymbol{w}_{ij}^L, \\
0 \leq \ & \hat{\boldsymbol{y}}(\boldsymbol{\pi}_j, \tau_i) - \hat{\boldsymbol{y}}(\boldsymbol{p}^E, \tau_i) + \boldsymbol{e}^U(\tau_i) + \boldsymbol{w}_{ij}^U, \\
\hat{\boldsymbol{y}}(\boldsymbol{p}, \tau_i) = \ & \boldsymbol{h}(\boldsymbol{x}(\tau_i), \boldsymbol{u}(\tau_i), \boldsymbol{p}), \\
\dot{\boldsymbol{x}}(t) = \ & \boldsymbol{f}(\boldsymbol{x}(t), \boldsymbol{u}(t), \boldsymbol{p}), \quad \forall t \in [t_0, t_f], \\
\boldsymbol{x}(t_0) = \ & \boldsymbol{g}(\boldsymbol{u}(t_0), \boldsymbol{p}),
\end{aligned}
\tag{18}
$$

where $L(\cdot)$ denotes the Lagrangian of the maximization problem in (17), $\boldsymbol{\lambda}^L$ and $\boldsymbol{\lambda}^U$ stand, respectively, for a $2Nn_yn_p$-dimensional vector of Lagrange multipliers, $\boldsymbol{w}^L$ and $\boldsymbol{w}^U$ are $2Nn_yn_p$-dimensional vectors of lifting (slack) variables and a positive constant ρ is chosen such that the contribution of the 1-norm terms in the objective function is insignificant yet the lifting variables are zero at the optimum. In total, the reformulated optimization problem in (18) has $2n_p^2 + N(n_u+1) + 8Nn_yn_p$ decision variables, $8Nn_yn_p$ inequality constraints and $4Nn_yn_p + 2n_p{}^2$ equality constraints. The presented dynamic optimization problem can be solved using various well-known algorithms such as sequential (Sargent, 2000), simultaneous (Biegler, 1984) or multiple-shooting (Bock and Plitt, 1984) approach.

Because of the nonconvex nature of the inner problem, there could exist several points (local minima and maxima, and saddle points) where the optimality conditions of the inner problem are satisfied. This situation requires a special treatment. For instance, the satisfaction of the second-order (sufficient) conditions could be enforced by appropriate inequality constraints. This approach, however, only guarantees reaching local optimality of the inner problem, which results in infeasibility of the outer problem. As a remedy, upon convergence of the NLP solver, which solves the resulting lifted MPEC (18), we check whether the resulting over-approximation of the set $\mathcal{P}_e$ is correct. This is done by fine over-approximation of (8) using the set-inversion approach (Jaulin and Walter, 1993; Paulen et al., 2016). If this test is positive, we can claim finding a locally optimal solution to the problem (17). Should this not be the case, a practical approach can be used to remedy this issue by the initialization of the MPEC using the results of a set-inversion algorithm. This might be repeated iteratively and can in many instances lead to convergence of the MPEC to a local optimum of the bilevel program. We discuss these issues further in the discussion section of this paper.

A scheme with improved computational tractability might be used, as proposed by Pronzato and Walter (1989), using an approximate (linearized) form of the problem (17) where the lower-level constraints of the problem are approximated as

$$
\boldsymbol{e}^L(\tau_i) \leq \frac{\partial \hat{\boldsymbol{y}}(\boldsymbol{p}, \tau_i)}{\partial \boldsymbol{p}}\Big|_{\boldsymbol{p}^E} (\boldsymbol{\pi}_j - \boldsymbol{p}^E) \leq \boldsymbol{e}^U(\tau_i). \tag{19}
$$

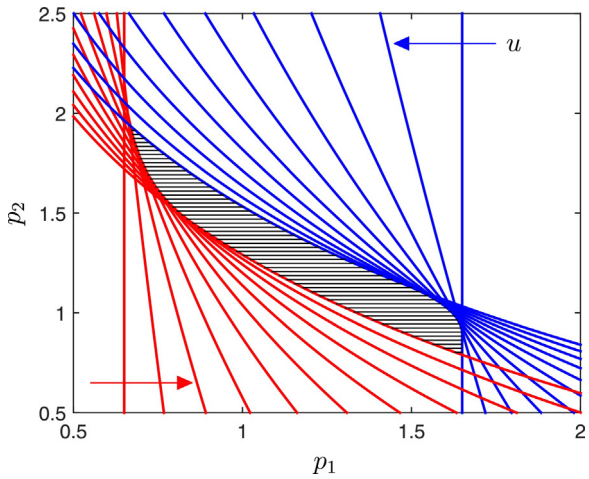


Fig. 4. Solution set of guaranteed parameter estimation with the corresponding constraints for eleven experiments with input equidistantly varied in its range and zero realization of the measurement noise.

This approximation holds if the set $\mathcal{P}_e$ is small enough (Pronzato and Walter, 1989). We will examine the properties of this reformulation in the following sections.

6. A case study in static optimal experiment design

We use a case study taken from Jaulin and Walter (1993) to demonstrate the optimal experiment design in the context of guaranteed parameter estimation for a dynamic time-invariant system. The problem represents fitting of a generic (e.g., thermodynamic or reaction-kinetic) correlation to the experimental data using an exponential dependence. The problem is to identify N optimal experiments that differ in inputs $u_\tau \in [0,1], \forall\, \tau \in \{1, 2\ldots, N\}$. The identified system can be described by the following mathematical model

$$
x_\tau = p_1 \exp(p_2 u_\tau), \tag{20a}
$$

$$
\hat{y}_\tau = x_\tau, \tag{20b}
$$

where p_1 and p_2 are the unknown parameters to be identified. The values for the expected parameters $\boldsymbol{p}^E$ are chosen to be $p_1^E = 1.15$ and $p_2^E = 1.28$. The upper and lower bounds of the measurement error are $[e^L(\tau_i), e^U(\tau_i)] = [-0.5, 0.5], \forall\, i \in \{1, 2\ldots, N\}$.

In order to gain some insight into this problem, let us first consider a series of eleven experiments ($N=11$) where the degree of freedom is varied equidistantly in its admissible range, i.e., $u_{\tau_i} = (i-1)/(N-1), \forall i \in \{1, 2\ldots, N\}$. We can plot the constraints set of (9). This is shown in Fig. 4 for the nominal assumption on the realization of the measurement noise, i.e., $e(\tau_i) = 0, \forall\, i \in \{1, 2\ldots, N\}$. Note that, in order to allow for comparisons, the GPE solution sets are always shown in this study with the zero realization of the measurement noise, unless specifically mentioned otherwise.

After observing the plot, one can conjecture that the series of optimal experiments should include those with (close to) extremal values of u as low values of u contribute more to better estimation of p_1 and, vice versa, the high values of u reveal more information on p_2.

To show the effect of the measurement noise on the GPE solution set, we determine the worst-case over-approximation of the GPE solution set that identifies the noise sequence $(\boldsymbol{e} = (-0.3434, -0.3808, -0.4288, -0.4731, -0.5000\,\mathbf{1}_7^T)^T$ where $\mathbf{1}_n = (1, \ldots, 1)^T \in \mathbb{R}^n)$ which yields the biggest volume of the

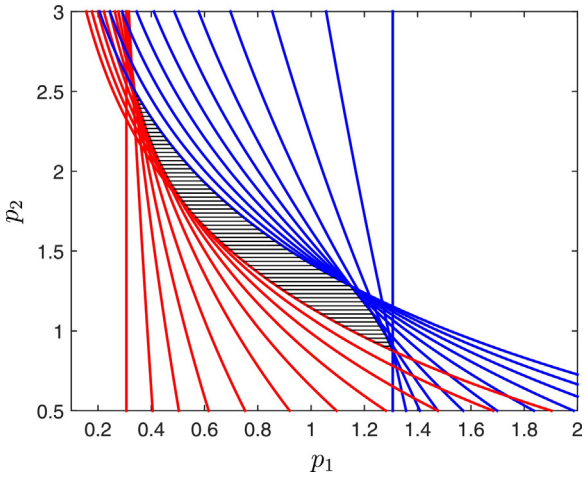


Fig. 5. Solution set of guaranteed parameter estimation with the corresponding constraints for eleven experiments with input equidistantly varied and worst-case realization of the measurement noise.

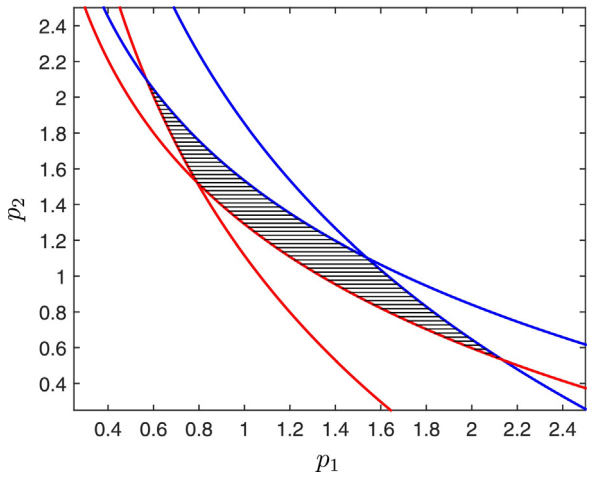


Fig. 6. Solution set of guaranteed parameter estimation with the corresponding constraints for modified E-optimal experiments ($N=2$) identified for nominal over-approximation of the GPE solution set.

orthotopic over-approximation of the set $\mathcal{P}_e$. This set is shown in Fig. 5.

We use the solver BARON (Tawarmalani and Sahinidis, 2005) to solve the optimization problems (7) and (18) for their various setups given in Table 1. This is done with default settings and only to local optimality in most cases as a compromise between CPU time and the quality of the obtained solution. We note that due to complex nature of the problem (18), the solver is generally not able to close the optimality gap within a reasonable threshold.

6.1. *Effect of the design criterion, the number of measurements, and the type of confidence region over-approximation on the optimal experiment*

We first compare the results of OED obtained for different design criteria. The results are dependent on the number of experiments (obtained measurements), as might be expected, and on the type of over-approximation of the GPE solution set. As the aforementioned problem analysis shows, the number of experiments to conduct is $N \geq 2$ in order to reach any reasonable estimation result. This is, of course, in line with the general consensus $N \geq n_p$ for parameter estimation (Bates and Watts, 1988; Pronzato and Walter, 1989).

As mentioned above the modified E design might yield a GPE set with very large volume as its focus is solely on the shape of the resulting confidence region. Moreover, due to nonlinearity and (possibly) induced irregularity of the GPE solution set, even in the case when the enclosing box is of square shape, the decorrelating effect of this criterion might not materialize on the actual GPE solution set. This is shown in Fig. 6 where the resulting constraint set of the modified E-optimal experiment for $N=2$ is shown.

For this reason we exclude modified E criterion from further considerations in this study.

The results of optimal experiment design for other criteria are summarized in Table 2. Here we use nominal over-approximation of the GPE solution set. It is found that for $N=2$, A and D designs coincide with $\boldsymbol{\tau}^*=(0.0540,\ 1)^T$. The resulting constraint set is shown in Fig. 7.

In case of three or more experiments the A, D and E designs coincide. The constraints set for the resulting $\boldsymbol{\tau}^*=(0, 0.1951, 1)^T$ is shown in Fig. 8. Our results show that all considered design criteria reach the best values for $N=3$ and the further increase of number of experiments does not improve the measures of the resulting GPE solution sets. Similar properties of the experiment design in

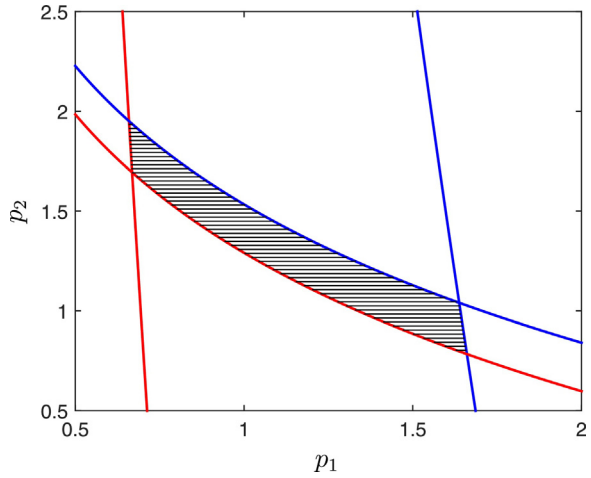


Fig. 7. Solution set of guaranteed parameter estimation with the corresponding constraints for two A- and D-optimal experiments identified for nominal over-approximation of the GPE solution set.

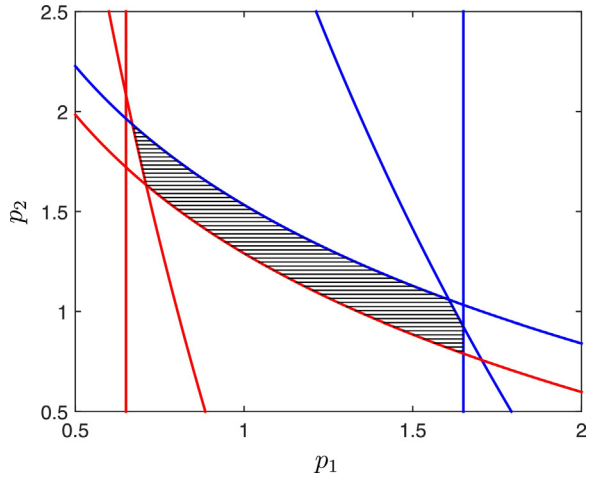


Fig. 8. Solution set of guaranteed parameter estimation with the corresponding constraints for optimal experiment design of three experiments.

Table 2
Comparison of optimal experiment designs for various design criteria and number of experiments N in case of nominal over-approximation of the GPE solution set.

Criterion	$N=2$			$N\geq 3$
	A design	D design	E design	A, D and E design
A	2.1684	2.1684	2.1714	2.1263
D	1.1686	1.1686	1.1724	1.1235
E	1.1670	1.1670	1.1655	1.1455
$\boldsymbol{\tau}^*$	$(0.0540, 1)^T$	$(0.0540, 1)^T$	$(0.0908, 1)^T$	$(0, 0.1951, 1)^T$

Table 3
Comparison of A-optimal experiment designs for $N=3$ experiments with nominal, worst-case and global over-approximation of the GPE solution set in terms of optimal experiment setup and optimality w.r.t. to random realization of measurement noise.

Over-approximation type	Nominal	Worst-case	Global
τ_2^*	0.1950	0.4513	0.3542
Optimality	63.9%	19.1%	17.9%

the context of guaranteed parameter estimation were discussed in Pronzato and Walter (1990).

The comparison of the optimal experiment designs that use different types of over-approximation of the set $\mathcal{P}_e$ (see Section 4.2) is detailed in Table 3. We show the comparison for A-optimal design and $N=3$, however, similar results are obtained for other design criteria and other values of N. We compare the value of A criterion (perimeter of the orthotopic over-approximation of the GPE solution set) for 1000 random, uniformly distributed, realizations of measurement noise. The results show that the resulting designs give two mutually same experiments and they differ in one experiment only. A-optimal design obtained with nominal over-approximation of the GPE solution set results in the best value of A criterion in roughly 60% of cases while the designs that use worst-case and global over-approximations are both best in approximately 20% of the random noise realizations. These results show the robustification of the optimal experiment design against the realization of the measurement noise. This can be achieved by fixing $\tau_1=0$ and $\tau_2=1$ (or at $\tau_{\max}$) and using up to three more experiments with the design variables τ_3, τ_4, ... taken from the aforementioned optimal designs.

6.2. Comparison to classical optimal experiment design

We compare the obtained results, presented in the previous section, to the results of the optimal experiment design for least-squares estimation. The influence of the design variables $\boldsymbol{\tau}$ on D-optimal design can be seen in Fig. 9 where the D-optimality criterion is plotted against the choice of how to perform two experiments ($N=2$). The plot shows negative of the determinant of the inverse of **FIM** which is equivalent in this case to D-optimality criterion and gives better visualization of the actual objective function. It is clear here that the minimum is achieved by picking $\tau_1^* = 1$ and $\tau_2^* = 1-(1/p_2)$ which is equivalent to the design for a similar problem obtained by Pronzato and Walter (1990). If we now consider to design three experiments and fix $\tau_1=1$, we can observe, in Fig. 10 (equivalent to Fig. 9 for $N=3$) that the vector $\boldsymbol{\tau}=(1-(1/p_2),\ 1)^T$ is again optimal. Summarizing, the classical optimal experiment design will result, for any N, in a repetition of the values from $\boldsymbol{\tau}^*=(1-(1/p_2),\ 1)^T$.

Other design criteria yield similar results and can be characterized as $\boldsymbol{\tau}*=(\tau_1^*,\ 1)^T$ where a great similarity is observed between A- and E-optimal design, unlike in the case of OED for guaranteed parameter estimation. All these presented observations show the structural difference between the classical OED and the OED proposed here for guaranteed parameter estimation.

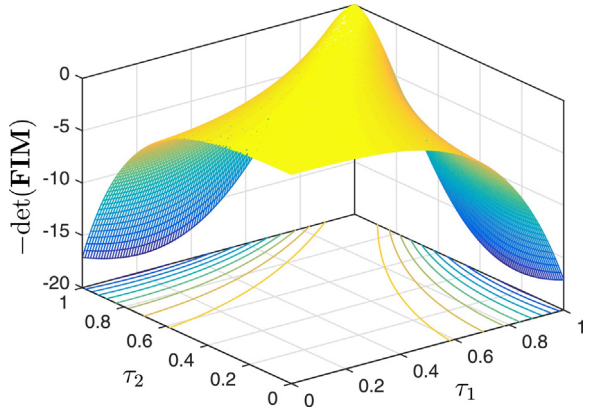


Fig. 9. Illustration of the objective function of classical D-optimal design for the design of two experiments.

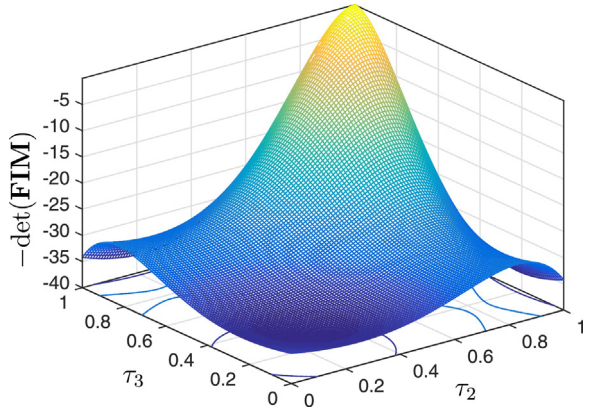


Fig. 10. Illustration of the objective function of classical D-optimal design for the design of three experiments (τ_1 is kept at the value of 1).

Fig. 11 compares the joint-confidence regions of least-squares estimates ($\alpha=0.95$) and the GPE solution sets for $N=2$ and $N=3$. The regions are obtained under application of D-optimal experiments obtained for classical OED and for OED in the context of guaranteed parameter estimation. For the fairness of comparison, the realization of the measurement noise is taken as zero. We can observe that despite their linear character, the joint-confidence ellipsoids capture the general shape of the GPE solution sets quite well. On the other hand, the joint-confidence ellipsoid for $N=2$ over-estimates the bounds (confidence intervals) of p_1 and under-estimates the bounds of p_2. In case $N=3$, the joint-confidence ellipsoid under-estimates the bounds of both p_1 and p_2. Similar observations can be made for $\alpha=0.99$. One should, however, bear in mind that the joint-confidence analysis of least-squares estimates holds only asymptotically so the presented analysis should only be used to visualize the differences between the joint-confidence regions. We do not intend to assess the quality of the resulting confidence intervals obtained by the presented methods.

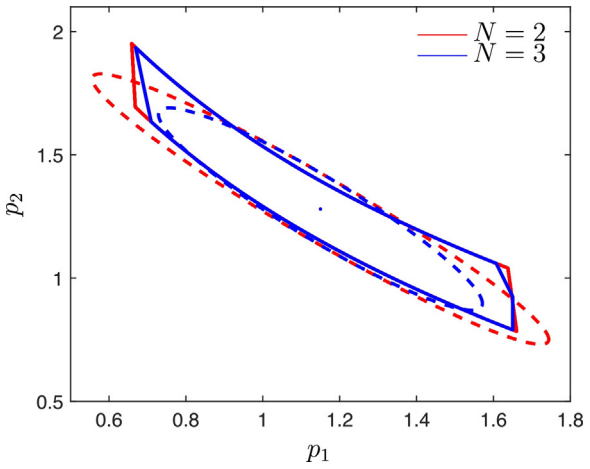


Fig. 11. Comparison of joint-confidence regions of least-squares estimates (dashed line) and of guaranteed parameter estimation (solid line) obtained under corresponding two and three optimal experiments.

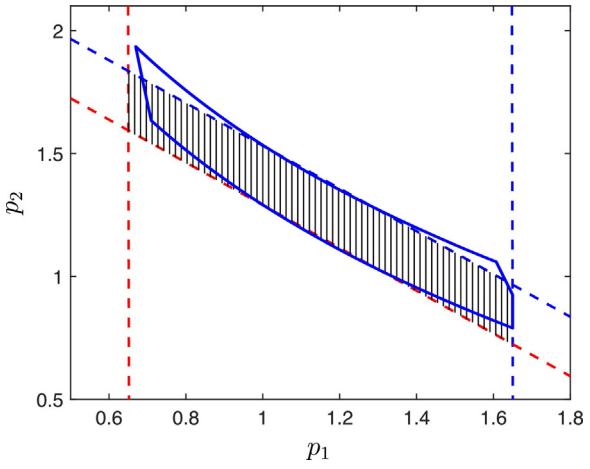


Fig. 12. Solution set of guaranteed parameter estimation with the corresponding constraints with linearized model and two corresponding optimal experiments (shaded region) compared to the GPE solution set resulting from optimal experiment of the original model.

6.3. Assessment of linearization-based identification of optimal experiments

Finally we evaluate the approximate solution to the problem (17) that is achieved after linearization of the nonlinear constraints of the problem. The resulting bilinear bilevel problem can be solved quite efficiently to global optimality by the BARON solver. On the other hand, the obtained solution does not coincide with the solution obtained for the original problem. This can be seen in Fig. 12 where we compare the GPE solution sets obtained for nonlinear and linearized model resulting from corresponding optimal experiments. Similarly to the case of classical experiment design, the linearization-based optimal experiment design results in a set of two measurements to be taken, $\boldsymbol{\tau}^* = (0.0012,\ 1.0000)^T$. We can observe that the resulting linearization-based GPE solution set differs from the true GPE solution set similarly to the joint-confidence ellipsoids discussed above. Our conclusion is that, despite the computational benefits of this approach, the linearization-based optimal experiment design yields suboptimal results of the problem (17).

7. A case study in dynamic optimal experiment design

This case study considers the Lotka–Volterra fishing problem (Telen et al., 2012; Sager, 2013), a system relevant to chemical engineering (Evans and Findley, 2000), and demonstrates the optimal experiment design in the context of guaranteed parameter estimation on a time-invariant system. The mathematical model of the Lotka–Volterra fishing problem reads as

$$\dot{x}_1(t) = x_1(t) - p_1 x_1(t) x_2(t) - 0.4 x_1(t) u(t), \qquad x_1(0) = 0.5, \tag{21a}$$

$$\dot{x}_2(t) = -x_2(t) + p_2 x_1(t) x_2(t) - 0.2 x_2(t) u(t), \qquad x_2(0) = 0.7, \tag{21b}$$

$$\hat{\boldsymbol{y}}(\boldsymbol{p}, \tau_i) = (x_1(\tau_i), x_2(\tau_i))^T, \qquad \forall i \in \{1, 2, \ldots, N\}, \tag{21c}$$

where x_1 and x_2 are the populations of the prey and of the predator, respectively. The manipulated variable $u(t)$ is restricted to take binary values. The true values of unknown model parameters $\boldsymbol{p} = (p_1, p_2)^T$ are $\boldsymbol{p}^E = (1, 1)^T$.

Based on the insights provided by Telen et al. (2012) and Sager (2013), the problem of designing an optimal experiment can be simplified to identification of the N switching times $\tau_i\ \forall\ i \in \{1, 2, \ldots, N\}$ of the sequence $u(t) = \{1, 0, 1, 0, \ldots\}$ where $\tau_{i+1} - \tau_i \geq 0.2$ and $\tau_N \leq 12$. It is assumed that the measurements are corrupted with measurement noise such that $[\boldsymbol{e}^L(\tau_i), \boldsymbol{e}^U(\tau_i)] = ([-0.3, 0.3], [-0.3, 0.3])^T$ and are available only at the time of the switch of $u(t)$.

Optimal experiment design for the guaranteed parameter estimation problem is formulated for all the three types of the over-approximation of GPE joint-confidence region (see Section 4.2). The resulting bilevel optimization problem is reformulated and regularized using the methods described in Section 5, which yields a well-posed nonlinear nonconvex optimization problem. The problem (18) is solved by multiple-shooting algorithm (Bock and Plitt, 1984) using IPOPT (Wächter and Biegler, 2006) via CasADi (Andersson et al., 2012).

7.1. Effect of the design criterion, the number of measurements, and the type of confidence region over-approximation on the optimal experiment

In order to obtain a good-quality estimates of the identifiable parameters of a linear model, n_p measurements are required (Bates and Watts, 1988; Pronzato and Walter, 1989). As evidenced in the previous case study, nonlinearity of the model might increase this number. Therefore, the value of N is selected to be $N \geq 2$ in this case study (note that if $N = 2$, four measurements $2n_y$ are obtained).

The problem (17) is first solved for $N = 2$ and A-, D- and E-optimal designs using the corresponding objective functions shown in Table 1. Fig. 13 shows the resulting optimal-experiment control trajectories of $u(t)$. The switching (measurement) instants are highlighted using square symbols. Similarly to the results obtained in the static case study, the optimal experiments coincide for the A and D designs. One can also observe a significant difference between A/D designs and E design. The GPE solution set corresponding to the A/D design is shown in Fig. 14. One can clearly observe that not all measurements contributed evenly to the identification of the parameters. While the measurements $x_1(\tau_1)$ and $x_2(\tau_1)$ do not falsify (almost) entire region $p_1 \times p_2 = [0.8, 1.2] \times [0.8, 1.2]$, the measurements taken in the second time instant (τ_2) determine the shape of the set $\mathcal{P}_e$. It is an interesting observation that the measurement $x_2(\tau_2)$ contributes the most to the falsification of values of p_1 but does not falsify any value of p_2 in the depicted range. This also shows how beneficial it is for the precision of the parameter estimation that both states are measured. If only x_1 is measured, Fig. 14 suggests the shape of the GPE solution set (given by $x_1(\tau_1)$ and $x_1(\tau_2)$) that is roughly three times larger.

Table 4 shows values of the objective functions for A, D and E designs obtained for the optimal switching times $\boldsymbol{\tau}^*$ from the

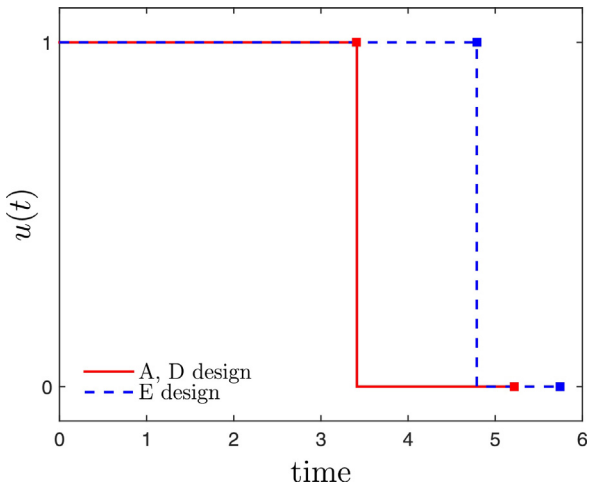


Fig. 13. Comparison of control input from A-, D- and E-optimal experiment designs for the guaranteed parameter estimation using nominal over-approximation of the GPE solution set.

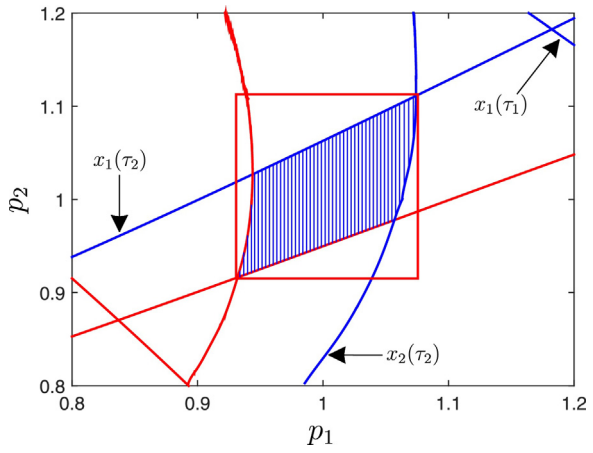


Fig. 14. Solution set of guaranteed parameter estimation with the corresponding constraints for optimal experiment design for $\boldsymbol{\tau}^*$ from D-optimal experiment design for GPE using nominal over-approximation of the GPE solution set.

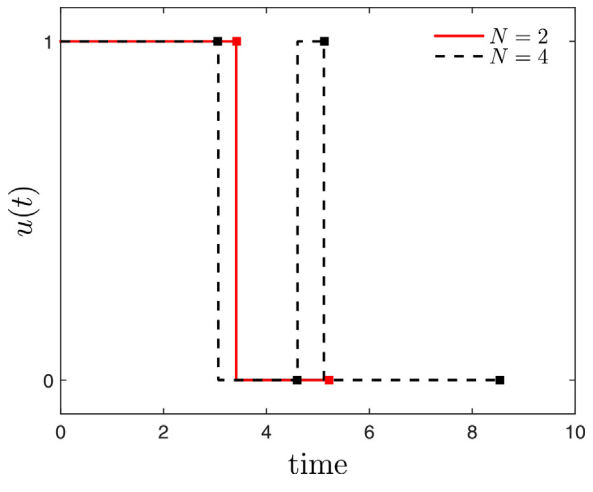


Fig. 15. Comparison of A-optimal experiments in input space with different number of measurements using nominal over-approximation of the GPE solution set.

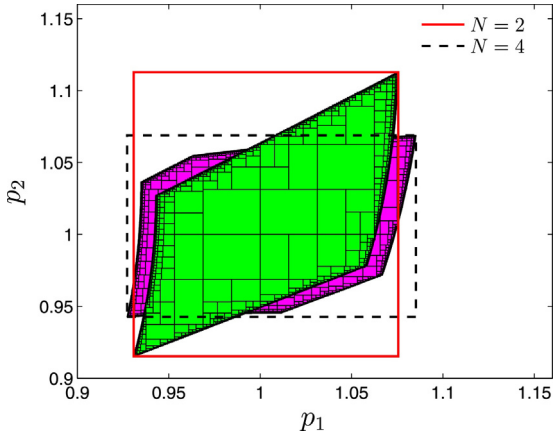


Fig. 16. Comparison of the GPE solution sets obtained in A-optimal experiments with different number of measurements.

corresponding optimal experiment design. The difference of A/D design to E design is not very large. However this might be expected after revisiting Fig. 14. Seen as a movement from A/D design, the E-optimal experiment design apparently compromises the minimization of $(p_2^U - p_2^L)$ w.r.t. to increase of $(p_1^U - p_1^L)$.

Let us now study the effect of the number of measurements on the quality of guaranteed parameter estimates. The optimal experiment design problem (17) is solved for A-design criterion using nominal over-approximation for $N=4$. The obtained results are compared with the case $N=2$ in Fig. 15 (optimal control inputs) and in Fig. 16 (GPE solution sets with the corresponding nominal over-approximating boxes). A similarity of the control inputs can only be observed w.r.t. the first switching time τ_1. As the control input trajectories significantly differ for $t > \tau_1$ and since it was observed, in the case $N=2$, that the measurements gathered at this time do not contribute significantly to the resulting GPE solution set, a significant difference between the GPE solution sets can be expected for $N=2$ and $N=4$ as evidenced in Fig. 16. Here we can clearly observe the improvement reached in the A-optimal criterion. Yet it can be seen that the volumes of the sets $\mathcal{P}_e$ do not differ much and, even, that the volume of GPE solution set obtained for $N=4$ is only slightly larger. It can be concluded that the A- nor D- optimal design are V-optimal, i.e., the designs do not generally achieve a minimal volume of the GPE solution set. This is in line with our discussion above (Section 4.3) and shows that, in applications where only bounds of the unknown parameters are or primary concern (robust design or robust control), the methodology proposed here would be favored over true V-optimal design.

Table 4
Comparison of optimal experiment designs for various design criteria calculated using nominal over-approximation of the GPE solution set.

Criterion	$N=2$			$N=4$
	A design	D design	E design	A design
A	0.3425	0.3425	0.3551	0.2844
D	0.0286	0.0286	0.0315	0.0200
E	0.1977	0.1977	0.1824	0.1581
$\boldsymbol{\tau}^*$	$\begin{pmatrix} 3.413 \\ 5.223 \end{pmatrix}$	$\begin{pmatrix} 3.413 \\ 5.223 \end{pmatrix}$	$\begin{pmatrix} 4.788 \\ 5.740 \end{pmatrix}$	$(3.0641, 4.6041, 5.1163, 8.5347)^T$

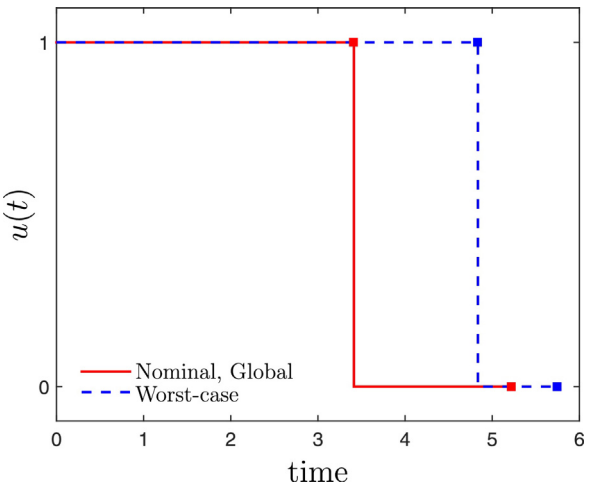


Fig. 17. Comparison of the solutions to A-optimal experiment design using different over-approximations of the GPE solution set.

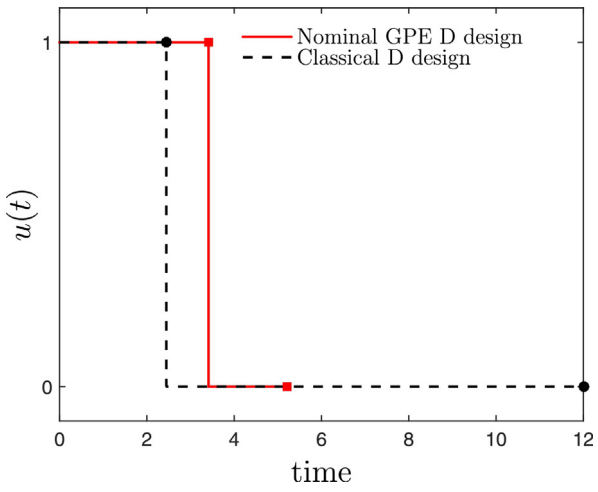


Fig. 18. Comparison of control inputs from classical D-optimal design and D-optimal experiment design for guaranteed parameter estimation using nominal over approximation.

Table 5
Comparison of A-optimal experiment designs for $N=2$ experiments with nominal, worst-case and global over-approximation of the GPE solution set in terms of optimal experiment set up and their optimality for 1000 random realizations of measurement noise.

Over-approximation type	Nominal and global	Worst-case
$\boldsymbol{\tau}^*$	$(3.413, 5.223)^T$	$(4.836, 5.749)^T$
Optimality	49.4%	50.6%

In Fig. 17, we show the optimal trajectory of the binary control variable $u(t)$ and we highlight its switching times $\boldsymbol{\tau}^*$. This is obtained by solving the A-optimal experiment design for guaranteed parameter estimation problem using various over-approximations of the GPE solution set. The same solutions as in the case of nominal over-approximation of the GPE solution sets are obtained for the *global* counterpart for all the considered (A-, D- and E-optimal) designs. This suggests some linearly proportional scaling of the constraints (cuts obtained by the measurements) in the parametric space. The use of worst-case over-approximation yields a significantly different control inputs as illustrated in Fig. 17 for A design. In this case, the control input is modified such that the measurements in τ_1, notice that only $\tau_1{}^*$ varies significantly between these designs, will bring more information about the parameters.

In Table 5, we show a comparison for the value of A-optimal criterion for 1000 random (sampled from uniform distribution) instances of measurement noise. The design obtained for A-optimal experiment using nominal and global over-approximation of the set $\mathcal{P}_e$ results in the better value for A-design criterion roughly in 50% of the cases while the design that uses worst-case over-approximation is the better for the rest of the random noise realizations. This shows that the robustification of the optimal experimentation against the realization of measurement noise can be achieved by conducting two dynamic experiments, each using a different switching sequence from Table 5.

7.2. Comparison with classical optimal experiment design

A similar comparison to the case study in static experiment design (Section 6) is performed. Fig. 18 compares the optimal control inputs obtained in classical D-optimal experiment design and in D-optimal experiment design for GPE using nominal over-approximation of the GPE solution set for $N=2$. We see that the optimal switching times $\boldsymbol{\tau}^*$ differ significantly between these two designs. Similar differences are observed when other design criteria are considered.

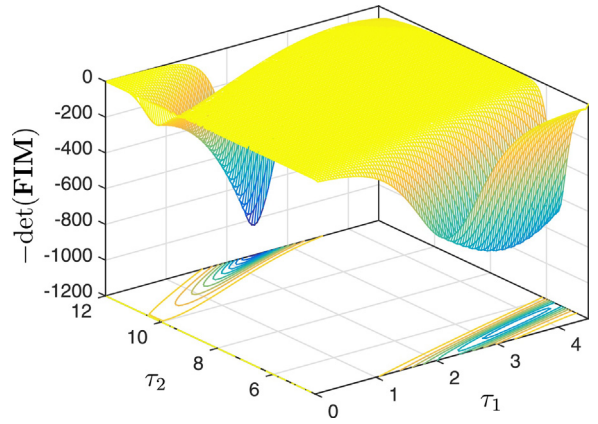


Fig. 19. Illustration of the objective function of classical D-optimal design for $N=2$ time steps.

The reason for these differences can be seen in Fig. 19, where we show $-\det(\mathbf{FIM})$ (equivalent to the determinant of the inverse of **FIM**) as function of $\boldsymbol{\tau}$ for $N=2$. The values of τ_1 and τ_2 are selected such as to cover the optimal designs (shown in Fig. 18) and to show both local optima of the function for any admissible $\boldsymbol{\tau}$. Clearly, the global optimum (classical D-optimal experiment design) is obtained for $\boldsymbol{\tau} \approx (2.2, 12)^T$ which exactly coincides with control input (black dashed curve) from Fig. 18. If this optimal experiment is used and guaranteed parameter estimation is performed, the resulting solution set is slightly bigger volume-wise and significantly worse in terms of A-optimal criterion as can be seen when Figs. 20 and 14 are compared. This represents an interesting situation where global optimizer of classical A-optimal experiment design is a local optimizer of the A-optimal experiment in the context of OED and vice versa. The situation is different in case of D and E designs where the optimal control trajectories are very similar for both classical design and design for GPE.

Fig. 21 compares the joint-confidence regions of least-squares estimates ($\alpha=0.95$) and the GPE solution sets for $N=2$ resulting from the discussed D-optimal designs. For the fairness of the comparison, the realization of the measurement noise is taken as zero. Similarly to the case study in static experiment design, we can observe that despite their linear character, the joint-confidence ellipsoids capture the general shape of the GPE solution sets quite well. We can also clearly see that while very good approximation is

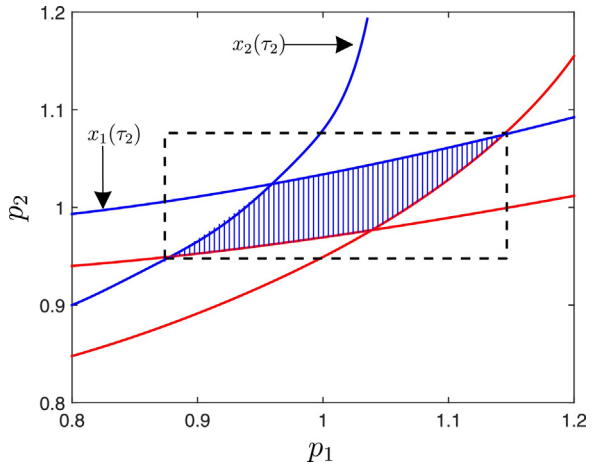


Fig. 20. Solution set of guaranteed parameter estimation with the corresponding constraints for optimal experiment design for τ^* from classical D design.

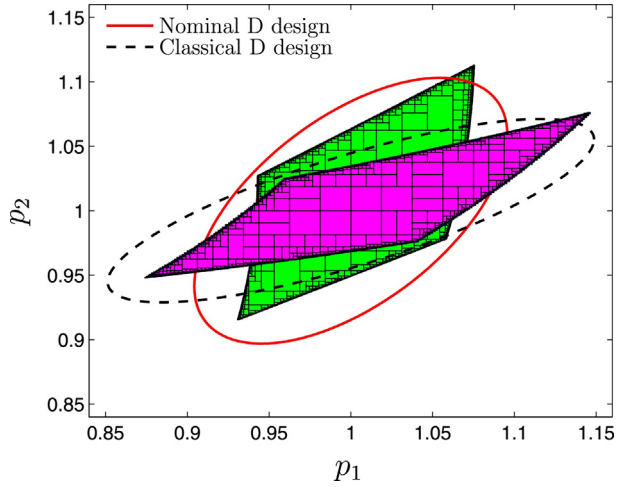


Fig. 21. Comparison of joint-confidence regions ($\alpha = 0.95$) for control input from classical D design and D-optimal experiment design for GPE using nominal over-approximation of the GPE solution set. The green and magenta sets represent the GPE solution set for the same control inputs.

achieved in the direction of the minor axes of the ellipsoids but this is rather unsatisfactory along the major axes. It can be concluded that a very fine approximation of the GPE solution set cannot be expected from the joint-confidence ellipse. As mentioned in the previous case study, we do not intend comparing the quality of the classical OED with our method and the comparison serves for illustration of the differences between classical OED and OED in the context of guaranteed parameter estimation.

7.3. Assessment of linearization-based identification of optimal experiments

An approximate (linearized) form of the OED for guaranteed parameter estimation problem (17) is formulated using linearized constraints (19) and solved for the A-optimal experiment design problem for $N = 2$ using nominal over-approximation of the GPE solution set.

The differences between the solutions to fully nonlinear and linearized problem of optimal experiment design are shown in Figs. 22 (input space) and 23 (parametric space). It must be first noted that if the constraints (measurement cuts) are simply linearized and A-optimal design obtained for a fully nonlinear problem is used, the resulting constraints intersection region

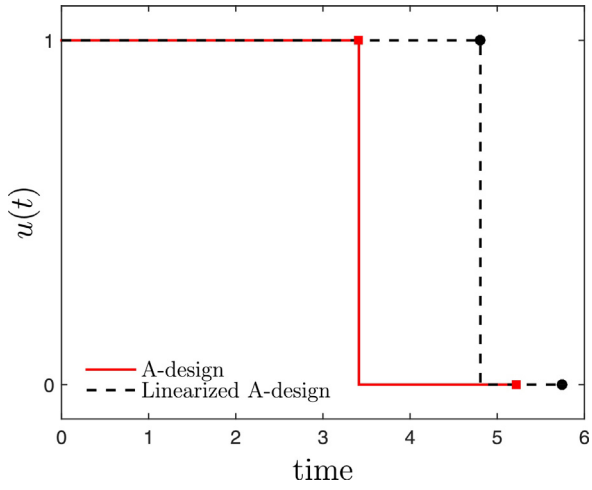


Fig. 22. Illustration of the effect of linearization on the solution of A-optimal experiment design using nominal over-approximation of the GPE solution set.

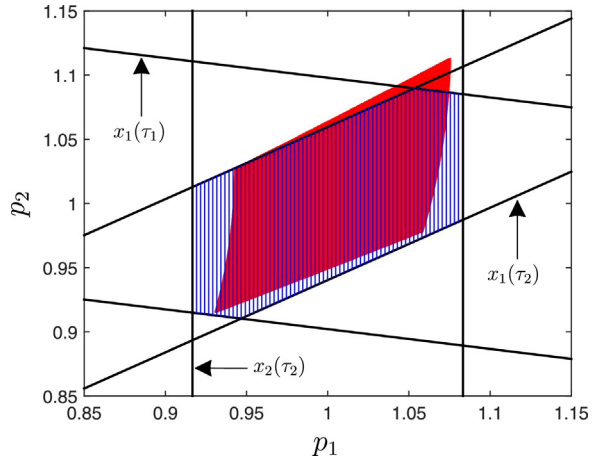


Fig. 23. Constraint set of guaranteed parameter estimation with linearized model and two corresponding optimal experiments (shaded region) compared to the GPE solution set resulting from optimal experiment of the original model.

(counterpart to the GPE solution set) approximates the actual GPE solution set (red region in Fig. 23) quite well. However, when the A-optimal design is determined that considers the linearized constraints instead of nonlinear ones, the optimal experiment identifies that some information can be gained from the measurements at τ_1, unlike the A-optimal design. This is the reason why the solution to linearization-based optimal experiment design resembles the solution obtained for the worst-case over-approximation of the GPE solution set (Fig. 17) and results in a loss of optimality, although only minor ($\approx$1%) in this case, compared to the original A-optimal experiment design problem. Similar results are achieved for other optimal experiment design problems.

8. Discussion

The results obtained and analyzed in the presented case studies provide some interesting insights into the problem of optimal experiment design in the context of GPE for nonlinear problems. It is obvious that this problem resembles the classical OED in many aspects such as that the same number of measurements (un)falsifies roughly the same portion of the parametric space or that the fundamental nature of the optimal experiments is similar. The later is evidenced by the general resemblance of the designed optimal experiments despite the differences that were

clearly pointed and that partially originate, on one hand, from the linear nature of the classical optimal experiment design and, on the other hand, from the different expectations of the realization of measurement noise.

The proposed methodology is built upon the assumptions that the solution set of the guaranteed parameter estimation is connected and that its bounding (tight over-approximation) using a box captures the general attributes of this nonconvex set. The first assumption is apparently violated in cases of identifiability problems of some parameters where no a priori information, which could exclude parts of parametric space that imply non-unique estimates, is available. These cases require specialized approaches of experiment designers, e.g., Müller et al. (2014), and are not considered in this study. They are, however, of a concern of modelers on daily basis and, thus, require attention and optional extension of the here presented methodology in the future research. The second main assumption is apparently justified in the applications, such as robust design and control, where the knowledge of joint-confidence intervals (bounds on the values of parameters) is paramount. As mentioned above the over-approximation of the GPE solution set does not necessarily capture the information about the linear correlation of parameters. This means that the decorrelation cannot be achieved in general using the methodology applied in this study and requires a fundamentally different approach. One could think here about constructing an ellipsoid that is centered at the Chebyshev center of the polytope, which is obtained using a linearization-based approach to GPE, or some nonlinear equivalent to this center, where the shape of this ellipsoid approximates the boundary of the GPE solution set in the best possible way. Should such construction be feasible, a decorrelating design of experiments would be similar to E-optimal design from classical OED. We note, however, that the use of ellipsoid brings necessarily a loose over-approximation of the parametric bounds (as noted e.g., in Walter and Piet-Lahanier, 1990). Another possible extension of this work is to consider other types of over-approximation of the set $\mathcal{P}_e$, e.g., using a zonotope or a polytope, giving an advantage of better (tighter) over-approximation. However, bounding based on either ellipsoid or polytope requires application of recursive algorithms (Pronzato and Walter, 1994), which, on one hand, might turn out to be computationally demanding and, on the other hand, would need to be taken into account by the proposed methodology which would require its major restructuring.

Another critical assumption of this paper is the availability and accuracy of the expected value of the parameters. This is not a novel issue in the model-based experiment design and several approaches, which could extend the present study, were developed including, for example, robustification of the experiment design (Telen et al., 2014) or subsequent optimal experiment design and parameter estimation (Baran et al., 2012). An interesting path would be to adopt the approach of Kurzhanski and Varaiya (2011) for solving the problem of robust experiment design. We would also like to mention in this context that if one considers the problem of the optimal experiment design to answer the question "What is the maximal possible improvement of the accuracy of parameters of the given model, given the accuracy of the sensors (data)?", then the assumption of knowing its true parameter is naturally a part of the problem instance. A question of model selection and model quality checking against experimental data arises here as well. It would be interesting to tackle the problem of model discrimination in the context of guaranteed parameter estimation using the approach similar to the one presented in this paper.

The optimal experiment design in the context of GPE is proposed to be found via a nonlinear nonconvex bilevel optimization problem with considerable number of decision variables and constraints (complexity aspects of the optimization problem were discussed above). In our computational experience, with the case studies treated in this contribution and in Gottu Mukkula and Paulen (2016a,b), the local solvers, which exploit sparsity of the problem (this appears because of $2n_p$ model instances) and use first- and second-order derivatives, are able to find a candidate (solution) point in a reasonable time, in order of minutes. Here the regularization of the problem (see (18)) helps greatly in further reduction of the CPU time. The experience with global solvers is worse. The globally optimal solution can be reached for lower-level-linearized form of the problem (17) (note the sub-optimality of this approach w.r.t. the original problem is discussed above) and for small-scale designs (N is small) of the static optimal experiment design. In other cases, the optimality gap is rarely closed but we admit that this analysis might be biased, as we did not perform any tuning of the global solvers, i.e., standard settings were used and no sophisticated initialization of the range of decisions variables was conducted, and we always restricted the solver to terminate within couple of hours.

Regarding its solvability, the nonconvexity of the inner problem of (17) stands for the most concerning issue in the proposed methodology as there exist currently, according to our knowledge, no mature software tools to tackle the problem of nonconvex dynamic bilevel programming. It is, of course, not the nonconvexity as such that is of concern but the existence of local optima (minima or maxima) and saddle points of the inner problem. In the case study on the static experiment design, despite the nonconvexity of the problem, it was always possible to identify the global optimum of the inner problem using reformulation of the inner problem into KKT conditions. It can be further studied how the collocation of the points that parameterize the orthotopic over-approximation of the GPE solution set ($\boldsymbol{\pi}$) can be exploited, e.g., via introduction of some redundant constraints. In the case study on dynamic experiment design the identification of local optima of the lower-level problem becomes a serious issue which was mitigated by the proper initialization of the optimization problem upon observation of the GPE solution set obtained by a set-inversion algorithm. This kind of infeasible-path heuristics might, of course, result in higher likeliness of identifying local optima of the problem (17). Another heuristics for obtaining a good local optimum can be employed, where the solutions to classical OED and linearization-based optimal experiment design are explored and effectively combined. Another path is the exploitation of multitude of algorithms for convex bilevel programs (Vicente and Calamai, 1994) (e.g., extreme point algorithms, complementarity pivot algorithm, descent methods, penalty function methods). Should one be interested in a feasible-path approach for solving a bilevel program, a nested approach could be used. Here a global optimization solver is employed to solve the inner problem directly and an upper-level optimization routine (e.g., gradient-based) is called to optimize the degrees of freedom of the outer problem. This scheme might, of course, lead to increased computational burden and it will be a subject of our further research to quantify the trade-off between the suggested feasible- and infeasible-path approaches.

9. Conclusions

In this paper we presented a novel methodology for designing the static and dynamic experiments in the framework of guaranteed parameter estimation. The proposed technique goes in line of classical design of experiments for least-squares estimation where the optimization of plant inputs is posed as an optimization problem over the (over-)approximation of the parametric joint-confidence region. We adopt the description (over-approximation) of the joint-confidence region using a box (orthotope) instead of the ellipse as done in a posteriori analysis of least-squares estimates or as suggested in earlier attempts on the experiment design for guaranteed estimation. We propose several ways of achieving

an over-approximation of the generally nonconvex set of guaranteed parameter estimates. An optimization problem is then posed over this over-approximation in order to identify optimal experimental conditions that shape the over-approximating box based on some measure (volume, perimeter, etc.). The resulting bilevel program is regularized by introduction of lifting variables for its well-posedness. The results obtained for the case studies of static and dynamic optimal experiment design provide many insights into the problem and solution of optimal experiment design in the context of guaranteed parameter estimation. They also show the differences between classical optimal experiment design and the optimal experiment design for guaranteed parameter estimation. They also highlight the need of fully nonlinear approaches to optimal experiment design as the linearization-based approach was shown to fail systematically to identify the optimal experiments. Due to the presence on nonconvex inner-level optimization problem in the bilevel program the proposed approach is hard to solve in general using state-of-the-art optimization solvers. The future work on this problem might thus consider development of the techniques that allow simplification of the proposed problem yet do not compromise the quality of the obtained solution (as shown for linearization-based approach) in order to be able to tackle large-scale nonlinear problems. Another path of the future work is to develop an efficient global optimization algorithm for solving the arising bilevel programs. Other possible extensions to this work are identified in the discussion section of this paper.

Acknowledgements

The research leading to these results has received funding from the European Commission under grant agreement number 291458 (ERC Advanced Investigator Grant MOBOCON).

References

Andersson, J., Åkesson, J., Diehl, M., 2012. CasADi – a symbolic package for automatic differentiation and optimal control. In: Forth, S., Hovland, P., Phipps, E., Utke, J., Walther, A. (Eds.), Recent Advances in Algorithmic Differentiation. Lecture Notes in Computational Science and Engineering. Springer, Berlin, pp. 297–307.

Bai, E.-W., Tempo, R., Cho, H., 1995. Membership set estimators: size, optimal inputs, complexity and relations with least squares. IEEE Trans. Circuits Syst. I: Fundam. Theory Appl. 42 (5), 266–277.

Baran, N., Wozny, G., Arellano-Garcia, H., 2012. Model-based design of experiments for model identification using closed-loop set-point response. In: Bogle, I.D.L., Fairweather, M. (Eds.), 22nd European Symposium on Computer Aided Process Engineering. Vol. 30 of Computer Aided Chemical Engineering. Elsevier, pp. 1337–1341.

Bates, D.M., Watts, D.G., 1988. Nonlinear Regression Analysis and Its Applications. John Wiley & Sons, Inc.

Bertsekas, D., Rhodes, I., 1971. Recursive state estimation for a set-membership description of uncertainty. IEEE IEEE Trans. Autom. Control 16 (2), 117–128.

Biegler, L.T., 1984. Solution of dynamic optimization problems by successive quadratic programming and orthogonal collocation. Comput. Chem. Eng. 8 (3/4), 243–248.

Bock, H., Plitt, K., 1984. A multiple shooting algorithm for direct solution of optimal control problems. In: Proc. of the 9th IFAC World Congress Budapest, pp. 242–247.

Borchers, S., Raković, S., Findeisen, R., 2011. Set membership parameter estimation and design of experiments using homothety. In: 18th IFA, World Congress. IFAC Proceedings Volumes, vol. 44(no. 1), pp. 9035–9040.

Evans, C., Findley, G., 2000. Analytic Solutions to the Lotka–Volterra Model for Sustained Chemical Oscillations., pp. 1–32.

Fisher, R., 1935. The Design of Experiments. Oliver & Boyd.

Fletcher, R., Leyffer, S., 2002. Numerical Experience With Solving MPECs as NLPs. Tech. Rep. Department of Mathematics and Computer Science, University of Dundee, Dundee.

Fletcher, R., Leyffer, S., Ralph, D., Scholtes, S., 2006. Local convergence of SQP methods for mathematical programs with equilibrium constraints. SIAM J. Optim. 17 (1), 259–286.

Franceschini, G., Macchietto, S., 2008. Model-based design of experiments for parameter precision: state of the art. Chem. Eng. Sci. 63 (19), 4846–4872.

Fung, K.Y., Ng, K.M., Zhang, L., Gani, R., 2016. A grand model for chemical product design. Comput. Chem. Eng.

Galvanin, F., Barolo, M., Bezzo, F., 2009. Online model-based redesign of experiments for parameter estimation in dynamic systems. Ind. Eng. Chem. Res. 48 (9), 4415–4427.

Gottu Mukkula, A.R., Paulen, R., 2016a. Optimal design of dynamic experiments for guaranteed parameter estimation. In: In: 2016 American Control Conference (ACC). IEEE, pp. 1826–1831.

Gottu Mukkula, A.R., Paulen, R., 2016b. Optimal dynamic experiment design for guaranteed parameter estimation. In: Kravanja, Z. (Ed.), European Symposium on Computer Aided Process Engineering ESCAPE 26. , pp. 757–762.

Gümüs, Z.H., Floudas, C.A., 2001. Global optimization of nonlinear bilevel programming problems. J. Glob. Optim. 20 (1), 1–31.

Hast, D., Findeisen, R., Streif, S., 2015. Detection and isolation of parametric faults in hydraulic pumps using a set-based approach and quantitative-qualitative fault specifications. Control Eng. Pract. 40, 61–70.

Hatz, K., Leyffer, S., Schlöder, J.P., Bock, H.G., 2013. Regularizing Bilevel Nonlinear Programs by Lifting, Preprint ANL/MCS-P 4076-0613.

Hatz, K., Schlöder, J.P., Bock, H.G., 2012. Estimating parameters in optimal control problems. SIAM J. Sci. Comput. 34 (3), A1707–A1728.

Houska, B., Telen, D., Logist, F., Diehl, M., Van Impe, J.F., 2015. An economic objective for the optimal experiment design of nonlinear dynamic processes. Automatica 51, 98–103.

Jaulin, L., Kieffer, M., Walter, E., Meizel, D., 2002. Guaranteed robust nonlinear estimation with application to robot localization. Trans. Syst. Man Cyber. Part C 32 (4), 374–381.

Jaulin, L., Walter, E., 1993. Set inversion via interval analysis for nonlinear bounded-error estimation. Automatica 29 (4), 1053–1064.

Kitsos, C., 2013. Optimal Experimental Design for Non-Linear Models: Theory and Applications. Springer, Berlin.

Kurzhanski, A.B., Sugimoto, K., Vályi, I., 1994. Guaranteed state estimation for dynamical systems: ellipsoidal techniques. Int. J. Adapt. Control Signal Process. 8 (1), 85–101.

Kurzhanski, A.B., Varaiya, P., 2011. Optimization of output feedback control under set-membership uncertainty. J. Optim. Theory Appl. 151 (1), 11–32.

Lin, Y., Stadtherr, M.A., 2007. Validated solutions of initial value problems for parametric ODEs. Appl. Numer. Math. 57 (10), 1145–1162.

Ljung, L., 1999. System Identification (2nd Ed.): Theory for the User. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Lucia, S., Andersson, J., Brandt, H., Diehl, M., Engell, S., 2014. Handling uncertainty in economic nonlinear model predictive control: a comparative case-study. J. Process Control 24, 1247–1259.

Lucia, S., Paulen, R., 2014. Robust nonlinear model predictive control with reduction of uncertainty via robust optimal experiment design. In: In: Proceedings of the 19th IFAC World Congress, Cape Town, South Africa, pp. 1904–1909.

Marco, M.D., Garulli, A., Lacroix, S., Vicino, A., 2000. A set theoretic approach to the simultaneous localization and map building problem. In: Proc. of the 39th IEEE Conf. on Decision and Control, pp. 833–838.

Milanese, M., Vicino, A., 1991. Optimal estimation theory for dynamic systems with set membership uncertainty. Automatica 27 (6), 997–1009.

Mitsos, A., Chachuat, B., Barton, P.I., 2009. Towards global bilevel dynamic optimization. J. Glob. Optim. 45 (1), 63–93.

Mitsos, A., Lemonidis, P., Barton, P.I., 2007. Global solution of bilevel programs with a nonconvex inner program. J. Glob. Optim. 42 (4), 475–513.

Müller, D., Esche, E., López, C.D.C., Wozny, G., 2014. An algorithm for the identification and estimation of relevant parameters for optimization under uncertainty. Comput. Chem. Eng. 71, 94–103.

Ogunnaike, B., Ray, W., 1994. Process Dynamics, Modeling, and Control. Topics in Chemical Engineering. Oxford University Press.

Pantelides, C., Renfro, J., 2013. The online use of first-principles models in process operations: review, current status and future needs. Comput. Chem. Eng. 51, 136–148.

Paulen, R., Villanueva, M.E., Chachuat, B., 2016. Guaranteed parameter estimation of non-linear dynamic systems using high-order bounding techniques with domain and CPU-time reduction strategies. IMA J. Math. Control Inf. 33 (3), 563–587.

Pronzato, L., Walter, E., 1989. Experiment design in a bounded-error context: comparison with D-optimality. Automatica 25 (3), 383–391.

Pronzato, L., Walter, E., 1990. Experiment design for bounded-error models. Math. Comput. Simul. 32 (5), 571–584.

Pronzato, L., Walter, E., 1994. Minimum-volume ellipsoids containing compact sets. Automatica 30 (11), 1731–1739.

Quarteroni, A., 2009. Mathematical models in science and engineering. Notices AMS 56 (1), 10–19.

Recker, S., Kerimoglu, N., Harwardt, A., Tkacheva, O., Marquardt, W., 2013. On the integration of model identification and process optimization. In: Proc. of the 23rd European Symposium on Computer Aided Process Engineering, pp. 1021–1026.

Safdarnejad, S.M., Gallacher, J.R., Hedengren, J.D., 2016. Dynamic parameter estimation and optimization for batch distillation. Comput. Chem. Eng. 86, 18–32.

Sager, S., 2013. Sampling decisions in optimum experimental design in the light of Pontryagin's maximum principle. SIAM J. Control Optim. 51 (4), 3181–3207.

Sargent, R., 2000. Optimal control. J. Comput. Appl. Math. 124, 361–371.

Schweppe, F., 1968. Recursive state estimation: unknown but bounded errors and system inputs. IEEE Trans. Autom. Control 13 (1), 22–28.

Streif, S., Kim, K.-K.K., Rumschinski, P., Kishida, M., Shen, D.E., Findeisen, R., Braatz, R.D., 2016. Robustness analysis, prediction, and estimation for uncertain biochemical networks: an overview. J. Process Control 42, 14–34.
Tanaskovic, M., Fagiano, L., Morari, M., 2014. On the optimal worst-case experiment design for constrained linear systems. Automatica 50 (12), 3291–3298.
Tawarmalani, M., Sahinidis, N.V., 2005. A polyhedral branch-and-cut approach to global optimization. Math. Program. 103 (2), 225–249.
Telen, D., Houska, B., Logist, F., Diehl, M., Van Impe, J.,2013. Guaranteed robust optimal experiment design for nonlinear dynamic systems. In: 2013 European Control Conference (ECC). IEEE, pp. 2939–2944.
Telen, D., Houska, B., Logist, F., Van Impe, J., 2016. Multi-purpose economic optimal experiment design applied to model based optimal control. Comput. Chem. Eng. 94, 212–220.
Telen, D., Logist, F., Derlinden, E.V., Tack, I., Impe, J.V., 2012. Optimal experiment design for dynamic bioprocesses: a multi-objective approach. Chem. Eng. Sci. 78, 82–97.
Telen, D., Vercammen, D., Logist, F., Impe, J.V., 2014. Robustifying optimal experiment design for nonlinear, dynamic (bio)chemical systems. Comput. Chem. Eng. 71, 415–425.
Vicente, L.N., Calamai, P.H., 1994. Bilevel and multilevel programming: a bibliography review. J. Glob. Optim. 5 (3), 291–306.
Wächter, A., Biegler, L., 2006. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. Math. Program. 106, 25–57.
Walter, E., Piet-Lahanier, H., 1990. Estimation of parameter bounds from bounded-error data: a survey. Math. Comput. Simul. 32 (5), 449–468.

Journal of Process Control 83 (2019) 187–195

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

Optimal experiment design in nonlinear parameter estimation with exact confidence regions

Anwesh Reddy Gottu Mukkula[a,*], Radoslav Paulen[a,b]

[a] Process Dynamics and Operations Group, Department of Chemical and Biochemical Engineering, Technische Universität Dortmund, Emil-Figge-Strasse 70, 44227 Dortmund, Germany
[b] Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Radlinského 9, 81237 Bratislava, Slovakia

ARTICLE INFO

Article history:
Received 24 September 2017
Received in revised form 6 January 2019
Accepted 14 January 2019
Available online 23 March 2019

Keywords:
Optimal experiment design
Parameter estimation
Least-squares estimation

ABSTRACT

A model-based optimal experiment design (OED) of nonlinear systems is studied. OED represents a methodology for optimizing the geometry of the parametric joint-confidence regions (CRs), which are obtained in an a posteriori analysis of the least-squares parameter estimates. The optimal design is achieved by using the available (experimental) degrees of freedom such that more informative measurements are obtained. Unlike the commonly used approaches, which base the OED procedure upon the linearized CRs, we explore a path where we explicitly consider the exact CRs in the OED framework. We use a methodology for a finite parametrization of the exact CRs within the OED problem and we introduce a novel approximation technique of the exact CRs using inner- and outer-approximating ellipsoids as a computationally less demanding alternative. The employed techniques give the OED problem as a finite-dimensional mathematical program of bilevel nature. We use two small-scale illustrative case studies to study various OED criteria and compare the resulting optimal designs with the commonly used linearization-based approach. We also assess the performance of two simple heuristic numerical schemes for bilevel optimization within the studied problems.

1. Introduction

At present, advanced industrial engineering and management strive for resource- and energy-efficient design and operation of systems, plants, and processes. Here a use of the model-based techniques is a leading paradigm. The employed models, whether mechanistic or data-based, include a finite number of parameters, whose values are related to the particular natural and system-wide phenomena and are thus commonly only known to belong to some interval or unknown completely. Therefore, as-precise-as-possible determination of the unknown (uncertain) model parameters is crucial for the deployment of the effective model-based solutions.

In the world, where the sensing technology becomes virtually present everywhere, the simplest way of obtaining the parameter values is to conduct a series of observations, experiments, during which the measurements of some quantities (output variables) are gathered. Subsequently, an estimation procedure is employed to find parameter values such that the model outputs match the observed data. As the measurements from a real plant are corrupted with some (random) noise, the uncertain model parameters cannot be identified precisely. This a posteriori uncertainty can be represented by a joint-confidence region (CR) that encompasses all the likely parameter estimates, given the probability distribution of the measurement noise.

The parametric uncertainty can be reduced by performing an experiment that appropriately sets the plant into an operating region, where more informative measurements are obtained. A way how to identify such experiment consists in performing an optimal experiment design (OED) [1,2]. OED identifies the best experiment in terms of initial conditions, control inputs, sampling times and/or locations of measurement devices. The model-based OED problem is usually formulated as a mathematical program, where a certain measure of the CR, such as volume, is minimized.

Some well-established methods exist for the OED for linear systems [3], wherein CR is an ellipsoid [4], so the task of its optimal shaping is greatly simplified. For nonlinear systems, the most common, yet approximate, approach is to resort to a system linearization and a use of the linear techniques. Beale [5] presented methodology for assessment of system nonlinearity in this respect. Other approaches, covered in [6] for convex problems, are of more or less approximate character and mostly rely on the convexity properties of the OED problem. Bates and Watts [7] presented a framework based on local nonlinear curvature properties, which is

* Corresponding author.
E-mail addresses: anweshreddy.gottumukkula@tu-dortmund.de (A.R. Gottu Mukkula), radoslav.paulen@stuba.sk (R. Paulen).

https://doi.org/10.1016/j.jprocont.2019.01.004
0959-1524/

also an approximate technique. The use of a finite parametrization of the exact CRs [8–10] is a relatively recent subject of study.

In this contribution, we study the framework for OED of nonlinear systems that is based on explicit consideration of the exact CRs. We also present a computationally simpler variant that is based on simultaneous inner- and outer-approximations of the CR by ellipsoids. The preliminary findings of this work were presented in the conference contribution [11]. We study various common design criteria and compare the performance of the presented techniques with the linearized OED.

We organized the paper as follows. The concepts of linear and nonlinear parameter estimation and construction of CRs are discussed first. Next, the formulation is presented of the experiment design criteria that directly use exact CRs (exact designs). We also present two simple heuristic numerical approaches taken from literature to solve the arising bilevel optimization problems. Finally results of two illustrative case studies are presented and discussed.

2. Parameter estimation problem

2.1. Mathematical model

In this paper, a mathematical model of a system is represented by

$$\hat{\boldsymbol{y}}(\boldsymbol{p},\tau)=\boldsymbol{F}(\boldsymbol{p},\boldsymbol{u}_\tau), \tag{1}$$

with $\hat{\boldsymbol{y}}$ as n_y measured variables, $\boldsymbol{u}_\tau$ as n_u degrees of freedom and $\hat{\boldsymbol{p}}$ as n_p uncertain parameters. Here τ represents an ordinal number of the data point taken in one or more experiments. Function $\boldsymbol{F}:\mathbb{R}^{n_p}\times\mathbb{R}^{n_u}\rightarrow\mathbb{R}^{n_y}$ is twice continuously differentiable mapping. Throughout the paper, we resort to a representation (1), which considers the system model as static and explicit. However, the presented findings can straightforwardly be extended to dynamic and implicit models.

We will assume throughout the paper that the model is not over-parametrized and that the parameters are identifiable. We consider that, upon the realization of an experiment or several experiments, N instances are gathered of n_y-dimensional vector of plant measurements $\boldsymbol{y}_m$ and are subsequently used for the estimation of unknown parameters. Throughout the paper, we will assume the Gaussian noise to be corrupting the measurements. In the following subsections, existing frameworks are presented for the identification of the unknown parameters and the corresponding exact CRs for both the linear and the nonlinear parameter estimation problems.

2.2. Linear parameter estimation

Parameter estimation with a linear model involved is a well-studied topic in the literature [4]. Assume a mathematical model with a mapping function $\boldsymbol{F}_l$ of the form

$$\hat{\boldsymbol{y}}(\boldsymbol{p},\tau)=\boldsymbol{F}_l(\boldsymbol{p},\boldsymbol{u}_\tau)=\mathbf{Q}(\boldsymbol{u}_\tau)\boldsymbol{p}, \tag{2}$$

where $\mathbf{Q}(\boldsymbol{u}_\tau)$ is a so-called regressors matrix with appropriate dimensions.

Under the assumption of uncorrelated and normally distributed measurement noise with known standard deviation vector $\boldsymbol{\sigma}$, the maximum-likelihood estimate is found via the least-squares estimation as

$$\hat{\boldsymbol{p}}\in\underset{\boldsymbol{p}}{\operatorname{argmin}} J_w(\boldsymbol{p}), \tag{3}$$

with

$$J_w(\boldsymbol{p}):=\sum_{i=1}^{n_y}\sum_{\tau=\tau_1}^{\tau_N}\sigma_i^{-2}(y_{i,m}(\tau)-\hat{y}_i(\boldsymbol{p},\tau))^2. \tag{4}$$

The CR of parameter estimates is then given by an ellipsoid [4]

$$(\boldsymbol{p}-\hat{\boldsymbol{p}})^T\mathbb{FIM}(\mathcal{U})(\boldsymbol{p}-\hat{\boldsymbol{p}})\leq\chi^2_{\alpha,n_p}, \tag{5}$$

where $\mathbb{FIM}$ is the so-called Fisher information matrix,

$$\mathbb{FIM}(\mathcal{U})=\sum_{\tau=\tau_1}^{\tau_N}\mathbf{Q}(\boldsymbol{u}_\tau)^T\operatorname{diag}^{-1}(\sigma_1^2,\ldots,\sigma_{n_y}^2)\mathbf{Q}(\boldsymbol{u}_\tau), \tag{6}$$

$\mathcal{U}:=(\boldsymbol{u}_{\tau_1}^T,\boldsymbol{u}_{\tau_2}^T,\ldots,\boldsymbol{u}_{\tau_N}^T)^T$, and χ^2_{α,n_p} represents the upper α quantile of the chi-squared statistical distribution with n_p degrees of freedom.

If the *variance of the measurement noise is unknown*, the covariance matrix $\operatorname{diag}(\sigma_1^2,\ldots,\sigma_{n_y}^2)$ is normally approximated by $s^2\mathbb{I}$ with

$$s^2:=\frac{J(\hat{\boldsymbol{p}})}{N-n_p}, \tag{7}$$

where the expected (most-likely) value of parameters $\hat{\boldsymbol{p}}$ is identified by solving

$$\hat{\boldsymbol{p}}\in\underset{\boldsymbol{p}}{\operatorname{argmin}}\ J(\boldsymbol{p})=\sum_{i=1}^{n_y}\sum_{\tau=\tau_1}^{\tau_N}(y_{i,m}(\tau)-\hat{y}_i(\boldsymbol{p},\tau))^2. \tag{8}$$

The joint-confidence ellipsoid is then given by

$$(\boldsymbol{p}-\hat{\boldsymbol{p}})^T\left(\sum_{\tau=\tau_1}^{\tau_N}\mathbf{Q}(\boldsymbol{u}_\tau)^T\left(s^2\mathbb{I}\right)^{-1}\mathbf{Q}(\boldsymbol{u}_\tau)\right)(\boldsymbol{p}-\hat{\boldsymbol{p}})\leq n_p\mathcal{F}_{n_p,N-n_p,\alpha}, \tag{9}$$

where $\mathbb{FIM}$ is replaced by its corresponding approximation and $\mathcal{F}$ represents the upper α quantile of the Fisher distribution with n_p and $N-n_p$ degrees of freedom in the numerator and in the denominator, respectively.

2.3. Nonlinear parameter estimation

Given a static nonlinear mathematical model

$$\hat{\boldsymbol{y}}(\boldsymbol{p},\tau)=\boldsymbol{F}_{nl}(\boldsymbol{p},\boldsymbol{u}_\tau), \tag{10}$$

one can identify the (exact) CR dependent upon the availability of information about the variance of the measurement noise. If the variance is known, the exact CR is given by [4]

$$J_w(\boldsymbol{p})-J_w(\hat{\boldsymbol{p}})\leq\chi^2_{\alpha,n_p}, \tag{11}$$

while, when the variance of the measurement noise is unknown, the exact CR is given by [4]

$$J(\boldsymbol{p})-J(\hat{\boldsymbol{p}})\leq n_p s^2\mathcal{F}_{n_p,N-n_p,\alpha}. \tag{12}$$

At this point we can define the sets $P_w:=\{\boldsymbol{p}|\text{Eq. (11)}\}$ and $P:=\{\boldsymbol{p}|\text{Eq. (12)}\}$. Unlike in the linear parameter estimation, the CR does not generally have a shape of an ellipsoid due to nonlinearity.

3. Optimal experiment design

We present a methodology for OED for both linear and nonlinear parameter estimation problems. We will assume that the CR is connected. For disjoint exact CRs, which result from non-identifiability issues, it is normally suggested to perform a re-parameterization of the model [7]. We will also assume that an estimate $\hat{\boldsymbol{p}}$ is available. The final assumptions which are inherent to the standard experiment design techniques is that there exists no structural plant-model mismatch and that the expected realization of the measurement noise is 0. In turn this results in $\boldsymbol{y}_m(\tau)=\hat{\boldsymbol{y}}(\hat{\boldsymbol{p}},\tau),\ \forall\tau$.

Several design criteria are proposed in the literature [3] such as A, D, E, Modified E, V, Q, M and so on. Each of these designs aims to tune a particular property of the confidence region. We will focus our study on the most used criteria, i.e., A, D, and E, but

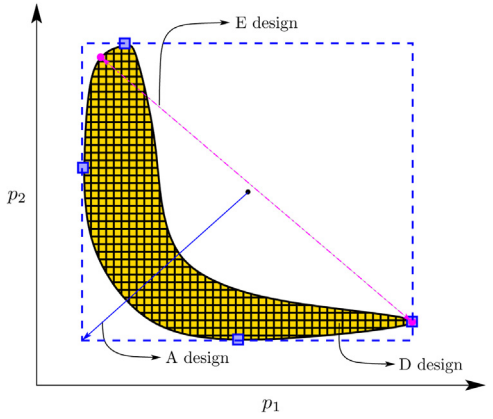


Fig. 1. Illustration of the design criteria in two-dimensional parametric space. The plot shows a generic exact CR (shaded), the over-approximating orthotope (dashed) of the exact CR identified using the anchor points $\boldsymbol{\pi}$ (□), ● mark the points that give the maximal Euclidean distance of two points in the CR.

other design criteria might be considered as well using the ideas presented herein.

3.1. A-optimal design

The idea behind the A design criterion is to minimize the perimeter of the box that encloses the exact CR [3], i.e., to minimize the sum of projections of the CR on the parameter axes. This idea is sketched in Fig. 1, where the shaded set represents a CR. The enclosing box is given by the four anchor points, marked as squares (□).

For a general CR, one can identify $2n_p$ anchor points

$$\boldsymbol{\pi}:=\left\{\begin{pmatrix} p_1^L \\ p_2^{1,L} \\ \vdots \\ p_{n_p}^{1,L} \end{pmatrix},\begin{pmatrix} p_1^U \\ p_2^{1,U} \\ \vdots \\ p_{n_p}^{1,U} \end{pmatrix},\begin{pmatrix} p_1^{2,L} \\ p_2^{L} \\ \vdots \\ p_{n_p}^{2,L} \end{pmatrix},\cdots,\begin{pmatrix} p_1^{n_p,U} \\ p_2^{n_p,U} \\ \vdots \\ p_{n_p}^{U} \end{pmatrix}\right\}, \tag{13}$$

where each point gives a lower or an upper limit of the value of a particular parameter in the exact CR. The anchor points can be identified by solving the following optimization problem

$$\phi_A(\mathcal{U}):=\max_{\boldsymbol{\pi}} \sum_{j=1}^{n_p} p_j^U - p_j^L \tag{14a}$$

$$\text{s.t.}\ \forall j \in \{1,\dots,2n_p\},\quad \forall \tau \in \{\tau_1,\dots,\tau_N\}: \tag{14b}$$

$$\hat{\boldsymbol{y}}(\boldsymbol{\pi}_j,\tau)=\boldsymbol{F}_{nl}(\boldsymbol{\pi}_j,\boldsymbol{u}_\tau), \tag{14c}$$

$$\boldsymbol{y}_m(\tau)=\hat{\boldsymbol{y}}(\hat{\boldsymbol{p}},\tau), \tag{14d}$$

$$J_w(\boldsymbol{\pi}_j)-J_w(\hat{\boldsymbol{p}})\le\chi^2_{\alpha,n_p}. \tag{14e}$$

Note that the problem of identifying the orthotopic enclosure of the CR is formulated for the case when the measurement-noise variance is known using the expression for the CR (11). This would simply be exchanged with the expression (12) in case when the variance is unknown.

Note that the problem (14) is separable and highly structured. On the other hand it is non-convex in general and the number of its optimization variables ($2n_p^2$) grows quadratically with the number of parameters. This means that identification of an orthotope might get challenging for the state-of-the-art solvers and for the high-dimensional problems.

The A-optimal experiment design (A design) can be identified by solving

$$\min_{\boldsymbol{u}_\tau\in[\boldsymbol{u}^L,\boldsymbol{u}^U],\,\forall\tau\in\{\tau_1,\dots,\tau_N\}}\phi_A(\mathcal{U}), \tag{15}$$

which is a special case of a bilevel program. The bounds $\boldsymbol{u}^L$ and $\boldsymbol{u}^U$ represent the lower and upper limits of the experimental degrees of freedom.

In the linear case, the A design is identified by

$$\min_{\substack{\boldsymbol{u}_\tau\in[\boldsymbol{u}^L,\boldsymbol{u}^U]\\ \forall\tau\in\{\tau_1,\dots,\tau_N\}}} \tilde{\phi}_A(\mathbb{FIM}):=$$

$$\min_{\substack{\boldsymbol{u}_\tau\in[\boldsymbol{u}^L,\boldsymbol{u}^U]\\ \forall\tau\in\{\tau_1,\dots,\tau_N\}}} \operatorname{trace}(\mathbb{FIM}^{-1}) \tag{16a}$$

$$\text{s.t.}\ \hat{\boldsymbol{y}}(\boldsymbol{p},\tau)=\boldsymbol{F}_l(\boldsymbol{p},\boldsymbol{u}_\tau),\quad \forall\tau\in\{\tau_1,\dots,\tau_N\}. \tag{16b}$$

In case of unknown variance of measurement noise, the approximate $\mathbb{FIM}$ from (9) will be used. We denote this linearization-based approach as *classical* in this study.

3.2. D-optimal design

The D design aims to find the experimental conditions such that the exact CR would have minimum volume. The D-optimal design problem then reads as

$$\min_{\substack{\boldsymbol{u}_\tau\in[\boldsymbol{u}^L,\boldsymbol{u}^U]\\ \forall\tau\in\{\tau_1,\dots,\tau_N\}}} \phi_D(\mathcal{U}):=$$

$$\min_{\substack{\boldsymbol{u}_\tau\in[\boldsymbol{u}^L,\boldsymbol{u}^U]\\ \forall\tau\in\{\tau_1,\dots,\tau_N\}}} \int\cdots\int_{P_w(\mathrm{U})} \mathrm{d}p_1\dots\mathrm{d}p_{n_p}. \tag{17}$$

As there is no finite-dimensional parameterization of the set $P_w(\mathcal{U})$ available in general, it is very hard in general to evaluate the volume integral. We propose to use a gridding-based approach, where the grid is evaluated inside the aforementioned orthotopic enclosure of the CR. The proposed optimization problem may be written as

$$\min_{\substack{\boldsymbol{u}_\tau\in[\boldsymbol{u}^L,\boldsymbol{u}^U]\\ \forall\tau\in\{\tau_1,\dots,\tau_N\}}} \hat{\phi}_D(\mathcal{U}):= \min_{\substack{\boldsymbol{u}_\tau\in[\boldsymbol{u}^L,\boldsymbol{u}^U]\\ \forall\tau\in\{\tau_1,\dots,\tau_N\}}} \sum_{\forall i\in\mathcal{I}_\Pi}\delta_i \tag{18a}$$

$$\text{s.t.}\ \delta_i=\begin{cases}1, & \text{if}\ \Pi_i\in P_w\\ 0, & \text{otherwise}\end{cases} \tag{18b}$$

$$\begin{aligned}\Pi=\ & \{p_1^L,p_1^L+\epsilon,p_1^L+2\epsilon,\dots,p_1^U\}\times\\ & \{p_2^L,p_2^L+\epsilon,p_2^L+2\epsilon,\dots,p_2^U\}\times\cdots\\ & \times\{p_{n_p}^L,p_{n_p}^L+\epsilon,p_{n_p}^L+2\epsilon,\dots,p_{n_p}^U\},\end{aligned} \tag{18c}$$

$$\max_{\pi}\sum_{j=1}^{n_p} p_j^U-p_j^L, \tag{18d}$$

$$\text{s.t.}\ \forall j\in\{1,\dots,2n_p\},\quad \forall\tau\in\{\tau_1,\dots,\tau_N\}: \tag{18e}$$

$$\hat{\boldsymbol{y}}(\boldsymbol{\pi}_j,\tau)=\boldsymbol{F}_{nl}(\boldsymbol{\pi}_j,\boldsymbol{u}_\tau), \tag{18f}$$

$$\boldsymbol{y}_m(\tau)=\hat{\boldsymbol{y}}(\hat{\boldsymbol{p}},\tau), \tag{18g}$$

$$J_w(\boldsymbol{\pi}_j)-J_w(\hat{\boldsymbol{p}})\le\chi^2_{\alpha,n_p}, \tag{18h}$$

where $\epsilon>0$ is the tuning parameter that determines the accuracy of the approximation and $\mathcal{I}_\Pi$ is the index set of Π. This approach

for approximating the volume of the CR is illustrated in Fig. 1 as a grid in the shaded set.

In principle, the identification of the orthotopic enclosure of the CR can be removed and the problem (18) can be modified to a single-level mathematical program. Nonetheless the optimization problem (18) is non-smooth due to the presence of indicator function (18b) and thus it can get very challenging and computationally highly expensive, especially in higher dimensions. An alternative approach can be exploited by approximation of the volume using the semi-algebraic sets [12]. Approaches to inner approximation of the CR based on an orthotope and on successive SDP approximations are presented by Streif et al. [13]. We propose a simpler approximation here, which uses an idea similar to the Löwner-John ellipsoids [14]. We construct the inner- and outer-approximating ellipsoids, which are the scaled counterparts of the linearized CRs. The proposed approximate D design is found by

$$\min_{\substack{\boldsymbol{u}_\tau \in [\boldsymbol{u}^L, \boldsymbol{u}^U] \\ \forall \tau \in \{\tau_1, \ldots, \tau_N\}}} \frac{\det(k_{out}\mathbb{FIM}^{-1})}{k_{out}} + \frac{\det(k_{in}\mathbb{FIM}^{-1})}{k_{in}} \tag{19a}$$

$$\text{s.t.} \max_{\boldsymbol{p}_{out}, \boldsymbol{p}_{in}, k_{out}, k_{in}} k_{out} - k_{in} \tag{19b}$$

$$\text{s.t.} \ \forall \tau \in \{\tau_1, \ldots, \tau_N\}: \tag{19c}$$

$$\hat{\boldsymbol{y}}(\boldsymbol{p}_{out}, \tau) = \boldsymbol{F}_{nl}(\boldsymbol{p}_{out}, \boldsymbol{u}_\tau), \tag{19d}$$

$$\hat{\boldsymbol{y}}(\boldsymbol{p}_{in}, \tau) = \boldsymbol{F}_{nl}(\boldsymbol{p}_{in}, \boldsymbol{u}_\tau), \tag{19e}$$

$$\boldsymbol{y}_m(\tau) = \hat{\boldsymbol{y}}(\hat{\boldsymbol{p}}, \tau), \tag{19f}$$

$$J_w(\boldsymbol{p}_{out}) - J_w(\hat{\boldsymbol{p}}) \leq \chi^2_{\alpha, n_p}, \tag{19g}$$

$$J_w(\boldsymbol{p}_{in}) - J_w(\hat{\boldsymbol{p}}) \leq \chi^2_{\alpha, n_p}, \tag{19h}$$

$$(\boldsymbol{p}_{out} - \hat{\boldsymbol{p}})^T \mathbb{FIM} (\boldsymbol{p}_{out} - \hat{\boldsymbol{p}}) \leq k_{out}, \tag{19i}$$

$$(\boldsymbol{p}_{in} - \hat{\boldsymbol{p}})^T \mathbb{FIM} (\boldsymbol{p}_{in} - \hat{\boldsymbol{p}}) \geq k_{in}, \tag{19j}$$

where $\boldsymbol{p}_{out}$ and $\boldsymbol{p}_{in}$ are intersection points between outer-/inner-approximating ellipsoids and the exact CR. The scaling factors k_{out} and k_{in} express the magnitude of deviation of the outer- and, respectively, inner-approximating ellipsoid from the linearized CR. The weighting in the cost function is then introduced to penalize the contribution of the most deviating ellipsoid. This prevents the design procedure from concentrating on shaping the ellipsoid that is potentially a very loose approximation of the CR and in practice avoids numerical and irregularity problems. Hence that the proposed problem also scales well, i.e., linearly, w.r.t. the number of the parameters as the lower-level problem optimizes $2n_p + 2$ variables. The proposed approximate D-optimal design is therefore computationally a less expensive problem when compared to the exact D-optimal design using the gridding-based approach. We will denote the proposed approximate D-optimal design approach as the *ellipsoidal D design*. The idea behind this method, slightly modified, could also be used for an approximate A design but we do not explore this path in the present study explicitly.

Note also that, if the CR can be expressed exactly as (5), the proposed optimization problem boils down to the classical D design where one solves

$$\min_{\substack{\boldsymbol{u}_\tau \in [\boldsymbol{u}^L, \boldsymbol{u}^U] \\ \forall \tau \in \{\tau_1, \ldots, \tau_N\}}} \tilde{\phi}_{\mathrm{D}}(\mathbb{FIM}) := \min_{\substack{\boldsymbol{u}_\tau \in [\boldsymbol{u}^L, \boldsymbol{u}^U] \\ \forall \tau \in \{\tau_1, \ldots, \tau_N\}}} \det(\mathbb{FIM}^{-1}) \tag{20a}$$

$$\text{s.t.} \ \hat{\boldsymbol{y}}(\boldsymbol{p}, \tau) = \boldsymbol{F}_l(\boldsymbol{p}, \boldsymbol{u}_\tau), \quad \forall \tau \in \{\tau_1, \ldots, \tau_N\}. \tag{20b}$$

3.3. E-optimal design

Objective of the E design is to minimize the Euclidean distance ($\|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_2$) between the two points (• in Fig. 1) that belong to the CR and their Euclidean distance is maximal. This can be expressed as

$$\min_{\substack{\boldsymbol{u}_\tau \in [\boldsymbol{u}^L, \boldsymbol{u}^U] \\ \forall \tau \in \{\tau_1, \ldots, \tau_N\}}} \phi_E(\mathcal{U}) \tag{21a}$$

$$\text{s.t.} \ \phi_E(\mathcal{U}) = \max_{\boldsymbol{\varphi}_1, \boldsymbol{\varphi}_2} \|\boldsymbol{\varphi}_1 - \boldsymbol{\varphi}_2\|_2^2 \tag{21b}$$

$$\text{s.t.} \ \forall j \in \{1, 2\}, \quad \forall \tau \in \{\tau_1, \ldots, \tau_N\}: \tag{21c}$$

$$\hat{\boldsymbol{y}}(\boldsymbol{\varphi}_j, \tau) = \boldsymbol{F}_{nl}(\boldsymbol{\varphi}_j, \boldsymbol{u}_\tau), \tag{21d}$$

$$\boldsymbol{y}_m(\tau) = \hat{\boldsymbol{y}}(\hat{\boldsymbol{p}}, \tau), \tag{21e}$$

$$J_w(\boldsymbol{\varphi}_j) - J_w(\hat{\boldsymbol{p}}) \leq \chi^2_{\alpha, n_p}. \tag{21f}$$

The E design is also known as a decorrelating design as it aims at finding the experimental conditions such that the CR is as spherical as possible. This criterion is illustrated in Fig. 1. It is noteworthy that the lower-level problem of (21) scales linearly w.r.t. the number of parameters as it optimizes $2n_p$ variables.

In the linear case the (classical) E design is identified by

$$\min_{\substack{\boldsymbol{u}_\tau \in [\boldsymbol{u}^L, \boldsymbol{u}^U] \\ \forall \tau \in \{\tau_1, \ldots, \tau_N\}}} \tilde{\phi}_{\mathrm{E}}(\mathbb{FIM}) := \min_{\substack{\boldsymbol{u}_\tau \in [\boldsymbol{u}^L, \boldsymbol{u}^U] \\ \forall \tau \in \{\tau_1, \ldots, \tau_N\}}} \max_i \lambda_i(\mathbb{FIM}^{-1}) \tag{22a}$$

$$\text{s.t.} \ \hat{\boldsymbol{y}}(\boldsymbol{p}, \tau) = \boldsymbol{F}_l(\boldsymbol{p}, \boldsymbol{u}_\tau), \quad \forall \tau \in \{\tau_1, \ldots, \tau_N\}, \tag{22b}$$

where $\lambda_i(\cdot)$ is the ith eigenvalue of a matrix.

4. Numerical Implementation

In this section we discuss the possible ways to solve the proposed optimization problems. We will exploit BARON [15] in this work in order to guarantee global optimality of the classical OED problems (16), (20) and (22). Special attention is devoted to the presented bilevel programs as the classical OED problems are single-level optimization problems and can straightforwardly be solved using a nonlinear program solver. We present two simple heuristic approaches taken from literature that can be used to solve the presented bilevel problems, which can be generalized in the form

$$\min_{\boldsymbol{x}_1} \ f(\boldsymbol{x}_1, \boldsymbol{x}_2^*) \tag{23a}$$

$$\text{s.t.} \ \boldsymbol{x}_2^* \in \operatorname*{argmax}_{\boldsymbol{x}_2} \ g(\boldsymbol{x}_2) \tag{23b}$$

$$\text{s.t.} \ \boldsymbol{0} = \boldsymbol{h}_E(\boldsymbol{x}_1, \boldsymbol{x}_2), \tag{23c}$$

$$\boldsymbol{0} \geq \boldsymbol{h}_I(\boldsymbol{x}_1, \boldsymbol{x}_2). \tag{23d}$$

A special care has to be taken w.r.t. the non-convexity of lower-level problem. Its global optimum has to be identified in order to guarantee feasibility of the upper-level problem [16].

4.1. Nested approach

The following nested approach is inspired by the formulation proposed in [17,18] for solving dynamic optimization problems and in [19] for solving a coordination control algorithm using a

price-driven coordination technique. The nested approach splits the bilevel optimization problem into a lower-level optimization problem

$$\boldsymbol{x}_2^* \in \underset{\boldsymbol{x}_2}{\operatorname{argmax}} g(\boldsymbol{x}_2) \tag{24a}$$

$$\text{s.t. } \mathbf{0} = \boldsymbol{h}_E(\boldsymbol{x}_1^*, \boldsymbol{x}_2), \tag{24b}$$

$$\mathbf{0} \geq \boldsymbol{h}_I(\boldsymbol{x}_1^*, \boldsymbol{x}_2). \tag{24c}$$

that is solved for a given $\boldsymbol{x}_1^*$ using a global solver (e.g. BARON) and a upper-level optimization problem

$$\boldsymbol{x}_1^* \in \underset{\boldsymbol{x}_1}{\operatorname{argmin}} f(\boldsymbol{x}_1, \boldsymbol{x}_2^*) \tag{25a}$$

$$\text{s.t. } \mathbf{0} = \boldsymbol{h}_E(\boldsymbol{x}_1, \boldsymbol{x}_2^*), \tag{25b}$$

$$\mathbf{0} \geq \boldsymbol{h}_I(\boldsymbol{x}_1, \boldsymbol{x}_2^*). \tag{25c}$$

that can then be solved for a given $\boldsymbol{x}_2^*$ using a local solver. We use IPOPT [20] as a local solver in this work.

The individual optimization problems are interconnected by the copy variables $\boldsymbol{x}_1^*$ and $\boldsymbol{x}_2^*$ that are exchanged between the problems. The problems are solved repeatedly and the convergence of the nested approach is claimed once the consecutive values of the copy variables satisfy $\|\boldsymbol{x}_{1,k+1}^* - \boldsymbol{x}_{1,k}^*\| \approx 0$ and $\|\boldsymbol{x}_{2,k+1}^* - \boldsymbol{x}_{2,k}^*\| \approx 0$, where k is an iteration counter.

If a gradient-based solver is used to determine the local optimum of the problem (25), the objective and constraint gradients are to be supplied. An approach from [21] can be used in this respect with $\boldsymbol{x}_k := (\boldsymbol{x}_{1,k}^T, \boldsymbol{x}_{2,k}^T)^T$

$$\begin{bmatrix} \nabla^2_{\boldsymbol{x}_2,\boldsymbol{x}_2} L|_{\boldsymbol{x}_k} & \nabla^T_{\boldsymbol{x}_2} \boldsymbol{h}|_{\boldsymbol{x}_k} \\ -\nabla_{\boldsymbol{x}_2} \boldsymbol{h}|_{\boldsymbol{x}_k} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \frac{\mathrm{d}\boldsymbol{x}_2^*}{\mathrm{d}\boldsymbol{x}_1} \\ \frac{\mathrm{d}\boldsymbol{v}^*}{\mathrm{d}\boldsymbol{x}_1} \end{bmatrix} = - \begin{bmatrix} \nabla^2_{\boldsymbol{x}_2,\boldsymbol{x}_1} L|_{\boldsymbol{x}_k} \\ \nabla_{\boldsymbol{x}_1} \boldsymbol{h}|_{\boldsymbol{x}_k} \end{bmatrix}, \tag{26}$$

where L represents the Lagrangian of the lower-level problem and $\boldsymbol{v} := (\boldsymbol{v}_E^T, \boldsymbol{v}_{I,i}^T)^T, \forall i \in \mathcal{I}_A$ is the vector of multipliers corresponding to the equality and active inequality constraints $\boldsymbol{h}(\boldsymbol{x}_1^*, \boldsymbol{x}_2) := (\boldsymbol{h}_E^T(\boldsymbol{x}_1^*, \boldsymbol{x}_2), \boldsymbol{h}_{I,i}^T(\boldsymbol{x}_1^*, \boldsymbol{x}_2))^T, \forall i \in \mathcal{I}_A$ of the lower-level problem and $\mathcal{I}_A$ is an index set of the active inequality constraints. It is not guaranteed that the nested approach always converge to a local minimum [18]. Instead it may sometimes converge to a local maximum or a saddle point. The obtained solution can be verified by evaluating the necessary and sufficient conditions for optimality.

4.2. KKT-based approach

Another heuristic approach for solving a bilevel optimization problem is to reformulate the lower-level problem using the KKT conditions [22–24]. The reformulated problem reads as

$$\min_{\substack{\boldsymbol{x}_1,\boldsymbol{x}_2 \\ \boldsymbol{v}_E,\boldsymbol{v}_I \leq 0}} \quad f(\boldsymbol{x}_1, \boldsymbol{x}_2) \tag{27a}$$

$$\text{s.t. } \mathbf{0} = \nabla_{\boldsymbol{x}} L(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{v}_E, \boldsymbol{v}_I), \tag{27b}$$

$$\mathbf{0} = \nabla_{\boldsymbol{v}} L(\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{v}_E, \boldsymbol{v}_I) = \boldsymbol{h}_E(\boldsymbol{x}_1, \boldsymbol{x}_2), \tag{27c}$$

$$0 = v_{I,i} h_{I,i}, \quad \forall i \in \mathcal{I}_I, \tag{27d}$$

where $\mathcal{I}_I$ is the index set of the inequality constraints of (23). As discussed above, the lower-level problem has to be solved to global optimality, which is not guaranteed by satisfaction of the KKT conditions. Reaching of global optimum of the lower-level problem has to be assured upon convergence in order to guarantee feasibility of the lower-level problem and thus a local optimum of the bilevel program. The feasibility test can be performed by solving (24) globally or by gridding or by set inversion [25] techniques with a subsequent comparison of obtained values for variables of lower-level problem.

We note that there are other approaches that can be employed to solve the problem (23). The solution methods proposed by Dutta et al. in [26] and by Dempe et al. in [27] assume a convex inner level optimization problem. Bard et al. [28], Dempe et al. [24] and Mitsos et al. [29] proposed solution methods considering a nonconvex inner level optimization problem. It is generally well known that there is a close connection between bilevel problem and semi-infinite programming (SIP) as discussed in [30]. Cutting-plane SIP algorithm is proposed by Blankenship and Falk [31], branch and bound algorithm by Bard and Moore [32] or double penalty function method by Ishizuka and Aiyoshi [33]. Recently, Walz et al. [34] presented a global SIP algorithm proposed in [35] that could be used in the context of optimal experiment design. Reference therein gives a complete picture about the use of SIP algorithms for the solution of bilevel programs.

Also various stochastic methods, such as genetic algorithms, simulated annealing, etc., could be used in principle, where these might be especially interesting for the D design problem (18) because of its non-smoothness. We only exploit the described nested and KKT-based approaches in this study.

5. Case studies

We apply the presented methodologies for finding OED for two small-scale illustrative case studies. The employed models are in the form of explicit step responses of linear time-invariant dynamic systems and the optimal experiment design should reveal the best sampling instants. We denote the designs that are based on the exact CRs (problems (15), (18), and (21)) as *exact* designs. The OED problems are solved for 2σ-confidence level ($\alpha = 0.9545$) using the following approaches:

1. Classical OED problems are solved globally.
2. The designs based on exact CRs and ellipsoidal D design are solved using nested approach where the lower-level problem is solved globally and the upper-level problem is solved using a local solver.
3. In order to study numerical efficacy of different algorithms, the A-optimal design is solved by KKT-based approach globally.

5.1. Case study 1

The mathematical model for biological oxygen demand (BOD) [7] is considered. The cumulative BOD of microorganisms at incubation time u_τ is given by

$$\hat{y}(\boldsymbol{p}, \tau) = p_1(1 - \exp(-p_2 u_\tau)), \quad u_\tau \in [0, 20], \tag{28}$$

which can also be interpreted as a step response of the first-order linear time-invariant system with static gain p_1 and time constant $1/p_2$. At this point it can be observed that p_1 enters the output equation linearly while p_2 enters nonlinearly.

The true values for the parameters p_1 and p_2 are, respectively, 2.5 and 0.5. These are also considered as expected least-squares estimates $\hat{\boldsymbol{p}}$ for all the studied OED problems. The measurements $y_m(\tau)$ are assumed to be corrupted by a zero-mean Gaussian noise with the standard deviation 0.1. We will assume here a case where variance or the standard deviation of the measurement noise is unknown. The exact CRs are then defined by (12). Additionally, we consider $J(\hat{p}) = 0$. The tolerance ϵ for the exact D design is set to 5×10^{-3}.

The optimal sampling times $\mathcal{U}^*$ with $N = \{4, 5\}$ for the classical and exact A-, D- and E-optimal designs and the ellipsoidal D design are reported in Table 1. The values of the objective function for exact

Table 1
Optimal designs ($\mathcal{U}^*$) as identified for classical, ellipsoidal and the exact OED using $N = \{4, 5\}$ and the values of objective function of exact designs ($\phi(\cdot)$) evaluated at the identified optimal designs ($\mathcal{U}^*$). In the case of D design, $\phi(\mathcal{U}^*) := \hat{\phi}_D(\mathcal{U}^*)$.

	Design	N	Solution ($\mathcal{U}^*$)	$\phi(\mathcal{U}^*)$
Classical OED	A	4	{1.69, 1.69, 20, 20}	1.610
	A	5	{1.77, 1.77, 20, 20, 20}	0.940
	D	4	{2, 2, 20, 20}	0.425
	D	5	{2, 2, 20, 20, 20}	0.155
	E	4	{1.61, 20, 20, 20}	1.016
	E	5	{1.75, 20, 20, 20, 20}	0.365
Ellips.	D	4	{1.42, 1.42, 20, 20}	0.414
	D	5	{1.69, 1.69, 19.99, 19.99, 20}	0.154
Exact OED	A	4	{1.37, 1.37, 20, 20}	1.585
	A	5	{1.60, 1.60, 20, 20, 20}	0.938
	D	4	{1.62, 1.62, 20, 20}	0.409
	D	5	{1.81, 1.82, 1.83, 19.99, 19.99}	0.154
	E	4	{1.04, 1.04, 20, 20}	0.974
	E	5	{1.22, 1.23, 20, 20, 20}	0.322

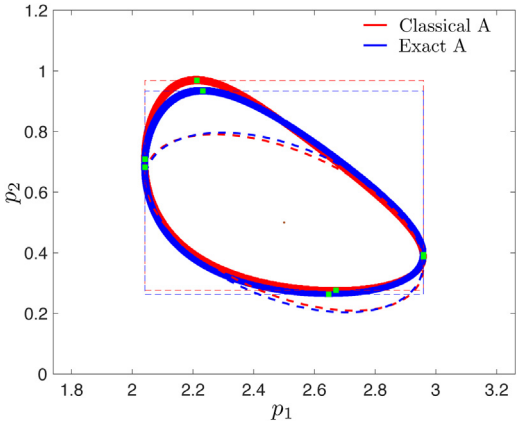


Fig. 2. The exact (solid) and linearized (dashed ellipsoid) CRs using $N = 4$ as obtained by classical and exact A designs. The plot shows the over-approximating orthotopes (dashed) of the exact CRs identified using the anchor points $\boldsymbol{\pi}$ represented by ■.

designs $\phi(\cdot)$ are evaluated at the identified optimal solution $\mathcal{U}^*$ for each design with $N = \{4, 5\}$. For all the designs, the exact OED has a superior performance when compared to the linearized design. The $\mathcal{U}^*$ as identified by the classical and the exact OED contain multiple common repetitive measurements at $u_\tau^* = 20$. This agreement between the designs can be attributed to the linear entry of p_1 into the model $\hat{y}(\tau)$. It can also further be reasoned by an obvious fact; that one can infer the steady-state gain irrespective of the value of time constant closer to the steady state. The classical and the exact A OED had same number of repetitions of $u_\tau^* = 20$ using $N = \{4, 5\}$, however this is not the case for D-optimal design where only using $N = 4$ the number of repetitions match. In case of the E-optimal design, the classical OED identified $\mathcal{U}^*$ in which $u_\tau^* = 20$ is repeated once more when compared to the solution identified by the exact OED, which again points to the linear decorrelation between p_1 and p_2 that classical E design tries to achieve. The performance of the proposed D-optimal design based on the inner- and the outer-approximation ellipsoids is better when compared to the classical D design for all values of N. It achieves a relatively small loss in performance when compared to the exact D design. This suggests that the proposed ellipsoidal D design is an interesting framework to perform approximate D design as compared to linearization-based alternative.

The exact CRs for the classical (—) and the exact (—) A OED for four measurements are compared in Fig. 2. The orthotopes enclosing the exact CRs are plotted using the anchor points $\boldsymbol{\pi}$ represented by ■. It is clear that the orthotope that encloses the exact CR identified by the

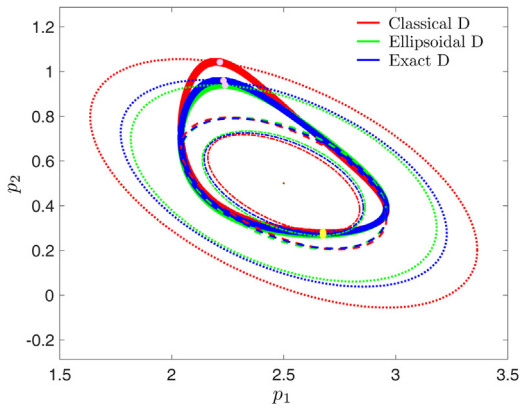


Fig. 3. The exact (solid) and linearized (dashed ellipsoid) CRs using $N = 4$ as obtained by classical, ellipsoidal and exact D designs. The plot shows the outer-/inner-approximating ellipsoids of the exact CRs (dotted and dash-dotted lines, respectively). ● and ● are the intersection points for the outer-/inner-approximating ellipsoids and the exact CRs.

classical A design (- - -) has a larger perimeter when one compares it with the orthotope identified by the exact A design. The reason for this can be found when looking at the linearized CRs for both designs (dashed ellipsoids). The linearized ellipsoid clearly does not approximate the exact CRs well, where, as it can be expected, the approximation is looser w.r.t. p_2 that enters nonlinearly in output equation (28). It is an interesting observation that the presented linearized CRs are very similar to each other, despite the fact that they give significantly different exact CRs.

Fig. 3 shows the exact CRs for the classical (—), ellipsoidal (—) and the exact (—) D designs using $N = 4$. The linearized CRs (dashed ellipsoids) are again very similar to each other. Among them the ellipsoid from the classical design (- - -) has the smallest volume, as might be expected. However, the exact CR for the classical design is the largest one (see Table 1), which again comes from disregarding of nonlinearity of the output equation in p_2 by the classical design.

We also present the inner- and the outer-approximation ellipsoids for all the three OED approaches by dash-dotted and dotted ellipsoids, respectively. The corresponding intersection points are marked by ● and ● respectively in Fig. 3. Here we can observe the benefits of weighting introduced in the objective function of the problem (19). While looking at the sizes of outer-approximating ellipsoids (especially the one constructed for exact D design), it might appear reasonable to minimize the volume of the over-approximating ellipsoid as a good approximation of the exact D design. This would correspond to setting $1/k_{in} \to 0$ while solving the problem (19). We have explored this path in our earlier study [11], but the obtained design results were unsatisfactory, since the over-approximation by an ellipsoid might become very loose. The proposed ellipsoidal D design therefore balances out the concentration on the size of the ellipsoid and the appropriateness of the over-approximation by an ellipsoid. It is clearly visible that the exact CR for the $\mathcal{U}^*$ identified by the proposed ellipsoidal OED has much smaller volume than the exact CR identified by the classical approach and, at the same time, it is very close to the optimal exact OED.

The exact CRs for the classical (—) and the exact (—) E designs are compared in Fig. 4 using $N = 4$ measurements. ● in Fig. 4 mark $\boldsymbol{\varphi}_1$ and $\boldsymbol{\varphi}_2$ (see (21)) obtained for the classical and the exact E-optimal designs. In this case, unlike for the previous designs, we observe a major discrepancy in the orientation between the linearized ellipsoids obtained for the classical and exact designs, despite the largest semi-axes, which correspond to the largest eigenvalues (see (22)) are very similar. Again this is attributed to the nonlinearity in p_2. The resulting difference of distances between the most dis-

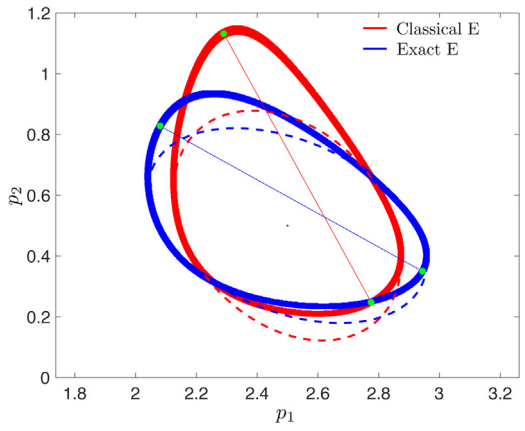


Fig. 4. The exact (solid) and linearized (dashed ellipsoid) CRs using $N = 4$ as obtained for classical and exact E OED. ● mark the points used to calculate the Euclidean distance of the CRs.

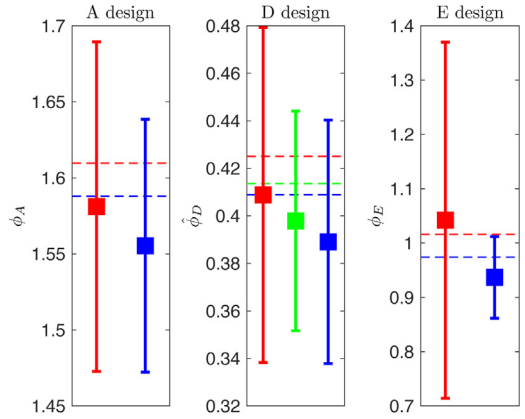


Fig. 5. Mean and variance of ϕ_A, $\hat{\phi}_D$ and ϕ_E for 1,000 random experiments with $N = 4$ noisy measurements at $\mathcal{U}^*$ of classical (—), ellipsoidal (—) and the exact (—) designs. Dashed line signifies the performance of the nominal design.

tant points that belong to P is significant, however, among the two designs. Similar behavior can be seen for the case with $N = 5$ (see Table 1).

Next, we study the performance of the obtained designs against a number of simulated experiments. The aim here is to evaluate robustness of the designs against random realization of noise that would be present in the real experiment. We exclude the dependence on the least-squares estimate here, i.e., we will use the nominal values for $\hat{\boldsymbol{p}}$. Such dependence is the subject of study for *robust* OED, which is not in the scope here. We simulated 1,000 experiments with each studied design using the obtained optimal incubation (measurement) times $\mathcal{U}^*$ and we corrupted the measurements $y_m(\tau) := \hat{y}(\hat{\boldsymbol{p}}, \boldsymbol{\tau}) + e$ with a Gaussian noise e of standard deviation 0.1.

Fig. 5 shows the mean and the variance of the objective value of the exact A, D and E designs. In this plot, we also include the nominal values (when noise-free measurements are gathered) of the different designs using dashed lines. With respect to the mean values of obtained ϕ_A, $\hat{\phi}_D$ and ϕ_E, it is confirmed that exact OED is the best option on average, despite the mean values do not match the expected nominal values of the design criteria. Regarding the obtained variances, it is noteworthy that the classical design exhibits the strongest sensitivity to noise as it can be concluded from the magnitude of the variances and thus appears to be the worst option. This again underpins the importance of consideration of the nonlinearity in the OED and it can be documented by comparing the worst-case value of the exact E design w.r.t. nominal mean obtained for the linearized design (right plot). The last interesting observation is that the variance of $\hat{\phi}_D$ obtained with ellipsoidal design is slightly smaller than the variance of the exact OED. This might point to the approximation error introduced (tolerance ϵ in (18)) in calculating $\hat{\phi}_D$, which is however not severe as the mean and worst-case values follow the expected trend.

Table 2
Optimal designs ($\mathcal{U}^*$) as identified for classical, ellipsoidal and the exact OED using $N = \{2, 3, 4\}$ and the values of objective function of exact designs ($\phi(\cdot)$) evaluated at the identified optimal designs ($\mathcal{U}^*$). In the case of D design, $\phi(\mathcal{U}^*) := \hat{\phi}_D(\mathcal{U}^*)$.

	Design	N	Solution ($\mathcal{U}^*$)	$\phi(\mathcal{U}^*)$
Classical OED	A	2	{1.91, 10}	1.666
	A	3	{1.86, 1.86, 10}	1.151
	A	4	{1.81, 1.81, 1.81, 10}	0.974
	D	2	{2, 10}	0.386
	D	3	{2, 2, 10}	0.231
	D	4	{2, 2, 10, 10}	0.148
	E	2	{1.90, 10}	1.225
	E	3	{1.82, 1.82, 10}	0.520
	E	4	{1.74, 1.74, 1.74, 10}	0.341
Ellipsoid	D	2	{1.70, 10}	0.363
	D	3	{1.73, 1.73, 10}	0.219
	D	4	{1.82, 1.82, 10, 10}	0.144
Exact OED	A	2	{1.63, 10}	1.584
	A	3	{1.67, 1.67, 10}	1.132
	A	4	{1.66, 1.66, 1.67, 10}	0.966
	D	2	{1.61, 10}	0.344
	D	3	{1.65, 1.66, 10}	0.218
	D	4	{1.74, 1.77, 10, 10}	0.144
	E	2	{1.62, 10}	1.094
	E	3	{1.63, 1.63, 10}	0.497
	E	4	{1.59, 1.59, 1.59, 10}	0.331

5.2. Case study 2

Here we consider a problem where the system output can be modeled as

$$\hat{y}(\tau) = b_0 \frac{p_1}{p_2^2} \left(\left[p_2 \frac{p_1 + p_2}{p_1} u_\tau + 1 \right] \exp(-p_2 u_\tau) - 1 \right), \tag{29}$$

which can also be interpreted as a step response of the second-order linear time-invariant system with a double pole at $-p_2$ and a zero at p_1. The corresponding transfer function can be given by

$$G(s) = \frac{b_0 (s - p_1)}{(s + p_2)^2}. \tag{30}$$

Clearly the constant b_0 is a parameter that influences the steady-state gain of the system and we will assume it, for simplicity, to be known $b_0 = -4$.

Notice that both p_1 and p_2 enter the output equation nonlinearly, so this problem can be considered as more challenging and even greater discrepancy might be expected between linearized and exact designs. The true values of the parameters p_1 and p_2 are 0.5 and 1.0, respectively, and are equal to $\hat{\boldsymbol{p}}$. The measurements $y_m(\tau)$ are assumed to be corrupted with a zero-mean Gaussian noise with known standard deviation of 0.4. For a fair comparison of the proposed framework, we assume $J_w(\hat{\boldsymbol{p}}) = 0$. The tolerance ($\epsilon$) for exact D design is 7.5×10^{-2}.

The classical and the proposed OED frameworks are applied to identify the optimal sampling times $\mathcal{U}^*$, where $u_\tau^* \in [0, 10]$. The previously discussed OED problems are solved for $N = \{2, 3, 4\}$ and $\alpha = 0.9545$ (2σ-CR) with the same numerical techniques as in case study 1. The results are presented in Table 2. The trends regarding the performance of the different designs are the same as described in case study 1. We observe in this case a more significant benefit of employing an ellipsoidal D design, which, for $N = \{3, 4\}$ almost

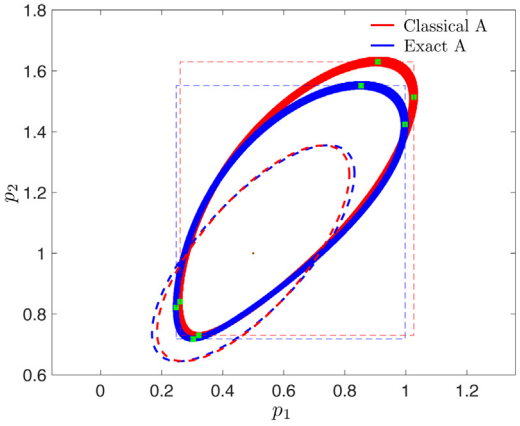


Fig. 6. The exact (solid) and linearized (dashed ellipsoid) CRs using $N=2$ as obtained for classical and exact A OED. The plot shows the over-approximating orthotopes (dashed) of the exact CRs identified using the anchor points $\boldsymbol{\pi}$ represented by ■.

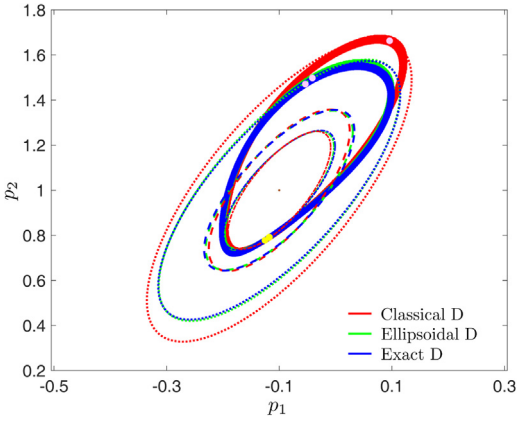


Fig. 7. The exact (solid) and linearized (dashed ellipsoid) CRs using $N=2$ as obtained for classical, ellipsoidal and exact D OED. The plot shows the outer-/inner-approximating ellipsoids of the exact CRs (dotted and dash-dotted lines, respectively). ● and ● are the intersection points for the outer-/inner-approximating ellipsoids and the exact CRs.

reaches the performance of exact OED and is superior to classical OED.

In Fig. 6, we compare the exact and linearized CRs obtained for the classical (▬) and the exact (▬) A design using $N=2$. We can see that the linearized CR captures well the orientation of the exact CR but due to nonlinearity the approximation is relatively poor. Similarly to the first case study we can observe that, despite very similar orientation of the linearized confidence regions, significant benefits of the exact design over the linearized one are obtained. Unlike in case study 1, we obtain reduction in the range on both parametric axes, which is caused by the nonlinearity of the output equation w.r.t. both parameters.

Fig. 7 shows the resulting CRs for D design criterion. This shows the reason for the good performance of the proposed ellipsoidal technique, which is able to tackle the nonlinearity of the CR far better than the linearization-based design.

The E-optimal designs for the classical and the exact OED with $N=2$ are compared in Figs. 8 and 9. Despite the fact that the linearized CRs show great similarity and they capture the orientation of the exact CR, the exact design tackles the nonlinearity far better and shows clear benefits w.r.t. the linearization-based counterpart.

Robustness of the obtained designs was tested against the random realization of the measurement noise for $N=4$ in the same way as in the previous case study. It is clear again that exact designs outperform the classical OED, which reaches the largest variances and

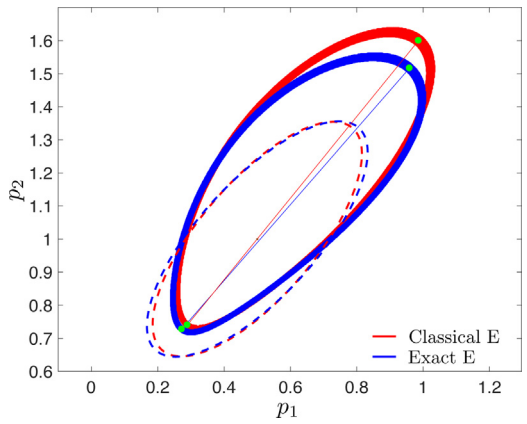


Fig. 8. The exact (solid) and linearized (dashed ellipsoid) CRs using $N=2$ as obtained for classical and exact E OED. ● mark the points used to calculate the Euclidean distance of the CRs.

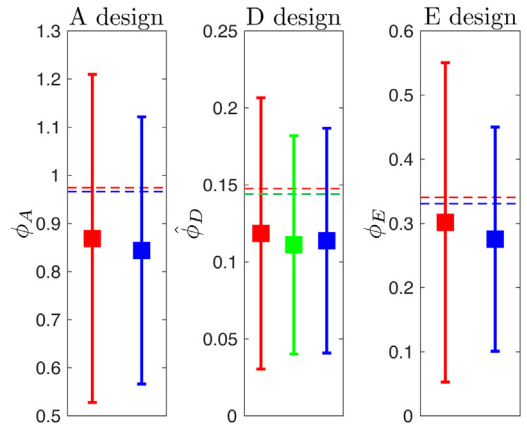


Fig. 9. Mean and variance of ϕ_A, $\hat{\phi}_D$ and ϕ_E for 1, 000 random experiments using $N=4$ noisy measurements at $\mathcal{U}^*$ of classical (▬), ellipsoidal (▬) and the exact (▬) OED. Dashed line signifies the performance of the nominal design.

the inferior means. The performance of the ellipsoidal D design is practically the same the performance of exact D design. In comparison with the case study 1, we observe larger variance of the exact D design, which we attribute to the higher nonlinearity.

5.3. Discussion

We studied cases where the CR is found for 2σ confidence. For a greater confidence, the CR increases in size and nonlinearity affects it stronger. That is why even bigger differences can be expected between classical and exact OED and even greater benefits can be obtained by using exact design (see [11]).

Regarding the computational efficacy of the different studied problems, it must be clearly pointed out that there exists a high sensitivity of model outputs w.r.t. the model parameters in both case studies due to the presence of highly nonlinear exponential terms. The presence of highly nonlinear terms and the need for inversion of the Fisher information matrix to formulate the objective function in classical OED prohibit BARON from closing the optimality gap and thus it returns locally optimal and sub-optimal solutions unless the optimization problem is properly initialized. However, the classical OED can be solved very efficiently, in few minutes on standard desktop workstation if initialized properly. The solution times for exact designs followed the expectations that result from the aforementioned complexity analysis (see Section 3). On average the solution procedure for exact A, D, and E design using nested approach took less than 10 min, ≈6 h, and 15–30 min, respectively. This shows that

the optimal exact A and E designs can be obtained with practically the same effort as in the case of the classical design for the small-scale problems. The exact A design procedure scales quadratically in n_p so it can get much more time-consuming in higher dimensions. We note that the reduction of CPU time obtained using ellipsoidal D design w.r.t. to exact D design was two-fold ($\approx$3 h), which is, on one hand, a considerable time saving but, on the other hand, it puts the potential user in question, whether the benefits prevail over the costs. We note, for completeness, that the KKT-based approach applied to problem of exact A design required the solution time of approximately one hour, which makes this approach clearly inferior.

In problems with small number of samples, it might be problematic to identify approximate (experimental) variance or to satisfy the asymptotic properties under which the CRs are defined. In this case, one might think of different approaches to experiment design. One such approach might be to relax the assumption of the presence of a white Gaussian noise in the measurements. This might in turn lead to set-membership estimation approach, also commonly known as guaranteed parameter estimation. A step in the direction of experimental design in set-membership context was taken in [36] and in the recent studies [34,37,38].

6. Conclusions

In this paper, exact and linearization-based methods were presented for the optimal experiment design of a nonlinear parameter estimation problem. The ellipsoidal method is proposed as a computationally less demanding counterpart to the exact D design, which is a computationally demanding problem since it requires a good approximation for the volume of a set. Two simple heuristic numerical methods are used here to solve the corresponding optimization problems, which are of bilevel nature. The OED framework is tested upon two illustrative small-scale nonlinear case studies, where the benefits of the exact design are showcased. The proposed ellipsoidal technique is shown to perform very well. Despite this study treated the case when the system model describes a static system in an explicit form, the methodology is straightforwardly applicable to dynamic systems and implicit model equations. An interesting direction for the future work lies, on one hand, in increasing the efficiency of the solution of the bilevel programs and, on the other hand, in the study of robust OED that relaxes the assumption of known (expected) least-squares estimates $\hat{\boldsymbol{p}}$, which might be relevant in practical tasks.

Acknowledgments

The research leading to these results has received funding from the European Commission under grant agreement number 291458 (ERC Advanced Investigator Grant MOBOCON). RP acknowledges the contribution of the European Commission under the project GuEst (grant agreement No. 790017), of the Slovak Research and Development Agency under the project APVV 15-0007, of the Scientific GrantAgency of the Slovak Republic under the grant 1/0004/17, and of the Research & Development Operational Programme for the project University Scientific Park STU in Bratislava, ITMS 26240220084, supported by the Research 7 Development Operational Programme funded by the ERDF.

References

[1] H. Hjalmarsson, From experiment design to closed-loop control, Automatica 41 (3) (2005) 393–438.
[2] L. Pronzato, Survey paper: optimal experimental design and some related control problems, Automatica 44 (2) (2008) 303–325.
[3] G. Franceschini, S. Macchietto, Model-based design of experiments for parameter precision: state of the art, Chem. Eng. Sci. 63 (19) (2008) 4846–4872.
[4] G.A.F. Seber, C.J. Wild, Nonlinear Regression, Wiley-Interscience, 2003.
[5] E. Beale, Confidence regions in non-linear estimation, J. R. Stat. Soc. Ser. B: Methodol. (1960) 41–88.
[6] L. Pronzato, A. Pázman, Design of Experiments in Nonlinear Models: Asymptotic Normality, Optimality Criteria and Small-Sample Properties, Springer, 2013.
[7] D.M. Bates, D.G. Watts, Nonlinear Regression Analysis and its Applications, Wiley Online Library, 1988.
[8] A.B. Kurzhanski, Ellipsoidal Calculus for Estimation and Feedback Control, Birkhäuser Boston, Boston, MA, 1997, pp. 229–243.
[9] W.C. Rooney, L.T. Biegler, Design for model parameter uncertainty using nonlinear confidence regions, AIChE J. 47 (8) (2001) 1794–1804.
[10] S. Streif, F. Petzke, A. Mesbah, R. Findeisen, R.D. Braatz, Optimal experimental design for probabilistic model discrimination using polynomial chaos, 19th IFAC World Congress, vol. 47, no. 3 (2014) 4103–4109.
[11] A.R. Gottu Mukkula, R. Paulen, Model-based optimal experiment design for nonlinear parameter estimation using exact confidence regions, 20th IFAC World Congress, vol. 50 (2017) 13760–13765.
[12] F. Dabbene, D. Henrion, C.M. Lagoa, Simple approximations of semialgebraic sets and their applications to control, Automatica 78 (2017) 110–118.
[13] S. Streif, N. Strobel, R. Findeisen, Inner approximations of consistent parameter sets by constraint inversion and mixed-integer programming, 12th IFAC Symposium on Computer Applications in Biotechnology, IFAC Proceedings Volumes, vol. 46, no. 31 (2013) 321–326.
[14] K. Ball, Ellipsoids of maximal volume in convex bodies, Geometr. Dedicat. 41 (2) (1992) 241–250.
[15] M. Tawarmalani, N.V. Sahinidis, A polyhedral branch-and-cut approach to global optimization, Math. Program. 103 (2005) 225–249.
[16] A. Mitsos, B. Chachuat, P.I. Barton, Towards global bilevel dynamic optimization, J. Global Optim. 45 (1) (2009) 63–93.
[17] P. Tanartkit, L. Biegler, A nested, simultaneous approach for dynamic optimization problems-I, Comput. Chem. Eng. 20 (6) (1996) 735–741, fifth International Symposium on Process Systems Engineering.
[18] P. Tanartkit, L. Biegler, A nested, simultaneous approach for dynamic optimization problems-II: the outer problem, Comput. Chem. Eng. 21 (12) (1997) 1365–1388.
[19] M. Ruben, S. Daniel, N. Daniel, D.P. Cesar, Coordination of distributed model predictive controllers using price-driven coordination and sensitivity analysis, 10th IFAC International Symposium on Dynamics and Control of Process Systems, IFAC Proceedings Volumes 46, no. 32 (2013) 215–220.
[20] A. Wächter, T.L. Biegler, On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming, Math. Program. 106 (1) (2006) 25–57.
[21] A.V. Fiacco, Y. Ishizuka, Sensitivity and stability analysis for nonlinear programming, Ann. Oper. Res. 27 (1) (1990) 215–235.
[22] S. Dempe, Bilevel Optimization: Reformulation and First Optimality Conditions, Springer Singapore, 2017, pp. 1–20.
[23] S. Dempe, V. Kalashnikov, G. Pérez-Valdés, N. Kalashnykova, Bilevel Programming Problems: Theory, Algorithms and Applications to Energy Networks, Energy Systems, Springer Berlin Heidelberg, 2015.
[24] S. Dempe, Foundations of Bilevel Programming, Nonconvex Optimization and its Applications, Springer, Boston, MA, 2002.
[25] M. Kieffer, E. Walter, Guaranteed estimation of the parameters of nonlinear continuous-time models: contributions of interval analysis, Int. J. Adapt. Control Signal Process. 25 (3) (2011) 191–207.
[26] J. Dutta, S. Dempe, Bilevel Programming with Convex Lower Level Problems, Springer US, Boston, MA, 2006, pp. 51–71.
[27] S. Dempe, S. Franke, On the solution of convex bilevel optimization problems, Comput. Optim. Appl. 63 (3) (2016) 685–703.
[28] J. Bard, Practical Bilevel Optimization: Algorithms and Applications, Nonconvex Optimization and its Applications, Springer US, 1998.
[29] A. Mitsos, P. Lemonidis, P.I. Barton, Global solution of bilevel programs with a nonconvex inner program, J. Global Optim. 42 (4) (2008) 475–513.
[30] O. Stein, G. Still, On generalized semi-infinite optimization and bilevel optimization, Eur. J. Oper. Res. 142 (3) (2002) 444–462.
[31] J.W. Blankenship, J.E. Falk, Infinitely constrained optimization problems, J. Optim. Theory Appl. 19 (2) (1976) 261–281.
[32] J.F. Bard, J.T. Moore, A branch and bound algorithm for the bilevel programming problem, SIAM J. Sci. Stat. Comput. 11 (2) (1990) 281–292.
[33] Y. Ishizuka, E. Aiyoshi, Double penalty method for bilevel optimization problems, Ann. Oper. Res. 34 (1) (1992) 73–88.
[34] O. Walz, H. Djelassi, A. Caspari, A. Mitsos, Bounded-error optimal experimental design via global solution of constrained min–max program, Comput. Chem. Eng. 111 (2018) 92–101.
[35] A. Mitsos, A. Tsoukalas, Global optimization of generalized semi-infinite programs via restriction of the right hand side, J. Global Optim. 61 (1) (2015) 1–17.
[36] D. Telen, B. Houska, F. Logist, M. Diehl, J. Van Impe, Guaranteed robust optimal experiment design for nonlinear dynamic systems, IEEE, 2013 European Control Conference (ECC) (2013) 2939–2944.
[37] A.R. Gottu Mukkula, R. Paulen, Model-based design of optimal experiments for nonlinear systems in the context of guaranteed parameter estimation, Comput. Chem. Eng. 99 (2017) 198–213.
[38] A.R. Gottu Mukkula, R. Paulen, Optimal design of dynamic experiments for guaranteed parameter estimation, 2016 American Control Conference (ACC) (2016) 1826–1831.

I&EC research
Industrial & Engineering Chemistry Research

Article

Cite This: Ind. Eng. Chem. Res. 2020, 59, 2396−2408

pubs.acs.org/IECR

Bayesian Approach to Probabilistic Design Space Characterization: A Nested Sampling Strategy

Kennedy P. Kusumo,[†] Lucian Gomoescu,[†,‡] Radoslav Paulen,[§] Salvador García Muñoz,[∥] Constantinos C. Pantelides,[†,‡] Nilay Shah,[†] and Benoît Chachuat*[,†]

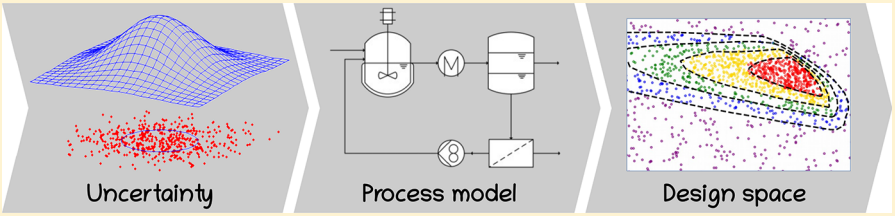
[†]Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, London SW7 2AZ, U.K.

[‡]Process Systems Enterprise, Ltd., London W6 7HA, U.K.

[§]Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, 812 43 Bratislava, Slovakia

[∥]Small Molecule Design and Development, Lilly Research Laboratories, Eli Lilly & Company, Indianapolis, Indiana 46285, United States

Supporting Information



ABSTRACT: Quality by design in pharmaceutical manufacturing hinges on computational methods and tools that are capable of accurate quantitative prediction of the design space. This paper investigates Bayesian approaches to design space characterization, which determine a feasibility probability that can be used as a measure of reliability and risk by the practitioner. An adaptation of nested sampling—a Monte Carlo technique introduced to compute Bayesian evidence—is presented. The nested sampling algorithm maintains a given set of live points through regions with increasing probability feasibility until reaching a desired reliability level. It furthermore leverages efficient strategies from Bayesian statistics for generating replacement proposals during the search. Features and advantages of this algorithm are demonstrated by means of a simple numerical example and two industrial case studies. It is shown that nested sampling can outperform conventional Monte Carlo sampling and be competitive with flexibility-based optimization techniques in low-dimensional design space problems. Practical aspects of exploiting the sampled design space to reconstruct a feasibility probability map using machine learning techniques are also discussed and illustrated. Finally, the effectiveness of nested sampling is demonstrated on a higher-dimensional problem, in the presence of a complex dynamic model and significant model uncertainty.

INTRODUCTION

Over the years the pharmaceutical industry has identified an increasing need for systematic and holistic approaches to drug development and manufacturing, which has led to a high penetration of process systems engineering (PSE) tools.[1,2] This need is prompted by safety concerns and regulations alongside growing pressure to increase efficiency, both in production and in process development. A recent estimate[3] amounts to US $2.6Bn for a single novel drug put on the market. To improve practices in the industry the International Council for Harmonization of Technical Requirements for Pharmaceuticals for Human Use (ICH) introduced the quality by design (QbD) initiative through a series of five guidelines: ICH Q8−12.[4] Historically, pharmaceutical development and manufacturing had emphasized checklist-based operations rather than scientific understanding. So, by promoting a scientific and risk-based approach to pharmaceutical product development and manufacturing, the QbD initiative and ICH guidelines triggered a paradigm shift in the industry and a complete new range of activities for the practitioners.

The QbD approach defines a quality target product profile (QTPP), which is a prospective summary of the quality characteristics of the pharmaceutical product that ensures the desired quality, safety, and efficacy. The (physical, chemical, biological, or microbiological) properties that should be within an appropriate limit, range, or distribution to ensure the desired product quality are called critical quality attributes (CQAs). A design space (DS) consists of "*the multidimensional combination and interaction of input variables (material attributes) and process parameters that have been demonstrated to provide assurance of quality*".[5] In practice a regulatory process

Special Issue: Christos Georgakis Festschrift

Received: September 7, 2019
Revised: November 25, 2019
Accepted: November 26, 2019
Published: November 26, 2019

DOI: 10.1021/acs.iecr.9b05006
Ind. Eng. Chem. Res. 2020, 59, 2396−2408

of postapproval change is not required so long as the process parameters vary within the limits of the approved DS.

A classical approach to DS characterization entails the following four steps:[6,7] (i) conduct a thorough experimental investigation of the relationships between process parameters and CQAs; (ii) analyze the sensitivity of the process parameters on the CQAs to only retain those parameters presenting a significant sensitivity; (iii) establish a mathematical or graphical representation of the DS using data-driven modeling and optimization techniques; and (iv) test the final DS by running further validation experiments, prior to submitting it for approval to the regulatory agencies. Multivariate statistical techniques, e.g., based on latent-variable modeling, have been proposed to reduce the cost and time needed to conduct these experiments.[8,9] Nevertheless, the presence of a large number of external disturbances and/or potential control manipulations lead to high-dimensional DS, whose effective characterization can be empowered by the use of process models.[10] Such process models may be either data-driven or knowledge-driven insofar as they describe the relationships between process parameters and CQAs accurately.

Characterizing a DS based on a process model is akin to the popular concepts of resilience and flexibility introduced by the PSE community during the 1980s.[11,12] Two classical problems in this literature are (i) the flexibility test,[13] which verifies that a feasible operation can be obtained for specific process parameters under a range of uncertainty scenarios—possibly by manipulating certain controls; and (ii) the flexibility index,[14] which represents the maximal deviation of the process parameters from given nominal values that can be tolerated in order for the operation to remain feasible under a range of uncertainty scenarios. A related concept is that of process operability,[15,16] which starts from an estimated range of the process uncertainty and calculates the desired ranges of the control variables so that a control strategy can guarantee a feasible operation.

Like the result of any model-based computation, the reliability of any DS determined on the basis of a process model depends on the accuracy of that model. If the uncertain parameters of the model follow a certain probability distribution then the DS itself is not a well-delineated region but a probabilistic one, where each point in the process parameter space is characterized by a probability of meeting all the established limits on CQAs.[17] Interestingly, a closely related concept to that of probabilistic DS was introduced in the PSE literature later during the 1990s under the name stochastic flexibility,[18,19] as a way of measuring the probability of a process design to operate feasibly in the presence of uncertainty. This probabilistic interpretation is in agreement with Peterson[20] who defines the DS as the set of process parameters such that the manufactured product satisfies all of the CQA limits alongside other process constraints with a probability greater than a given threshold.

Existing computational approaches to probabilistic DS characterization differ in how they account for process model uncertainty and how they approximate the DS itself. Process model uncertainty is represented in either one of two ways: (i) set-membership, most commonly a joint confidence ellipsoid at a given confidence level as derived from a frequentist parameter estimation;[10,21] or (ii) sampled distribution, for instance, the sampled parameter posterior in Bayesian estimation or a bootstrap parameter distribution.[9,20,22] Caution must be exerted in both cases since assuming an ellipsoidal region for non-normally distributed parameters or representing a probability distribution with an insufficient number of samples can lead to large errors in the predicted DS.

In terms of DS representation, one may broadly categorize the existing approaches as (i) design-centering algorithms; and (ii) sampling algorithms. The former consider a class of parametrized shapes—most commonly a box—and transform the DS characterization into an optimization problem that seeks to inscribe the largest possible shape within the DS. The flexibility-index problem[14,23] falls into this category, and it was recently applied to probabilistic DS computation.[21,24] A drawback with this approach is that a simple shape may not provide a good approximation of the DS and thus introduce significant conservatism. Another drawback is that design-centering problems give rise to complex mathematical programs with either robust (semi-infinite) or chance constraints that are computationally hard to tackle rigorously.[25,26] By contrast, sampling algorithms discretize the process parameter range and return a subset of the samples that satisfy all of the CQA limits and other constraints up to the desired reliability value. Exhaustive sampling may be achieved via a fine uniform gridding or (quasi) Monte Carlo sampling. The assessment of each parameter sample can be made via the solution of an optimization problem, which is akin to a flexibility test and is especially suited to a set-membership description of the model uncertainty.[17,21,27] If the model parameter uncertainty is represented by a sampled distribution instead, both Monte Carlo and Bayesian techniques can be used for propagating the uncertainty to the CQAs and estimate a feasibility probability.[9,22] These sampling techniques have proven effective in practice but they are computationally expensive and mainly tractable for low-dimensional DS at present. For computationally demanding process models surrogate-based approaches can be applied, including kriging, radial basis functions, and high-dimensional model representation.[28] For instance, surrogate-based adaptive sampling[29−31] has been shown to speed-up the computation of flexibility indices. But while it can be computationally cheaper to characterize the DS using a simple surrogate, one needs to account for the additional burden of constructing the surrogate itself as well as any additional approximation error that may be introduced.

Despite the large body of research on DS computation, there is a clear need for algorithms with improved computational efficiency to tackle industrially relevant problems that have more design parameters or greater uncertainty. The focus of this paper is on nested sampling, a Monte Carlo technique that was introduced by Skilling[32] for estimating the Bayesian evidence in parameter estimation. This algorithm proceeds by progressively sampling in nested contours of increasing likelihood, so as to maintain a dense enough sample in regions of higher likelihood, in the manner of adaptive sampling. An advantage of nested sampling over Markov chain Monte Carlo (MCMC) techniques is its better ability to handle multimodal posteriors.[33]

Our main contribution herein is an adaptation of nested sampling for the characterization of a probabilistic DS. In this context nested sampling determines a set of samples with a feasibility probability larger than a given reliability threshold. A byproduct of the algorithm is a second set of samples with a feasibility probability below the desired threshold, which provides an approximation of the entire probabilistic DS. In

DOI: 10.1021/acs.iecr.9b05006
Ind. Eng. Chem. Res. 2020, 59, 2396−2408

this manner nested sampling can offer a greater flexibility in selecting the probability threshold to provide assurance of quality. An important motivation behind the adaptation of nested sampling is the availability of efficient strategies for generating replacement proposals as the algorithm progresses,[33,34] which we can leverage to the benefit of DS characterization. Another appeal is the ability to readily exploit a sampled joint posterior distribution of the model parameters: by combining the proposed approach with a Bayesian estimation procedure for the model parameters, one can arrive at a truly Bayesian approach to DS characterization directly from experimental data. Last but not least importantly nested sampling is, like other sampling-based techniques, nonintrusive in the sense that it relies on the result of model simulations at given process parameter values only. In principle, this allows for black-box models such as a process flowsheet or a CFD model to be used for DS characterization.

The rest of the paper is organized as follows: In the following section we review the mathematical formulation of the design space characterization problem. Then we present the nested sampling approach and use a simple example to illustrate its main features. We also test the method on two case studies of increasing complexity and discuss the results, before concluding the paper.

PROBLEM STATEMENT

Consider a manufacturing process for a pharmaceutical product that has its quality defined by some CQAs, denoted by $\mathbf{s} \in \mathbb{R}^{n_s}$. Assume that a mathematical model of the process (either knowledge- or data-driven) is available that predicts the CQAs corresponding to a set of process parameters, denoted by $\mathbf{d} \in \mathcal{K}$ within the knowledge space $\mathcal{K} \subset \mathbb{R}^{n_d}$:

$$\mathbf{s} = \mathbf{f}(\mathbf{d}, \boldsymbol{\theta}) \tag{1}$$

The model parameters, $\boldsymbol{\theta} \in \mathbb{R}^{n_\theta}$ may represent physical constants, coefficients in a regression model, or disturbances that affect the CQAs. The mapping $\mathbf{f}: \mathbb{R}^{n_d} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_s}$ needs not be given in closed-form but could be implicitly defined by a set of algebraic or differential equations or even black-box functions such as a process simulator. Assume furthermore that the CQA limits are represented alongside other process constraints by the following inequalities:

$$\mathbf{G}(\mathbf{d}, \boldsymbol{\theta}) := \mathbf{g}(\mathbf{d}, \mathbf{f}(\mathbf{d}, \boldsymbol{\theta})) \leq \mathbf{0} \tag{2}$$

Notice that the variable $\mathbf{s}$ can be abstracted away from the inequality constraints without loss of generality. Similarly to the mapping $\mathbf{f}$ above, the constraint function $\mathbf{G}: \mathbb{R}^{n_d} \times \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_g}$ is not necessarily available in closed form.

Ignoring the uncertainty in the model parameters leads to a nominal design space:

$$\mathcal{D}_{\text{nom}} := \{\mathbf{d} \in \mathcal{K}: \mathbf{G}(\mathbf{d}, \boldsymbol{\theta}_{\text{nom}}) \leq \mathbf{0}\} \tag{3}$$

for a given nominal value $\boldsymbol{\theta}_{\text{nom}}$ of the model parameters. However, the value of $\boldsymbol{\theta}$ is inherently uncertain in practice by the nature of the modeling exercise. A Bayesian framework considers $\boldsymbol{\theta}$ as random variables with a joint distribution $p(\boldsymbol{\theta})$ that describes the belief on the value of $\boldsymbol{\theta}$. For instance, $p(\boldsymbol{\theta})$ could be estimated from experimental data using Bayesian inference. In this framework the model may only be used to predict the probability that the manufacturing process is feasible for a given $\mathbf{d} \in \mathcal{K}$, which is akin to a stochastic flexibility test:[18]

$$\mathbb{P}[\mathbf{G}(\mathbf{d}, \cdot) \leq \mathbf{0} \mid p(\boldsymbol{\theta})] := \int_{\{\boldsymbol{\theta}: \mathbf{G}(\mathbf{d}, \boldsymbol{\theta}) \leq \mathbf{0}\}} p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \tag{4}$$

The problem of interest throughout this paper is to determine the probabilistic DS given by

$$\mathcal{D}_\alpha := \{\mathbf{d} \in \mathcal{K}: \mathbb{P}[\mathbf{G}(\mathbf{d}, \cdot) \leq \mathbf{0} \mid p(\boldsymbol{\theta})] \geq \alpha\} \tag{5}$$

where $0 < \alpha \leq 1$ is the so-called reliability value.[20] Notice that the following set-membership counterpart to the probabilistic DS is often computed instead of $\mathcal{D}_\alpha$ in practice:[21]

$$\hat{\mathcal{D}}_\alpha := \{\mathbf{d} \in \mathcal{K}: \forall\, \boldsymbol{\theta} \in \Theta_\alpha, \mathbf{G}(\mathbf{d}, \boldsymbol{\theta}) \leq \mathbf{0}\} \tag{6}$$

with Θ_α chosen as the highest posterior density (HPD) set such that $\int_{\boldsymbol{\theta} \in \Theta_\alpha} p(\boldsymbol{\theta}) = \alpha$. However, $\mathcal{D}_\alpha$ and $\hat{\mathcal{D}}_\alpha$ are not equivalent in general when the constraints are nonlinear or the model parameters are not normally distributed.

According to the monotonicity property that $\mathcal{D}_\alpha \supset \mathcal{D}_{\alpha'}$ whenever $\alpha < \alpha'$, a higher reliability value increases conservatism by shrinking the DS. A practical choice for α entails trading-off the risk of violating the CQA limits against the loss of operational flexibility: a lower α value increases false positives in $\mathcal{D}_\alpha$ by including design parameters that will not fulfill all of the CQA limits and process constraints as expected; while a higher α value increases false negatives by excluding design parameters that are in fact feasible. Because false positives pose a threat to the assurance of quality, practitioners are prompted to be conservative. But the choice of α remains specific to the nature of the process at hand and to the risks involved.

It is also worth noting an important assumption underlying the definition of $\mathcal{D}_\alpha$ (or $\hat{\mathcal{D}}_\alpha$), namely, that the process model is either structurally correct or that any structural mismatch can be reported in terms of parametric uncertainty. A practitioner's lacking confidence in the process model structure can always increase conservatism by opting for a higher reliability value α. However, the probabilities conveyed by $\mathcal{D}_\alpha$ could become misleading in the presence of large structural mismatch. A possible remedy entails the consideration of multiple candidate models under the Bayesian framework and the restatement of $\mathbb{P}[\mathbf{G}(\mathbf{d}, \cdot) \leq 0 \mid p(\boldsymbol{\theta})]$ to explicitly consider structural uncertainty.

Illustrative Example. We consider a simple case study with two design variables, $\mathbf{d} := (d_1, d_2)^{\mathrm{T}}$ and a single CQA, s. We assume the following relationship between the design variables and the CQA:

$$s = \theta d_1^2 + d_2 \tag{7}$$

with the uncertain parameter θ. The goal is to characterize the probabilistic DS inside the knowledge space $\mathcal{K} := [-1,1]^2$ imposed by the following CQA limits:

$$0.20 \leq s \leq 0.75 \tag{8}$$

We consider the nominal case to be given by $\theta_{\text{nom}} = 1$, which gives the following nominal design space:

$$\mathcal{D}_{\text{nom}} := \{\mathbf{d} \in [-1,1]^2: 0.20 \leq d_1^2 + d_2 \leq 0.75\} \tag{9}$$

Under a normality assumption for the model parameters, $\theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta)$ the probabilistic DS can be expressed analytically as

DOI: 10.1021/acs.iecr.9b05006
Ind. Eng. Chem. Res. 2020, 59, 2396−2408

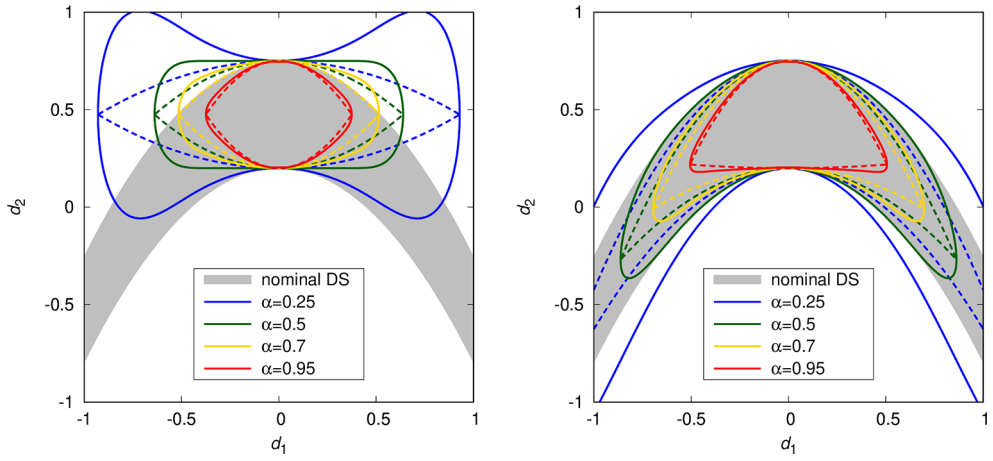


Figure 1. Depiction of the design space for the illustrative example. The nominal design space in eq 9 is represented with the gray-shaded area. The probabilistic design spaces corresponding to different reliability values α in eq 10 are shown in solid lines with different colors. The set-membership counterparts in eq 11 are shown in dashed lines. The left and right plots are for the uncertainty scenarios $\theta \sim \mathcal{N}(0,1)$ and $\theta \sim \mathcal{N}(1, \sqrt{0.3})$, respectively.

Algorithm 1 Standard Monte Carlo Sampling for Design Space Characterization

1: **input** $\mathcal{S}_d = \{\mathbf{d}_i \in \mathcal{K} : i = 1, \ldots, N_d\}$, $\mathcal{S}_\theta = \{(\boldsymbol{\theta}_j, \omega_j) \sim p(\cdot) : j = 1, \ldots, N_\theta\}$
2: $\mathcal{DS} \leftarrow \emptyset$
3: **for all** $\mathbf{d}_i \in \mathcal{S}_d$ **do**
4: $\quad \mathcal{DS} \leftarrow \mathcal{DS} \cup \left\{\left(\mathbf{d}_i\,;\, \widehat{\mathbb{P}}[\mathbf{G}(\mathbf{d}_i, \cdot) \le \mathbf{0} \mid \mathcal{S}_\theta]\right)\right\}$
5: **end for**
6: **return** $\mathcal{DS}$

$$\begin{aligned}\mathcal{D}_\alpha &:= \{\mathbf{d} \in [-1,1]: \mathbb{P}[0.20 \le \theta d_1^2 + d_2 \le 0.75 \mid \theta \sim \mathcal{N}(\mu_\theta, \sigma_\theta)] \ge \alpha\} \\ &= \left\{\mathbf{d} \in [-1,1]: \operatorname{erf}\left(\frac{0.75 - \mu_\theta d_1^2 - d_2}{\sqrt{2}\sigma_\theta d_1^2}\right) - \operatorname{erf}\left(\frac{0.20 - \mu_\theta d_1^2 - d_2}{\sqrt{2}\sigma_\theta d_1^2}\right) \ge 2\alpha\right\}\end{aligned} \tag{10}$$

Figure 1 compares $\mathcal{D}_\alpha$ for two parameter uncertainty scenarios, $\theta \sim \mathcal{N}(0,1)$ in the left plot and $\theta \sim \mathcal{N}(1, \sqrt{0.3})$ in the right plot, and different reliability values α. As expected, a more precise and accurate description of the model uncertainty enables a more favorable trade-off between CQA satisfaction and operational flexibility. With an imprecise and inaccurate uncertainty description (left plot), avoiding false positives requires a high reliability value (e.g., α = 0.95), but this results in a small DS and many false negatives. As the precision and accuracy of the model parameter improves (right plot), a lower reliability value may be selected without generating false positives (e.g., α = 0.5) and at the same time enable a much larger DS.

For comparison, Figure 1 also shows the DS counterpart for a set-membership uncertainty description based on eq 6, which can also be expressed analytically as

$$\begin{aligned}\hat{\mathcal{D}}_\alpha &:= \{\mathbf{d} \in [-1,1]: \forall\, \theta \in \mu_\theta \pm z_\alpha \sigma_\theta,\ 0.20 \le \theta d_1^2 + d_2 \le 0.75\} \\ &= \{\mathbf{d} \in [-1,1]: 0.20 - (\mu_\theta - z_\alpha\sigma_\theta)d_1^2 \le d_2 \le 0.75 - (\mu_\theta + z_\alpha\sigma_\theta)d_1^2\}\end{aligned} \tag{11}$$

with the quantile $z_\alpha = \sqrt{2}\,\mathrm{erf}^{-1}(\alpha)$. For the same reliability value α, $\hat{\mathcal{D}}_\alpha$ is generally more conservative than $\mathcal{D}_\alpha$. And the mismatch between $\mathcal{D}_\alpha$ and $\hat{\mathcal{D}}_\alpha$ furthermore reduces as α gets closer to 1. The rest of the paper focuses on computational approach to characterizing the Bayesian design space $\mathcal{D}_\alpha$.

METHODOLOGY

Our main focus is on computational methods that rely on sampling to characterize a probabilistic design space. These methods are appealing in that they allow for the use of a process model in explicit or implicit form and they can accommodate any probability distribution for the model parameters. A pseudocode for the standard Monte Carlo approach is presented in Algorithm 1, where both the process parameter space range and the model uncertainty distribution are discretized. At each process parameter sample, $\mathbf{d}_i \in \mathcal{S}_d$, the probability of CQA satisfaction is estimated via an ensemble of simulations over the uncertainty scenarios in $\mathcal{S}_\theta$:

$$\widehat{\mathbb{P}}[\mathbf{G}(\mathbf{d}_i, \cdot) \le \mathbf{0} \mid \mathcal{S}_\theta] := \sum_{(\boldsymbol{\theta}_j, \omega_j) \in \mathcal{S}_\theta} \mathbb{I}[\mathbf{G}(\mathbf{d}_i, \boldsymbol{\theta}_j)]\omega_j \tag{12}$$

where $\mathbb{I}[\cdot]$ stands for the indicator function, such that $\mathbb{I}[\mathbf{g}] = 1$ if $g_k \le 0, \forall k$ and $\mathbb{I}[\mathbf{g}] = 0$ otherwise; and the samples $\boldsymbol{\theta}_j$ with corresponding weights ω_j are drawn from the uncertainty distribution $p(\cdot)$. Although effective, this approach can prove computationally prohibitive because the number of model simulations grows as the product between the numbers of process parameter samples, N_d and uncertainty scenarios, N_θ. A major challenge for making such methods more efficient thus entails reducing the number of model simulations. The nested sampling algorithm described next aims precisely at sampling

DOI: 10.1021/acs.iecr.9b05006
Ind. Eng. Chem. Res. 2020, 59, 2396−2408

Table 1. Comparison of Nested Sampling for Parameter Estimation and Design Space Characterization

Bayesian parameter estimation	Bayesian design space characterization
Outcome	
An estimate of the Bayesian evidence Z	A set of samples of given size within the probabilistic design space at the desired reliability value α^*, $\{\mathbf{d}_i \in \mathcal{D}_{\alpha^*}: i = 1, ..., N_d\}$
A set of weighted samples from the posterior, $\{(\boldsymbol{\theta}_j,\omega_j) \sim p(\cdot): j = 1, ..., N_\theta\}$	A set of samples at a lesser reliability value $\alpha < \alpha^*$, $\{\mathbf{d}_k \in \mathcal{D}_{\alpha_k}: \alpha_k < \alpha^*, k = 1, ..., N'_d\}$
Challenge	
How to efficiently sample the prior distribution subject to a likelihood constraint, $\mathcal{L}(\boldsymbol{\theta}) > \mathcal{L}_{\min}$ so as to have a high acceptance rate and uniform spread	How to efficiently sample the design space subject to a feasibility probability constraint, $\hat{\mathbf{P}}[\mathbf{G}(\mathbf{d}, \cdot) \le 0] > \hat{\mathbf{P}}_{\min}$ so as to have a high acceptance rate and uniform spread

Algorithm 2 Nested Sampling for Design Space Characterization

1: **input** $\mathcal{K}$, $\mathcal{S}_L = \{\mathbf{d}_i \in \mathcal{K} : i = 1, \ldots, N_L\}$, $\mathcal{S}_\theta = \{(\boldsymbol{\theta}_j, \omega_j) \sim p(\cdot) : j = 1, \ldots, N_\theta\}$, α^*, N_R
2: $\mathcal{DS} \leftarrow \emptyset$
3: **while** $\exists \mathbf{d}_i \in \mathcal{S}_L : \widehat{\mathbb{P}}[\mathbf{G}(\mathbf{d}_i, \cdot) \le \mathbf{0} \mid \mathcal{S}_\theta] < \alpha^*$ **do**
4: $\quad \widehat{\mathbb{P}}_{\min} \leftarrow \min\{\widehat{\mathbb{P}}[\mathbf{G}(\mathbf{d}_i, \cdot) \le \mathbf{0} \mid \mathcal{S}_\theta] : \mathbf{d}_i \in \mathcal{S}_L\}$
5: $\quad \mathbf{d}_{\min} \leftarrow \arg\min\{\widehat{\mathbb{P}}[\mathbf{G}(\mathbf{d}_i, \cdot) \le \mathbf{0} \mid \mathcal{S}_\theta] : \mathbf{d}_i \in \mathcal{S}_L\}$
6: $\quad \mathcal{S}_R \leftarrow \{\mathbf{d}_k \in \mathcal{K} : k = 1, \ldots, N_R\}$
7: $\quad$ **for all** $\mathbf{d}_k \in \mathcal{S}_R$ **do**
8: $\quad\quad$ **if** $\widehat{\mathbb{P}}[\mathbf{G}(\mathbf{d}_k, \cdot) \le \mathbf{0} \mid \mathcal{S}_\theta] > \widehat{\mathbb{P}}_{\min}$ **then**
9: $\quad\quad\quad \mathcal{S}_L \leftarrow \mathcal{S}_L \cup \{\mathbf{d}_k\} \setminus \{\mathbf{d}_{\min}\}$
10: $\quad\quad\quad \mathcal{DS} \leftarrow \mathcal{DS} \cup \left\{\left(\mathbf{d}_{\min} ; \widehat{\mathbb{P}}_{\min}\right)\right\}$
11: $\quad\quad$ **end if**
12: $\quad$ **end for**
13: **end while**
14: **for all** $\mathbf{d}_i \in \mathcal{S}_L$ **do**
15: $\quad \mathcal{DS} \leftarrow \mathcal{DS} \cup \left\{\left(\mathbf{d}_i ; \widehat{\mathbb{P}}[\mathbf{G}(\mathbf{d}_i, \cdot) \le \mathbf{0} \mid \mathcal{S}_\theta]\right)\right\}$
16: **end for**
17: **return** $\mathcal{DS}$

the feasibility probability map more uniformly in order to generate a larger number of points with high reliability value.

Nested Sampling for Design Space Characterization. Nested sampling[32] is a Monte Carlo technique for estimating the evidence in Bayesian parameter estimation. Specifically, given a prior distribution $\pi(\boldsymbol{\theta})$ and a likelihood function $\mathcal{L}(\boldsymbol{\theta})$, the posterior distribution $p(\boldsymbol{\theta})$ can be inferred from Bayes' theorem:

$$p(\boldsymbol{\theta}) = \frac{\mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{Z}$$

where $Z := \mathbb{E}[\mathcal{L}(\boldsymbol{\theta}) \mid \boldsymbol{\theta} \sim \pi(\cdot)] = \int \ldots \int_{\mathbb{R}^{n_\theta}} \mathcal{L}(\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$ denotes the Bayesian evidence. The algorithm proceeds by progressively sampling in nested contours of increasing likelihood, so as to maintain a dense enough sample in regions of higher likelihood—refer to Appendix A of the Supporting Information for further details. This feature is an important motivation for investigating nested sampling in probabilistic DS applications. Before delving into the details of the algorithm, a quick comparison of nested sampling for parameter estimation and DS characterization is presented in Table 1.

A pseudocode of nested sampling for probabilistic DS characterization is presented in Algorithm 2. The inputs to this algorithm comprise (line 1): the knowledge space, $\mathcal{K}$; the desired reliability value α^*; the number of replacement candidates per iteration N_R; an initial sample set of design parameters of size N_L within the knowledge space $\mathcal{S}_L = \{\mathbf{d}_i \in \mathcal{K}: i = 1, ..., N_L\}$; and set of weighted samples of size N_θ describing the model parameter uncertainty $\mathcal{S}_\theta = \{(\boldsymbol{\theta}_j, \omega_j) \sim p(\cdot): j = 1, ..., N_\theta\}$. Recall that the latter may either be drawn from a closed-form distribution (with equal weights), e.g., a Gaussian distribution with a given mean and covariance matrix, or be the result of a Bayesian parameter estimation using MCMC or nested sampling again.

The N_L points in the set $\mathcal{S}_L$ are called live points in the nested sampling literature and may be initialized via uniform sampling in the knowledge space $\mathcal{K}$. Nested sampling starts by estimating the feasibility probability at each of these points according to eq 12, which uses the discretized uncertainty set $\mathcal{S}_\theta$. Then each iteration of the algorithm proceeds by sampling $N_R \ge 1$ new points within an envelope that encloses the current live points and substituting them with the live point having the least feasibility probability in case of improvement. The replaced live points are called dead points and stored with their feasible probability in the result set $\mathcal{DS}$. Notice that the main benefit of using $N_R > 1$ is in regards of vectorizing the feasibility probability evaluation, although this parallelization capability is not used for the case studies in this paper.

Key to the efficiency of nested sampling is the ability to generate replacement candidates that have a satisfactory acceptance rate (line 6). An important motivation behind the adaptation of nested sampling for DS characterization is that a variety of techniques have been developed by the Bayesian estimation community over the years to generate such points. This includes (i) sampling from an enlarged

DOI: 10.1021/acs.iecr.9b05006
Ind. Eng. Chem. Res. 2020, 59, 2396−2408

ellipsoid or multiple enlarged ellipsoids enclosing the current live points;[33,34] or (ii) running a short MCMC from a randomly selected live point.[33,35]

Estimating the feasibility probability of the constraints for given process parameter values requires N_θ process model runs and is by far the most computationally demanding aspect of the algorithm. For N_{iter} nested sampling iterations, the total number of process model runs amounts to $N_\theta(N_L + N_{\text{iter}}N_R)$. An acceleration strategy consists in interrupting the evaluation of the feasibility probability $\hat{\mathbb{P}}[\mathbf{G}(\mathbf{d}_k, \cdot) \leq \mathbf{0} \mid \mathcal{S}_\theta]$ of a replacement candidate $\mathbf{d}_k$ (line 8) if the feasibility probability mass accumulated so far and the mass of the remaining samples add up to less than $\hat{\mathbb{P}}_{\text{min}}$ (line 4)—e.g., by summing in decreasing order of probability weights ω_j in eq 12 for maximal efficiency. Such an interruption is indeed a guarantee that the replacement candidate should be rejected. Of course, the consequence is that the feasibility probability $\hat{\mathbb{P}}[\mathbf{G}(\mathbf{d}_k, \cdot) \leq \mathbf{0} \mid \mathcal{S}_\theta]$ of any rejected point will not be estimated accurately when this acceleration strategy is used. This is the main reason why we do not append the rejected points to the result set $\mathcal{DS}$ in Algorithm 2, which is in agreement with our primary objective of uncovering N_L points with a feasibility probability greater than α^*.

The main iteration terminates when the feasibility probability of all the live points is no less than the target reliability value α^*. All of these points are appended with their corresponding feasibility probabilities to the result set $\mathcal{DS}$ (line 15) that already contains the dead points. It may happen that the number of points with a feasibility probability greater than α^* in $\mathcal{DS}$ is larger than N_L when multiple live point replacements occur during the final iteration. Also, Algorithm 2 will terminate prematurely in the event that $\mathcal{D}_{\alpha^*}$ is empty, which occurs when the model uncertainty is too large or the reliability value α^* is too high. This situation can nonetheless be detected, e.g., by monitoring the rate of improvement in the feasibility probability over a given number of iterations, and the algorithm can be stopped when progress is too slow.

An implementation of Algorithm 2 in a Python package called DEUS (standing for design under uncertainty using sampling techniques) is available as part of the Supporting Information to the paper. At each iteration the replacement candidates are generated by sampling in a single ellipsoid enclosing the current live points. Accordingly, the tuning parameters for a probabilistic DS computation in DEUS include (i) the number of live points, N_L; (ii) the number of replacement candidates at each iteration, N_R; (iii) the initial enlargement factor of the ellipsoid; and (iv) the shrinking rate of that enlargement factor at each iteration. This package is used to solve all of the numerical case studies presented below, using as default parameters an initial enlargement factor of 30% and a shrinking rate of 0.2 for the ellipsoids[33] as well as $N_R = 8$ replacement proposals. Despite the fact that no parallelization is implemented in DEUS at present, choosing an N_R value greater than 1 can reduce the overhead caused by reconstructing an ellipsoid around the live points too frequently. Finally, an appropriate N_L value is highly dependent on the volume ratio between target design space and knowledge space, with a larger volume ratio calling for a larger N_L to maintain a suitable sample density across the DS. In our experience it is advisable to start with a low N_L to establish that $\mathcal{D}_{\alpha^*}$ is not empty, then refine the characterization by increasing N_L. Like any other sampling-based approach, there is no guarantee that DEUS will not miss part of the feasible region. But the likelihood of this happening can be reduced by either increasing the enlargement factor of the ellipsoids, decreasing the related shrinking rate, or increasing the number of replacement proposals per iteration.

Exploitation of the Results. Sampling-based approaches to probabilistic DS characterization return a set of points and their corresponding feasibility probabilities. Turning these results into a format that can be exploited by the practitioner is clearly important. It is straightforward to represent a probabilistic design space in one or two dimensions graphically, for instance using a range of colors to depict different feasibility probability levels. This representation may also be used in three or four dimensions by splitting the range of the extra dimensions onto a trellis chart—see for instance the Suzuki reaction case study below.

But, more generally, a graphical presentation might be insufficient for the practitioner to easily test the feasibility probability of a given set of design parameters or to convey the resulting design space to third-parties in a concise way. Alternative representations include:

- Inscribing a simple shape—e.g., a box or an ellipsoid—within the envelope defined by the samples above a given reliability level in the data set $\mathcal{DS}$ returned by Algorithm 1 or Algorithm 2. Under the assumption that the samples describe a convex set, one may proceed by first constructing a polyhedral representation of the set then fitting the desired shape within that polyhedron. The latter problem may be solved efficiently using convex optimization, but constructing a set of hyperplanes to describe the convex hull of a set of points can be computationally burdensome (NP-hard).[36] This approach is furthermore inadequate for feasibility probability maps presenting nonconvex iso-contours or multiple modes.
- Fitting a nonparametric model to approximate the full feasibility probability map across the knowledge space. For instance, a multilayer perceptron[37,38] (MLP)—a class of feed-forward artificial neural network with multiple (hidden) layers and nonlinear activation functions—can be trained on the labeled data set $\mathcal{DS}$. The trained MLP provides a computationally cheap surrogate, transforming the DS samples into a suitable form for exploitation. It may be used readily to predict the feasibility probability for any design parameters $\mathbf{d} \in \mathcal{K}$. It may also be embedded into a design-centering optimization problem[26] for finding a subset of points, e.g., in the form of a box or an ellipsoid, with feasibility probability above a desired reliability value. This is supported by recent advances in deterministic global optimization with neural networks embedded.[39]
- Fitting a multinomial classifier[37] to separate the knowledge space into two or more subregions corresponding to different feasibility probability ranges—e.g., below and above the reliability value α^*. Similar to nonparametric regression an MLP has the ability to distinguish data sets that are not linearly separable and may be trained on the labeled data set $\mathcal{DS}$. The trained classifier then provides a cheap way of estimating the probability range of any design parameters $\mathbf{d} \in \mathcal{K}$, and a softmax function may be used in the final layer of the MLP to estimate the

DOI: 10.1021/acs.iecr.9b05006
Ind. Eng. Chem. Res. 2020, 59, 2396−2408

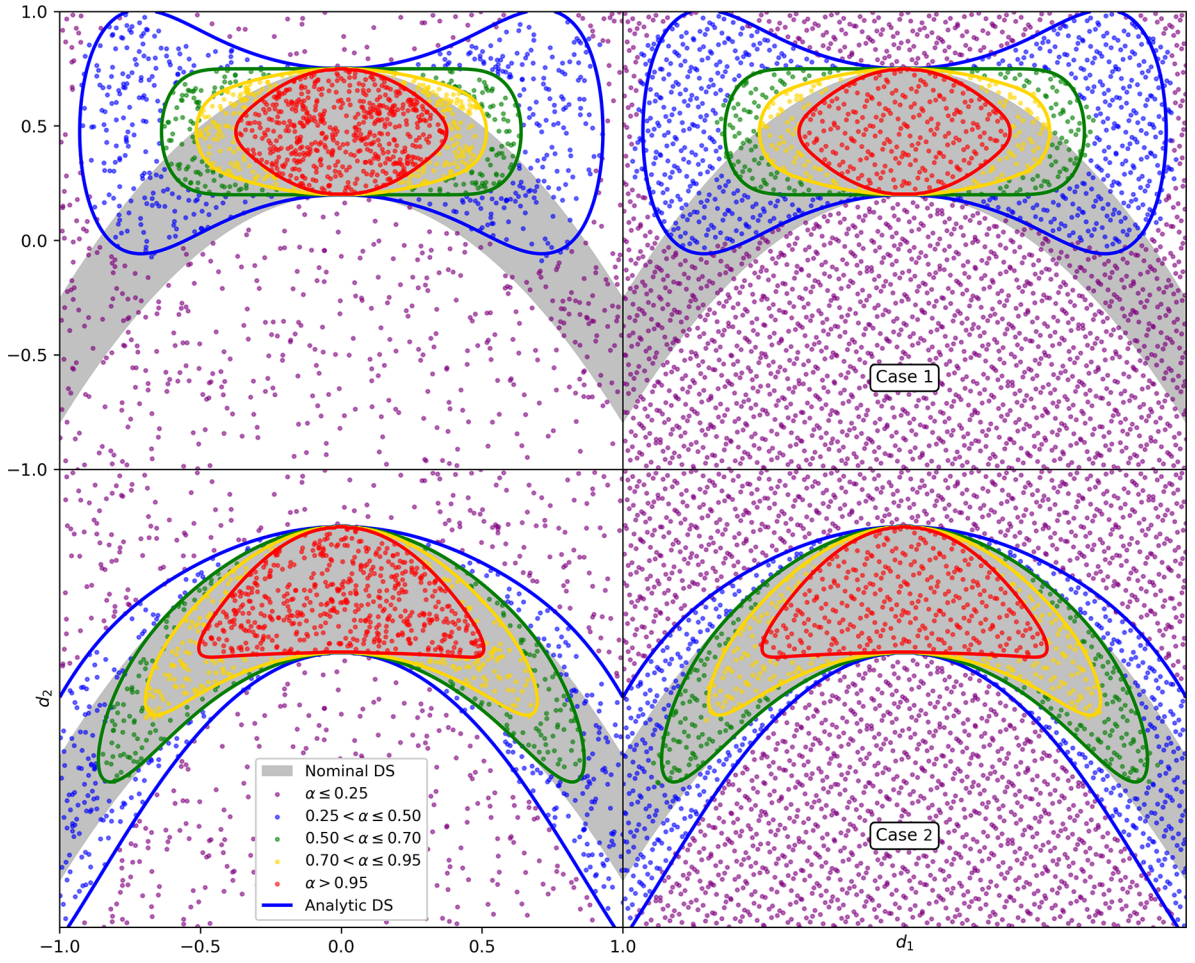


Figure 2. Comparison of sampling-based techniques to determine the probabilistic design space in eq 10. The top and bottom plots are for the uncertainty scenarios $\theta \sim \mathcal{N}(0,1)$ and $\theta \sim \mathcal{N}(1, \sqrt{0.3})$, respectively. The left plots show the results of nested sampling using Algorithm 2. The right plots show the results of the standard Monte Carlo in Algorithm 1 using quasi-random Sobol sampling. The nominal design space (eq 9) is represented with the gray-shaded area. Iso-reliability contours for the actual probabilistic design space at different reliability values α (eq 10) are shown in solid lines with different colors.

probability of $\mathbf{d} \in \mathcal{K}$ belonging to a given range. Like a nonparametric regression model such a classifier may also be embedded into a design-centering problem.

Our focus in the remainder of the paper is on MLP to approximate the full feasibility probability map. The MLP of interest has a single neuron in its output layer—whose state represents the feasibility probability—and its input layer comprises exactly n_d neurons—one for each process parameter. A key, yet arduous, decision is selecting the numbers of hidden layers and hidden neurons in the MLP. At least one hidden layer is necessary since the feasibility probability map is generally nonlinear. An MLP with a single hidden layer is capable of approximating any continuous functions under mild assumptions on the activation function, but it can take an arbitrary large number of neurons in that hidden layer to meet a desired accuracy.[40] Deep learning mitigates this problem by including additional hidden layers. However, it is important to keep in mind that the feasibility probability is inherently noisy due to the model uncertainty discretization. Therefore, the MLP may also be called upon to play a role in filtering this noise and avoiding overfitting of the data by restricting its number of hidden layers or neurons. Since the numerical case studies below have a handful of process parameters we consider hidden layers with a few dozen neurons as a rule of thumb. A more systematic analysis is beyond the scope of this paper.

Illustrative Example (Continued). We revisit our simple case study concerned with the determination of the design space defined by the CQA constraint (eq 8) using the uncertain model (eq 7). The left plots in Figure 2 are the results of nested sampling (Algorithm 2) with a reliability target of $\alpha^* = 0.95$, $N_L = 500$ live points, and $N_\theta = 100$ samples drawn from either $\mathcal{N}(0,1)$ (top plot) or $\mathcal{N}(1, \sqrt{0.3})$ (bottom plot). Overall, the resulting samples—depicted in a different color according to reliability membership—are in excellent agreement with the exact subregions delimited by the solid contours (eq 10). Nevertheless a few of the points are misclassified due to the approximation of the feasibility probability based on eq 12, which uses a finite number of uncertainty samples. This discrepancy can be removed by increasing the number of uncertainty samples N_θ, yet at the cost of a higher computational burden since the total number of model evaluations is directly proportional to N_θ.

The right plots in Figure 2 are generated using the Monte Carlo Algorithm 1 with Sobol sampling. These results are again in excellent agreement with the exact subregions, despite a few samples being misclassified for the same reason as earlier with nested sampling since the same approach is used to estimate

DOI: 10.1021/acs.iecr.9b05006
Ind. Eng. Chem. Res. 2020, 59, 2396−2408

the feasibility probability. For a quantitative comparison between the two algorithms, notice that the total number of model evaluations is identical in both cases (322 000 and 369 200 model evaluations under the uncertainty scenarios $\theta \sim \mathcal{N}(0,1)$ and $\theta \sim \mathcal{N}(1, \sqrt{0.3})$, respectively) and no acceleration strategy is applied for nested sampling. A breakdown of the samples generated with Algorithm 1 and Algorithm 2 within different feasibility probability ranges is furthermore reported in Table 2. Recall that unsuccessful proposals in the nested sampling approach are currently discarded, which is the reason why the total number of samples is lower with this approach compared to Monte Carlo sampling. Regardless, nested sampling enables a much denser sampling of the targeted reliability region, as controlled by the number of live points and the gradual sampling toward higher reliability regions. It would take three to four times more samples—and hence model evaluations—for the standard Monte Carlo approach to achieve a similar concentration of points within the desired reliability region of 0.95 in this case study.

Table 2. Comparison between the Samples Generated via Standard Monte Carlo with Sobol Sampling (Algorithm 1) and Nested Sampling (Algorithm 2) within Different Reliability Ranges

reliability value	samples drawn by standard Monte Carlo (Algorithm 1)	samples drawn by nested sampling (Algorithm 2)
uncertainty scenario: $\theta \sim \mathcal{N}(0,1)$		
$0.95 \leq \alpha$	247	500
$0.70 \leq \alpha < 0.95$	139	231
$0.50 \leq \alpha < 0.70$	172	181
$0.25 \leq \alpha < 0.50$	666	418
$\alpha < 0.25$	2025	532
total	**3249**	**1862**
uncertainty scenario: $\theta \sim \mathcal{N}(1, \sqrt{0.3})$		
$0.95 \leq \alpha$	331	500
$0.70 \leq \alpha < 0.95$	194	210
$0.50 \leq \alpha < 0.70$	212	200
$0.25 \leq \alpha < 0.50$	490	321
$\alpha < 0.25$	2022	572
total	**3249**	**1803**

The left plot in Figure 3 compares several iso-contours of the exact feasibility probability map (eq 9) with those of an MLP trained on the nested sampling results under the uncertainty scenario $\theta \sim \mathcal{N}(0,1)$. A good agreement is obtained in the upper reliability region where the concentration of samples is high, while the MLP predictions somewhat deviate in the lower reliability region where the sample points are scarcer—compare Table 2. The MLP under consideration here has four hidden layers, with 16, 32, 32, and 16 artificial neurons, respectively. Its training was conducted in the Python package Keras[41] that is implemented on top of the package TensorFlow,[42] using mini-batch gradient descent. The resulting parity plot is shown on the right plot in Figure 3. The regression quality is generally good but for a few of the points that present a rather large deviation from the training set. Part of this mismatch is attributed to the fact that the data points in the training set itself are noisy due to the uncertainty discretization. Moreover, the actual feasibility probability map presents discontinuities at the points $\mathbf{d} = (0,0.20)^T$ and $\mathbf{d} = (0,0.75)^T$ so the MLP has difficulties capturing the very stiff probability variations near such points.

Industrial Case Studies. We demonstrate the applicability of the proposed nested sampling approach on two case studies of industrial relevance. The first one considers a Michael addition reaction in a continuous reactor and enables a comparison with recent work on flexibility-based algorithms[21] for probabilistic DS characterization. The second one investigates a biphasic Suzuki coupling reaction performed in fully batch mode[10] and gives the opportunity to determine a four-dimensional DS with a more complex mechanistic process model.

Michael Addition Reaction. This case study considers the following Michael addition reaction in a continuous stirred-tank reactor (CSTR):[21]

$$\mathrm{AH} + \mathrm{B} \xrightarrow{k_1} \mathrm{A}^- + \mathrm{BH}^+$$

$$\mathrm{A}^- + \mathrm{C} \underset{k_3}{\overset{k_2}{\rightleftharpoons}} \mathrm{AC}^-$$

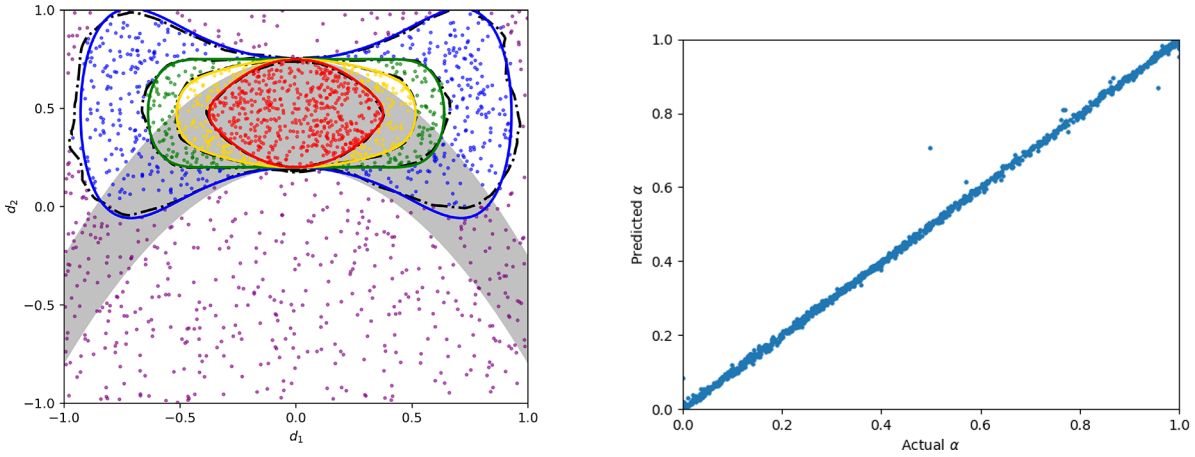


Figure 3. Exploitation of the nested sampling results using an MLP to fit the full feasibility probability map in the uncertainty scenario $\theta \sim \mathcal{N}(0,1)$. The left plot compares the contours of the fitted MLP (dashed lines) with the actual DS (solid color lines) at different reliability values α. The gray-shaded area depicts the nominal design space (eq 9). The right plot shows the parity plot of the trained MLP.

DOI: 10.1021/acs.iecr.9b05006
Ind. Eng. Chem. Res. 2020, 59, 2396−2408

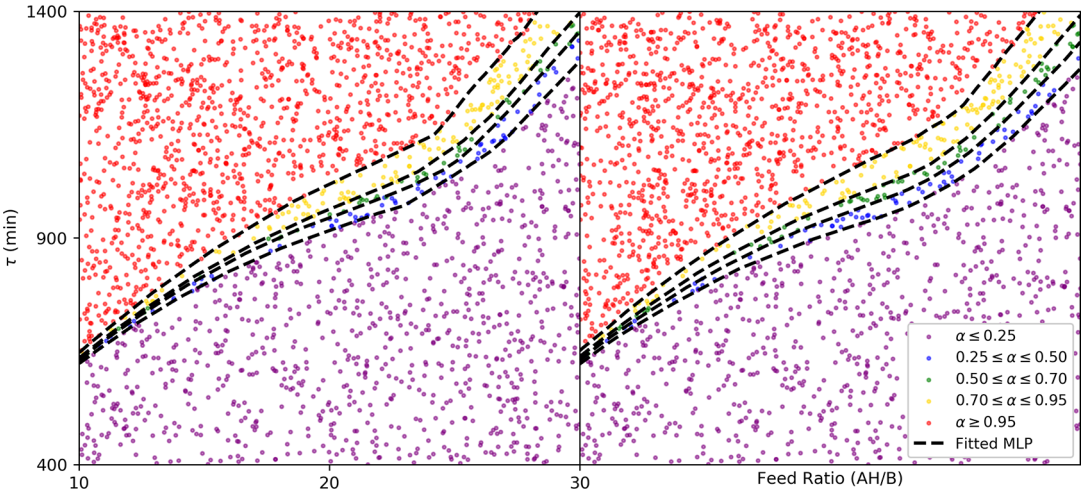


Figure 4. Comparison of probabilistic DS for the Michael addition reaction computed using nested sampling (Algorithm 2) with N_θ = 100 (left) and 1000 (right) uncertainty scenarios. Samples from the probabilistic design space belonging to different reliability ranges are shown in different colors. Iso-reliability contours from the fitted MLP are shown in black dashed lines.

$$AC^- + AH \xrightarrow{k_4} A^- + P$$

$$AC^- + BH^+ \xrightarrow{k_5} P + B$$

where AH is the Michael donor; C is the Michael acceptor; B is a base; BH^+, A^-, and AC^- are reaction intermediates; and P is the product. A complete statement of the steady-state process model can be found in Appendix B of the Supporting Information. All of the kinetic constants are considered uncertain in this model, following a multivariate normal distribution $(k_1, ..., k_5) \sim \mathcal{N}(\mu_k, \Sigma_k)$.

The knowledge space is defined in terms of two process parameters: (i) the molar ratio between the concentration of AH and B in the feed, $R_{AH|B} \in [10, 30]$; and (ii) the residence time in the CSTR, $\tau \in [400, 1400]$ (min). The feasible operating region is furthermore limited by two CQA constraints: (i) conversion of feed C greater than 90%; and (ii) residual concentration of AC^- smaller than 2 mmol L^{-1}.

Like in the illustrative example above, we apply nested sampling (Algorithm 2) with a reliability target of $\alpha^* = 0.95$, now considering $N_L = 1000$ live points to enable a high density of points. The resulting sample set ($\mathcal{DS}$) is then regressed using an MLP comprising 4 hidden layers (with 16, 32, 32, and 16 neurons in each layer) to approximate the full feasibility probability map. The two plots in Figure 4 are for N_θ = 100 and 1000 uncertainty samples. The corresponding computational statistics are reported in Table 3, both without and with the acceleration strategy. The effect of increasing the number of uncertainty scenarios by 10-fold is rather small in Figure 4, while the increase in CPU time is naturally close to a factor of 10 due to the corresponding increase in the number of function evaluations. Though differences can be noted between both fitted MLPs, with the MLP on the right plot (N_θ = 1000) predicting smoother iso-reliability contours due to a lower noise in the estimated feasibility probabilities compared to the MLP on the left plot (N_θ = 100). It is also worth noting that the effect of the acceleration—about 5% reduction in the number of function evaluations and 10% reduction in CPU time—is rather modest. This is attributed to a relatively low fraction of rejected proposals due to the shape of the DS.

Table 3. Computational Statistics for the Michael Addition Reaction Using Nested Sampling (Algorithm 2) with Different Model Uncertainty Representations

	nested sampling with acceleration		nested sampling without acceleration	
uncertainty scenarios, N_θ	# model eval.	CPU sec.[a]	# model eval.	CPU sec.[a]
100	226 884	114	236 000	128
1000	2 287 174	1070	2 400 000	1207

[a]CPU times obtained on a single core of AMD Ryzen 5 2600X processor.

Finally, we compare the nested sampling results to the results obtained by Laky et al.[21] using an optimization approach based on flexibility analysis. It is worth reiterating that the algorithms therefrom determine or approximate the set-membership counterpart $\hat{\mathcal{D}}_\alpha$ (eq 6) to the probabilistic design space $\mathcal{D}_\alpha$, which happens to be quite conservative.[21] These optimization-based methods are consistently faster than nested sampling (by a factor of 10 or more) when solving the optimization problems using a (local) gradient-based algorithm. But nested sampling is nevertheless competitive against these methods when a global optimization solver is used.

Suzuki Coupling Reaction. This case study investigates the Suzuki coupling reaction between a boronic ester (SM1) and an organohalide (SM2) to produce a desired pharmaceutical intermediate (P1) and a dimeric impurity (Imp1) related to SM1.[10] The reaction is biphasic and conducted in batch mode. The gaseous phase consists of an inert gas with traces of O_2; the liquid phase consists of a mixture of water and tetrahydrofuran (THF) as solvent for 17 chemical species that participate in 12 reactions—3 of which are reversible and 1 is considered instantaneous. A mechanistic, kinetic-based model is available describing the changes in composition of the liquid and gas phases during the batch—see Appendix C of the Supporting Information for a complete statement. The temperature-dependent reaction rates are modeled with Arrhenius equations and kinetic parameters that were verified experimentally.[10] The pre-exponential factors of all 14 kinetically limited reactions are considered to be uncertain

DOI: 10.1021/acs.iecr.9b05006
Ind. Eng. Chem. Res. 2020, 59, 2396−2408

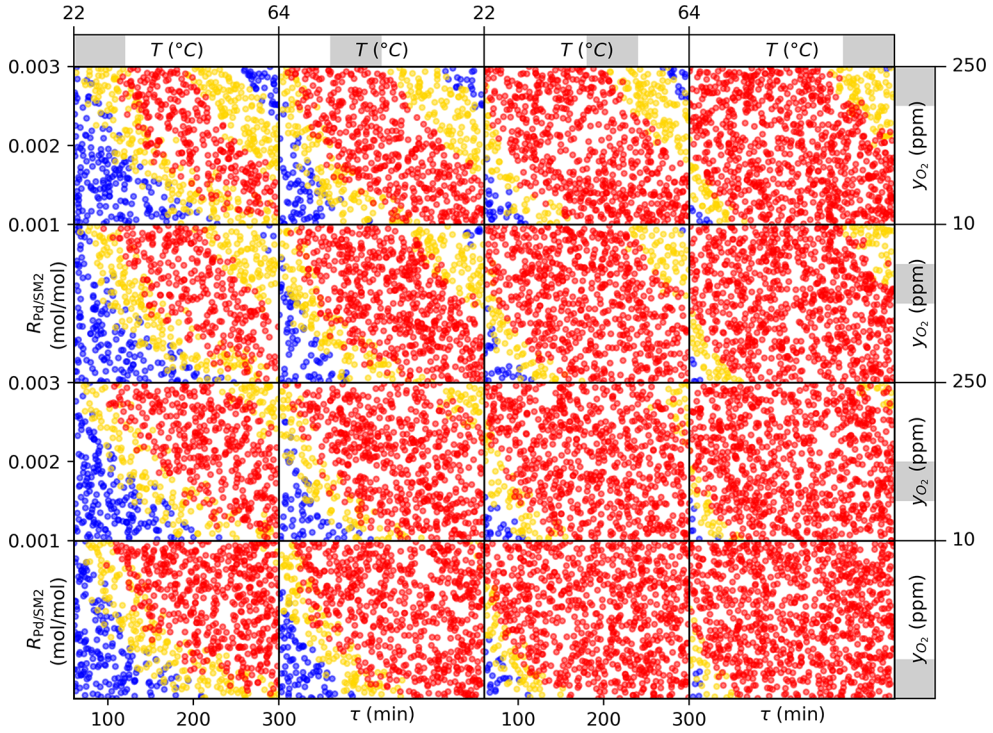


Figure 5. Probabilistic DS for the Suzuki coupling reaction computed using nested sampling (Algorithm 2) with N_L = 10 000 live points and N_θ = 1000 uncertainty scenarios. The outer axes on the trellis chart correspond to the oxygen fraction in head space (y_{O_2}) and reactor temperature (T); the inner axes on each subplot, to the batch duration (τ) and catalyst equivalent ($R_{Pd|SM2}$). Samples from the probabilistic design space belonging to different reliability ranges are shown in different colors (red: $\alpha \geq 0.85$; yellow: $0.05 \leq \alpha < 0.85$; blue: $\alpha < 0.05$).

here, following normal distributions with standard-deviations equal to 15% of the nominal values.

The reactor has a large number of design and operation parameters so we consider a reduced DS characterization problem in terms of four parameters only: (i) the batch duration, $\tau \in [75, 300]$ (min); (ii) the equivalent of catalyst, $R_{Pd|SM2} \in [0.001, 0.003]$(mol mol^{-1}); (iii) the reactor temperature, $T \in [22, 64]$ (°C); and (iv) the molar fraction of O_2 in reactor's head space $y_{O_2} \in [10, 250]$(ppm). Moreover, two CQAs limit the feasible operating region at the end of the batch: (i) maximum amount of unreacted SM2 of 0.001 mol mol^{-1} for the reaction to be considered complete; and (ii) maximum level of impurity Imp1 below 0.0015 mol mol^{-1} for the batch product to be downstream processable.

In order to determine the probabilistic DS we apply nested sampling (Algorithm 2) with a reliability target of $\alpha^* = 0.85$. Because the DS now comprises four process parameters the numbers of live points needs increasing to ensure a high density of points, so we use N_L = 5000 and 10 000 next. We furthermore consider two uncertainty descriptions with N_θ = 200 and 1000 samples, respectively. For the DS representation we use trellis charts where the ranges of oxygen concentration and temperature are split into four-by-four intervals—indicated by gray bars on the two outer axes; then each subplot is a projection of the points that belong to the particular intervals of oxygen fraction in head space y_{O_2} and temperature T on the plane defined by the batch duration τ and catalyst equivalent $R_{Pd|SM2}$—the two inner axes. For instance, the chart on Figure 5 is a representation of the probabilistic DS computed with 10 000 live points and 1000 uncertainty scenarios. Notice in particular the effect of the process parameters on the chart, whereby the DS expands upon increasing temperature at constant oxygen level and upon increasing oxygen fraction in head space at constant temperature. Computational statistics for the probabilistic DS are shown in Table 4. Overall, it takes over 4 days to determine the probabilistic DS with 10 000 live points and 1000 uncertainty scenarios. A 12-fold decrease in the CPU time (to about 9 h) is observed upon reducing to 5000 live points and 200 uncertainty scenarios. The number of function evaluations is only reduced by about 3% in applying the acceleration strategy since the fraction of rejected proposals is again low. The corresponding reduction in CPU time is between 4−10%. A vectorized implementation of Algorithm 2 that would compute simulation ensembles or replacement proposals in parallel

Table 4. Computational Statistics for the Suzuki Coupling Reaction Using Nested Sampling (Algorithm 2) with Different Model Uncertainty Representations and Live Points

		nested sampling with acceleration		nested sampling without acceleration	
live points, N_L	uncertainty scenarios, N_θ	# model eval.	CPU min.[a]	# model eval.	CPU min.[a]
5000	200	1 412 390	525	1 451 200	574
10 000	1000	14 132 975	6464	14 544 000	6741
5000	1	–	–	6392	3.5
10 000	1	–	–	12 680	6.8

[a]CPU times obtained on a single core of AMD Ryzen 5 2600X processor.

DOI: 10.1021/acs.iecr.9b05006
Ind. Eng. Chem. Res. 2020, 59, 2396−2408

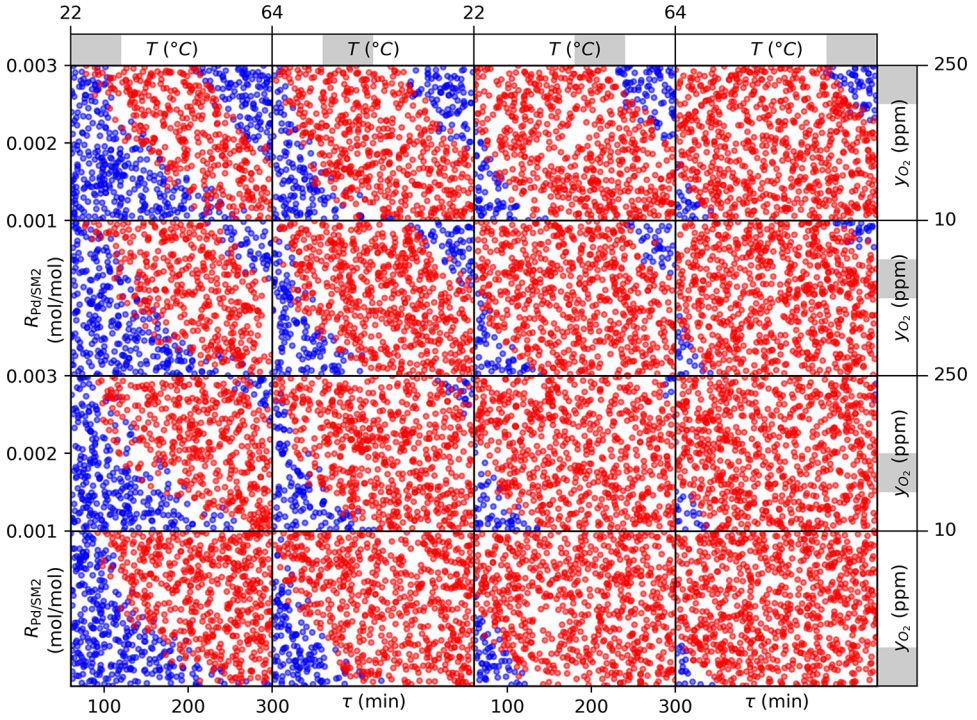


Figure 6. Nominal DS for the Suzuki coupling reaction computed using nested sampling (Algorithm 2) with N_L = 10 000 live points. The outer axes on the trellis chart correspond to the oxygen fraction in head space (y_{O_2}) and reactor temperature (T); the inner axes on each subplot, to the batch duration (τ) and catalyst equivalent ($R_{Pd|SM2}$). Samples within the nominal design space are depicted in red and those outside are in blue.

could readily reduce the overall time needed and will be the focus of future work.

For comparison, we also compute the nominal DS (eq 3) by applying nested sampling (Algorithm 2), in this case with a single uncertainty scenario (N_θ = 1) corresponding to the nominal kinetic parameters and setting the reliability target to α^* = 1. The chart in Figure 6 is a representation of this nominal DS with 10 000 live points. The results in Table 4 for 5000 and 10 000 live points show a dramatic reduction in CPU time to just a few minutes, which is clearly attributed to the fact that a single simulation is needed to assess whether or not a point belongs to the nominal DS instead of a large ensemble of simulations. However, the results in Figures 5 and 6 differ significantly as certain points within the nominal DS have a feasibility probability as low as 10−20%—an extremely poor reliability level in practice. Despite their popularity among practitioners, approaches based on nominal parameter values or mean responses fail to quantify reliability and risk, and therefore, their results can be misleading.[10,22]

CONCLUDING REMARKS

Design space is a key concept in pharmaceutical QbD, helping practitioners develop a better understanding of their manufacturing processes and enhancing regulatory flexibility. In this context, it is of paramount importance to develop computational methods and tools that can provide a clear quantitative representation of the DS—in agreement with the ICH Q8 guideline. Our main focus throughout the paper has been on Bayesian approaches to DS characterization, which determine a feasibility probability that can be used as a measure of reliability and risk.

We have contributed a nested sampling algorithm tailored to the characterization of a probabilistic DS. A hallmark feature of nested sampling is its ability to maintain a given set of live points through regions with increasing probability feasibility until reaching a desired reliability level (or showing it cannot be reached). The algorithm furthermore leverages efficient strategies from Bayesian parameter estimation for generating replacement proposals during the search. Through a simple illustrative example and the case study of a Michael addition reaction, we have established that nested sampling can outperform standard Monte Carlo sampling and be competitive with optimization methods relying on process flexibility concepts, even in low-dimensional DS characterization problems. We have also showcased the use of machine learning techniques to reconstruct a feasibility probability map based on the sampled design space, which can be more easily exploited by the practitioner. In the second case study of a Suzuki coupling reaction we have shown that nested sampling is effective for larger DS characterization with a handful of process parameters, in the presence of a complex dynamic model and realistic model uncertainty—a class of problems currently out of reach for flexibility-based optimization techniques.

A major impediment facing nested sampling—and other sampling-based techniques—for probabilistic DS characterization in higher-dimensional problems, for instance, in multiunit integrated plants, is the very large number of process simulations required. This stems from the large ensemble of process simulations needed to capture the effect of model uncertainty in conjunction with maintaining a large number of live points. We have implemented an acceleration strategy as part of nested sampling, which reduces the overall number of process model runs without impairing a design space's accuracy. A recommended follow-up to this work would entail a vectorized implementation of the nested sampling algorithm in order to further reduce the time needed to characterize a

DOI: 10.1021/acs.iecr.9b05006
Ind. Eng. Chem. Res. 2020, 59, 2396−2408

complex DS, for instance by computing simulation ensembles or replacement proposals in parallel.

Lastly, the characterization of the probabilistic DS in our work did not include adjustable control actions, although these may increase the size of the probabilistic design space. This practice is acceptable insofar as many pharmaceutical processes still use minimal online measurement and control of the CQAs—the process is instead carried to completion, followed by testing of the final product. As part of future work, it would be interesting to extend the nested sampling approach in order to encompass such recourse actions.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.iecr.9b05006.

Appendix A: Nested Sampling for Bayesian Evidence; Appendix B: Michael Addition Model; Appendix C: Suzuki Coupling Model; Appendix D: DEUS Python Package (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: b.chachuat@imperial.ac.uk.

ORCID

Radoslav Paulen: 0000-0002-1599-2634

Benoît Chachuat: 0000-0003-4780-9686

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work is supported by Eli Lilly and Company through the Pharmaceutical Systems Engineering Lab (PharmaSEL) program. Lucian Gomoescu is a Marie Skłodowska-Curie early stage researcher at Process Systems Enterprise, Ltd., enrolled in the European Union's Horizon 2020 research and innovation program under grant agreement 675585 (Marie Skłodowska-Curie ITN SyMBioSys). Radoslav Paulen gratefully acknowledges the contribution of the European Commission under grant 790017 (GuEst), the contribution of the Slovak Research and Development Agency (project APVV 15-0007), and the support of Slovak Ministry of Education, Science, Research and Sport under the project STU as the Leader of Digital Coalition 002STU-2-1/2018. The authors are grateful to Carla Vanesa Luciani for her assistance with the Suzuki reaction case study.

■ NOMENCLATURE

Acronyms

CFD = computational fluid dynamics
CPU = central processing unit
CQA = critical quality attribute
CSTR = continuous stirred-tank reactor
DS = design space
HPD = highest posterior density
ICH = International Council for Harmonization
MCMC = Markov chain Monte Carlo
MLP = multilayer perceptron
PSE = process systems engineering
QbD = quality by design
QTPP = quality target product profile
DEUS = design under uncertainty using sampling techniques

Main Symbols

α = reliability value
$\mathcal{D}$ = design space
$\mathcal{K}$ = knowledge space
$\mathcal{DS}$ = sampled design space
$\mathcal{L}(\cdot)$ = likelihood function
$\mathcal{N}(\mu, \sigma)$ = Normal distribution with mean μ and standard deviation σ
$\mathcal{S}$ = sampling set
$\mathbb{I}[\cdot]$ = indicator function
$\mathbb{E}[\cdot]$ = expected value
$\mathbb{P}[\cdot]$ = probability
ω = weight
$\pi(\cdot)$ = prior distribution
θ = model parameter
τ = residence time [min]
d = process parameter
G = critical quality attribute constraint
k = kinetic rate constant
N = number of samples
n = dimension
$p(\cdot)$ = posterior distribution
R = molar ratio [mol mol^{-1}]
s = critical quality attribute
T = temperature [°C]
Y = gas molar fraction [ppm]
Z = evidence

Subscripts and Superscripts

* = target value
iter = iterations
L = live points
min = minimal value
nom = nominal
R = replacement proposals

■ REFERENCES

(1) Troup, G. M.; Georgakis, C. Process systems engineering tools in the pharmaceutical industry. *Comput. Chem. Eng.* **2013**, *51*, 157−171.

(2) Reklaitis, G. V., Seymour, C., García-Muñoz, S., Eds.; *Comprehensive Quality by Design for Pharmaceutical Product Development and Manufacture*; John Wiley & Sons, 2017.

(3) DiMasi, J. A.; Grabowski, H. G.; Hansen, R. W. Innovation in the pharmaceutical industry: New estimates of R&D costs. *Journal of Health Economics* **2016**, *47*, 20−33.

(4) ICH *Quality Guidelines*. https://www.ich.org/products/guidelines/quality/article/quality-guidelines.html, 2019 (accessed 15 August 2019).

(5) Holm, P.; Allesø, M.; Bryder, M. C.; Holm, R. In *ICH Quality Guidelines*; Teasdale, A., Elder, D., Nims, R. W., Eds.; John Wiley & Sons, 2017; Chapter 20, pp 535−577.

(6) Boukouvala, F.; Muzzio, F. J.; Ierapetritou, M. G. Design space of pharmaceutical processes using data-driven-based methods. *Journal of Pharmaceutical Innovation* **2010**, *5*, 119−137.

(7) Chatterjee, S.; Moore, C. M. V.; Nasr, M. M. In *Comprehensive Quality by Design for Pharmaceutical Product Development and Manufacture*; Reklaitis, G. V., Seymour, C., García-Muñoz, S., Eds.; John Wiley & Sons, 2017; Chapter 2, pp 9−24.

(8) Facco, P.; Dal Pastro, F.; Meneghetti, N.; Bezzo, F.; Barolo, M. Bracketing the design space within the knowledge space in pharmaceutical product development. *Ind. Eng. Chem. Res.* **2015**, *54*, 5128−5138.

DOI: 10.1021/acs.iecr.9b05006
Ind. Eng. Chem. Res. 2020, 59, 2396−2408

(9) Bano, G.; Facco, P.; Bezzo, F.; Barolo, M. Probabilistic Design space determination in pharmaceutical product development: A Bayesian/latent variable approach. *AIChE J.* **2018**, *64*, 2438−2449.
(10) García-Muñoz, S.; Luciani, C. V.; Vaidyaraman, S.; Seibert, K. D. Definition of design spaces using mechanistic models and geometric projections of probability maps. *Org. Process Res. Dev.* **2015**, *19*, 1012−1023.
(11) Biegler, L. T.; Grossmann, I. E.; Westerberg, A. W. *Systematic Methods of Chemical Process Design*; Prentice Hall, 1997.
(12) Grossmann, I. E.; Calfa, B. A.; Garcia-Herreros, P. Evolution of concepts and models for quantifying resiliency and flexibility of chemical processes. *Comput. Chem. Eng.* **2014**, *70*, 22−34.
(13) Halemane, K. P.; Grossmann, I. E. Optimal Process Design Under Uncertainty. *AIChE J.* **1983**, *29*, 425−433.
(14) Swaney, R. E.; Grossmann, I. E. An index for operational flexibility in chemical process design. Part I: Formulation and theory. *AIChE J.* **1985**, *31*, 621−630.
(15) Georgakis, C.; Uztürk, D.; Subramanian, S.; Vinson, D. R. On the operability of continuous processes. *Control Engineering Practice* **2003**, *11*, 859−869.
(16) Lima, F. V.; Jia, Z.; Ierapetritou, M.; Georgakis, C. Similarities and differences between the concepts of operability and flexibility: The steady-state case. *AIChE J.* **2009**, *56*, 702−716.
(17) Pantelides, C. C.; Shah, N.; Adjiman, C. S. Design Space, Models and Model Uncertainty. AIChE Annual Meeting: Nashville, TN, 2009.
(18) Straub, D. A.; Grossmann, I. E. Integrated stochastic metric of flexibility for systems with discrete state and continuous parameter uncertainties. *Comput. Chem. Eng.* **1990**, *14*, 967−985.
(19) Pistikopoulos, E. N.; Mazzuchi, T. A. A novel flexibility analysis approach for processes with stochastic parameters. *Comput. Chem. Eng.* **1990**, *14*, 991−1000.
(20) Peterson, J. J. A Bayesian approach to the ICH Q8 definition of design space. *Journal of Biopharmaceutical Statistics* **2008**, *18*, 959−975.
(21) Laky, D.; Xu, S.; Rodriguez, J. S.; Vaidyaraman, S.; García Muñoz, S.; Laird, C. An Optimization-Based Framework to Define the Probabilistic Design Space of Pharmaceutical Processes with Model Uncertainty. *Processes* **2019**, *7*, 96.
(22) Peterson, J. J.; Yahyah, M.; Lief, K.; Hodnett, N. In *Comprehensive Quality by Design for Pharmaceutical Product Development and Manufacture*; Reklaitis, G. V., Seymour, C., García-Muñoz, S., Eds.; John Wiley & Sons, 2017; Chapter 4, pp 55−70.
(23) Straub, D. A.; Grossmann, I. E. Design optimization of stochastic flexibility. *Comput. Chem. Eng.* **1993**, *17*, 339−354.
(24) Ochoa, M. P.; Luciani, C.; Stamatis, S. D.; García-Muñoz, S.; Grossmann, I. E. *New Developments in Flexibility Analysis in the Framework of Design Space Definition*; AIChE Annual Meeting: Pittsburgh, PA, 2018.
(25) Floudas, C. A.; Gümüş, Z. H. Global Optimization in Design under Uncertainty: Feasibility Test and Flexibility Index Problems. *Ind. Eng. Chem. Res.* **2001**, *40*, 4267−4282.
(26) Harwood, S. M.; Barton, P. I. How to solve a design centering problem. *Mathematical Methods of Operations Research* **2017**, *86*, 215−254.
(27) Boukouvala, F.; Muzzio, F. J.; Ierapetritou, M. G. In *Comprehensive Quality by Design for Pharmaceutical Product Development and Manufacture*; Reklaitis, G. V., Seymour, C., García-Muñoz, S., Eds.; John Wiley & Sons, 2017; Chapter 6, pp 95−123.
(28) Wang, Z.; Ierapetritou, M. G. Global sensitivity, feasibility, and flexibility analysis of continuous pharmaceutical manufacturing processes. *Comput.-Aided Chem. Eng.* **2018**, *41*, 189−213.
(29) Boukouvala, F.; Ierapetritou, M. G. Feasibility analysis of black-box processes using an adaptive sampling Kriging-based method. *Comput. Chem. Eng.* **2012**, *36*, 358−368.
(30) Rogers, A.; Ierapetritou, M. Feasibility and flexibility analysis of black-box processes. Part 1: Surrogate-based feasibility analysis. *Chem. Eng. Sci.* **2015**, *137*, 986−1004.
(31) Rogers, A.; Ierapetritou, M. Feasibility and flexibility analysis of black-box processes. Part 2: Surrogate-based flexibility analysis. *Chem. Eng. Sci.* **2015**, *137*, 1005−1013.
(32) Skilling, J. Nested Sampling. *AIP Conf. Proc.* **2004**, *735*, 395−405.
(33) Feroz, F.; Hobson, M. P.; Bridges, M. MultiNest: an efficient and robust Bayesian inference tool for cosmology and particle physics. *Mon. Not. R. Astron. Soc.* **2009**, *398*, 1601−1614.
(34) Mukherjee, P.; Parkinson, D.; Liddle, A. R. A nested sampling algorithm for cosmological model selection. *Astrophys. J.* **2006**, *638*, L51.
(35) Handley, W.; Hobson, M.; Lasenby, A. PolyChord: nested sampling for cosmology. *Mon. Not. R. Astron. Soc.: Lett.* **2015**, *450*, L61−L65.
(36) Boyd, S.; Vandenberghe, L. *Convex Optimization*; Cambridge University Press: New York, NY, 2004.
(37) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed.; Springer: New York, NY, 2009.
(38) Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, 2016.
(39) Schweidtmann, A. M.; Mitsos, A. Deterministic Global Optimization with Artificial Neural Networks Embedded. *Journal of Optimization Theory & Applications* **2019**, *180*, 925−948.
(40) Hornik, K. Approximation capabilities of multilayer feedforward networks. *Neural Networks* **1991**, *4*, 251−257.
(41) Chollet, F.; et al. *Keras*, 2015; https://keras.io.
(42) Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G. S.; Davis, A.; Dean, J.; Devin, M.; Ghemawat, S.; Goodfellow, I.; Harp, A.; Irving, G.; Isard, M.; Jia, Y.; Jozefowicz, R.; Kaiser, L.; Kudlur, M.; Levenberg, J.; Mané, D.; Monga, R.; Moore, S.; Murray, D.; Olah, C.; Schuster, M.; Shlens, J.; Steiner, B.; Sutskever, I.; Talwar, K.; Tucker, P.; Vanhoucke, V.; Vasudevan, V.; Viégas, F.; Vinyals, O.; Warden, P.; Wattenberg, M.; Wicke, M.; Yu, Y.; Zheng, X. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*, 2015; http://tensorflow.org/.

DOI: 10.1021/acs.iecr.9b05006
Ind. Eng. Chem. Res. 2020, 59, 2396−2408

Computers and Chemical Engineering 153 (2021) 107437

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Data-based design of inferential sensors for petrochemical industry

Martin Mojto[a,*], Karol Ľubušký[b], Miroslav Fikar[a], Radoslav Paulen[a]

[a] Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Radlinského 9, Bratislava 812 37, Slovakia
[b] Slovnaft, a.s., Vlčie hrdlo 1, Bratislava 824 12, Slovakia

ARTICLE INFO

Article history:
Received 18 January 2021
Revised 19 June 2021
Accepted 1 July 2021
Available online 14 July 2021

Keywords:
Inferential (soft) sensors
Data pre-treatment
Petrochemical industry
Process monitoring

ABSTRACT

Inferential (or soft) sensors are used in industry to infer the values of imprecisely and rarely measured (or completely unmeasured) variables from variables measured online (e.g., pressures, temperatures). The main challenge, akin to classical model overfitting, in designing an effective inferential sensor is the selection of a correct structure of the sensor. The sensor structure is represented by the number of inputs to the sensor, which correspond to the variables measured online and their (simple) combinations. This work is focused on the design of inferential sensors for product composition of an industrial distillation column in two oil refinery units, a Fluid Catalytic Cracking unit and a Vacuum Gasoil Hydrogenation unit. As the first design step, we use several well-known data pre-treatment (gross error detection) methods and compare the ability of these approaches to indicate systematic errors and outliers in the available industrial data. We then study effectiveness of various methods for design of the inferential sensors taking into account the complexity and accuracy of the resulting model. The effectiveness analysis indicates that the improvements achieved over the current inferential sensors are up to 19%.

1. Introduction

The accuracy and reliability of industrial measurements have a huge impact on the effectiveness of industrial process control (Khatibisepehr et al., 2013). Especially, the control performance of advanced process controllers (Qin and Badgwell, 2003) is highly related to the indication quality of controlled variables (CVs). It is often the case that the crucial CVs (e.g., distillate purity) are too expensive or impossible to measure at the frequency required for an effective feedback control. This gave rise to a use of so-called inferential (or soft) sensors (Mejdell and Skogestad, 1991; Kordon et al., 2003; Curreri et al., 2020).

The purpose of an inferential sensor is to infer the CV value (output) using the data from other measured variables (inputs). The design procedure aims at (a) identifying the sensor structure and (b) at estimating the sensor parameters. While the latter problem can be solved relatively easily and there are standard academic and industrial tools even for situations of parameter-varying sensors (e.g., recursive estimation or simple bias update King, 2011), the former issue of structure selection can be much more challenging in practice.

The effectiveness and reliability of inferential sensors are highly related to the quality of data used for the design. Subsequently, the data quality is affected by the amount of systematic and random errors (Su et al., 2009). There are several methods dealing with the pre-treatment methods for industrial data (Alves and Nascimento, 2007). The group of well-known and popular multivariate data treatment methods includes Hotelling's T^2 distance (Hotelling, 1931), k-means clustering (Forgy, 1965), or minimum covariance determinant (MCD) technique (Rousseeuw, 1984). Several applications of these methods were reported in the industrial context (Alameddine et al., 2010; Xu et al., 2017; Frumosu and Kulahci, 2019; Azzaoui et al., 2019; Fontes et al., 2021).

There are several approaches to the design of inferential sensors (Fortuna et al., 2007; Liu, 2010; Sun and Braatz, 2021). According to the way of inferential sensor modeling, one can divide these methods into two main types: model-based and data-driven. The former category usually uses a first-principles model (Torgashov and Skogestad, 2019) and therefore it requires fundamental knowledge about the process behavior and characteristics. Several contributions have been published in the field of model-based approaches and their combination with the extended Kalman filter (Gryzlov et al., 2013) or with the neural networks (Chen et al., 2000). However, the behavior of industrial process is often too complicated and requires much effort to develop a first-principles model with an acceptable accuracy. In such a case, data-driven approaches provide less demanding yet effective solution. The popularity of the data-driven methods also increases in the process industry because of the increased availability of modern and cheap online sensors.

* Corresponding author.
E-mail address: martin.mojto@stuba.sk (M. Mojto).

https://doi.org/10.1016/j.compchemeng.2021.107437
0098-1354/

Nomenclature

List of acronyms

AIC_C	corrected Akaike information criterion
BC	bias correction
BIC	Bayesian information criterion
CV	controlled variable
FCC	fluid catalystic cracking
GF	gasoline fraction
HGO	hydrogenated gasoil
LASSO	least absolute shrinkage and selection operator
MCD	minimum covariance determinant
MIQP	mixed-integer quadratic programming
NIPALS	nonlinear iterative partial least squares
OLSR	ordinary least squares regression
PCA	principal component analysis
PLS	partial least squares
Ref	current (reference) inferential sensor
RMSE	root mean square error
RSS	residual sum of squares
SS	subset selection
SS-CV	subset selection with cross-validation
SVD	singular value decomposition
VGH	vacuum gasoil hydrogenation

List of symbols

a	vector of inferential sensor parameters, $a \in \mathbb{R}^{n_p}$
F	flow rate
H_v	heat of vaporization
n	number of measurements
$\tilde{n}_p$	number of principal components
n_p	total number of available inputs
n_p^*	number of inputs selected for the sensor structure
m	vector of input variables, $\mathbb{R}^{n_p}$
M	matrix of input dataset, $\mathbb{R}^{n \times n_p}$
M_C	matrix of input dataset centered, $\mathbb{R}^{n \times n_p}$
M_N	matrix of input dataset normalized, $\mathbb{R}^{n \times n_p}$
p	pressure
PCT	pressure compensated temperature
Q	energy flow rate
R	universal gas contant
RX	gas/liquid phase ratio
T	termodynamic temperature
x	concentration
y	vector of measurements of output variable,$\mathbb{R}^n$
$\hat{y}$	inferred CV by the inferential sensor
y_C	vector of measurements of output variable centered, $\mathbb{R}^n$
y_N	vector of measurements of output variable normalized, $\mathbb{R}^n$

Currently, the most popular data-based methods for inferential sensor design are based on Principal Component Analysis (PCA) regression (Pearson, 1901) and on Partial Least Squares (PLS) (Wold et al., 1984; 2001). The principle of PCA regression is an application of unsupervised learning to input-variable space reduction and subsequent regression on the reduced space. The use of PCA has a long history yet its use is still very frequent in industry (Kadlec et al., 2009; Yuan et al., 2015; Yu et al., 2020). The characteristics of PLS are similar to PCA regression (Dunn et al., 1989), yet unlike PCA it takes into account also the output space (supervised learning approach). The selection between the use of PCA or PLS is dependent on the availability and quality of infrequently measured output variables.

Both PCA and PLS partially avoid overfitting of the inferential sensor by performing regression in the reduced dimensions. The structure of the resulting sensor is, however, not sparse which might be undesirable or prohibitive, e.g., in case of using the designed inferential sensor for advanced process control. As a response, sparsifying data-driven approaches for the soft-sensor design were developed. The Least Absolute Shrinkage and Selection Operator (LASSO) (Santosa and Symes, 1986; Tibshirani, 2011) uses 1-norm penalization balance between the soft-sensor accuracy and its complexity. The concept of sparse soft-sensor design is further developed in so-called subset selection (SS) methods, which aim at selecting the best subset of explanatory variables from the multivariate input space. The original methodology was proposed to select suitable input variables from the whole set of input candidates according to various (backward, forward, bi-directional) stepwise approaches (Efroymson, 1960; Smith, 2018). Several studies (Mencarelli et al., 2020; Miyashiro and Takano, 2015) proved that SS can be enhanced by using model-overfitting criteria such as adjusted R^2 (R^2_{adj}), corrected Akaike Information Criterion (AICC), or Bayesian Information Criterion (BIC). The performance of SS can also be improved by emulation of the cross-validation process (Takano and Miyashiro, 2020).

In this work, we deal with the data treatment and with the subsequent design of inferential sensors on the pre-treated data. The purpose of the data treatment is to remove the outliers and systematic errors from measurements to make the design of inferential sensors more accurate and reliable. We compare the effectiveness of Hotelling's T^2 distance, MCD and k-means clustering to indicate outliers in a multivariate industrial dataset. Our methodology uses only data-based treatment while the model-based techniques such as data reconciliation also have the potential to enhance the quality of the final designed soft sensors (Manenti et al., 2011; Xenos et al., 2014). The development of first-principles models for the case studies presented in this contribution would be rather complex and is often cumbersome in real industrial conditions. Therefore the data reconciliation is not considered in this paper.

We design linear inferential sensors using various data-driven techniques. The studied methods involve variance-covariance approaches (PCA and PLS) and relatively recent model-sparsity enforcing methods (LASSO and SS). The main contribution of this study is the comparison of these methods in the context of industrial (inferential) soft-sensor design. We analyze the effectiveness of these methods investigating a soft-sensor performance in two industrial use cases provided by the refinery Slovnaft, a.s. in Bratislava, Slovakia. The examples differ in complexity yet they both aim at monitoring the product composition of a distillation column in a crucial processing unit. Recently, there have been several publications (de Morais et al., 2019; Humod et al., 2020; Luo et al., 2020) dealing with the design of inferential sensors for similar industrial units. The soft sensors presented in these publications show satisfying performance in the particular petrochemical process or part of the refinery. While these contributions are focused on a specific soft-sensor design method, we analyze the performance of several methods based on different principles.

The structure of the paper is organized as follows. At first, the basic description of industrial use cases is introduced. Subsequently, the key aspects and relations of well-known data treatment methods are reviewed. Next, the advantages and important characteristics of soft-sensor design methods are briefly introduced. Case studies present results and compare obtained soft sensors for a Fluid Catalytic Cracking (FCC) unit and a Vacuum Gasoil Hydrogenation (VGH) unit. The obtained results are finally discussed and the paper is concluded.

2. Problem description

Our goal is to identify models of inferential sensors in the following linear form:

$$\hat{y}_i = m_i^\intercal (a_1, a_2, \ldots, a_{n_p})^\intercal = m_i^\intercal a, \tag{1}$$

where $\hat{y}$ stands for the desired CV inferred (estimated) by the sensor, m is the vector of available input variables, $a \in \mathbb{R}^{n_p}$ represents the vector of sensor parameters, and index i represents measurements index.

2.1. FCC unit

This unit serves to convert heavy hydrocarbon fractions (vacuum distillates) of the crude oil incoming from the entire refinery to more valuable products, such as gasoline or olefins. The FCC unit is separated into several individual sections (sub-units). One of these sub-units includes several interconnected distillation columns (e.g., debutanizer or depropanizer) to process light hydrocarbons C2–C6. The desired variable ($\hat{y}$) to be inferred by the soft sensor is the composition (main impurity) of the bottom product x_B of the depropanizer column shown in Fig. 1.

The studied depropanizer column processes the feed mixture of nine hydrocarbons C3–C5. The purpose of this column is to separate the feed into C3-fraction-rich distillate product x_D and to C4/C5-fraction-rich bottom product x_B. The available operational degrees of freedom are feed flowrate F, bottom product flowrate B, distillate flowrate D, reflux flowrate R, heat duty in the reboiler Q_B, and heat duty in the condenser Q_D. Most of these variables are available as historical data. These are marked correspondingly in Fig. 1. The plant measurements, also available from historical data, are pressure at the top of the column p_D, pressure at the bottom of the column p_B, and temperatures of distillate T_D, of bottoms T_B, at the top of the distillation column $T_{C,D}$ and at the bottom of the distillation column $T_{C,B}$. The vector of eleven available input variables is given as:

$$m = \left(F, R, Q_B, p_D, p_B, T_D, T_B, T_{C,D}, T_{C,B}, \frac{R}{F}, \frac{Q_B}{F}\right)^\intercal. \tag{2}$$

The use of the thermodynamic properties model to monitor top/bottom stream compositions is prohibitive in this case, even under any appropriate ideality assumptions. This is because there are too many degrees of freedom for the treated multi-component mixture that cannot be inferred from plant data. The current inferential sensor (denoted as Ref), applied in the refinery, uses three out of eleven variables and is designed according to King (2011) as follows:

$$x_B = a_0 + a_1 p_B + a_2 T_{C,B} + a_3 \frac{Q_B}{F}, \tag{3}$$

where a_0 is an intercept, a so-called bias term.

This problem represents a rather standard and well-studied case study of designing an inferential sensor.

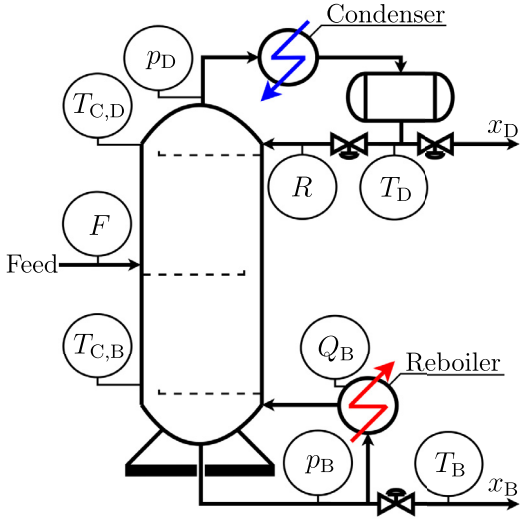


Fig. 1. A schematic diagram of the depropanizer column.

2.2. VGH unit

The purpose of this unit is to process the vacuum distillates by hydrotreating. This unit is separated into a high-pressure reaction section and a low-pressure fractionation section (see scheme in Fig. 2). The main part of the reaction section is represented by the main reactor that hydrogenates the feed. This operation refines the feed from impurities, e.g., nitrogen and sulfur. The reaction section feeds the downstream fractionation section. Here the products are separated into a gasoline fraction (GF), a hydrogenated gasoil (HGO) and other (secondary) products.

Beside the main reactor, the VGH unit involves dozens of low-/high-pressure tanks, heat exchangers, coolers and several distillation columns and furnaces. Furthermore, the unit contains many sensors and control (mostly PI controllers) devices and instrumentation to provide desired operating conditions and products. Overall, there are approximately 1000 historical values available. Therefore, the inferential sensor design for the VGH unit represents a much more challenging problem compared to the case of the FCC unit (11 variables measured at one distillation column).

The variable to be inferred by the soft sensor is HGO product purity expressed in terms of 95% point of distillation curve $T_{95\%,HGO}$. The design of an inferential sensor is performed on the subset of the input variables selected from the whole available dataset. The candidate inputs are selected based on consultation with operators and plant management. The resulting set of 30 candidate inputs is following:

$$\begin{aligned} m = (&PCT_{HGO}, PCT_{GF}, T_{ex,1}, T_{ex,2}, T_{ex,3}, T_{ex,4}, \\ &T_{wabt,1}, T_{wabt,2}, T_{wabt,3}, T_{wabt,4}, T_{wabt,5}, \\ &RX_1, RX_2, RX_3, RX_4, RX_5, RX_6, RX_7, \\ &x_{H2}, T_{frac,1}, T_{frac,2}, F_{frac,heat}, p_{frac}, \\ &F_{f,rec}, F_f, x_{f,N2}, x_{f,S}, T_{f,5p}, T_{f,50p}, T_{f,95p})^\intercal, \end{aligned} \tag{4}$$

with pressure-compensated temperature PCT, exotherms for the reactors T_{ex}, weighted average bed temperatures in the reactors T_{wabt}, ratios of gas/liquid phases in different sections RX, content of the hydrogen in the reaction section x_{H_2}, temperatures in the main fractionator T_{frac}, flow rate of heat medium for main fractionator $F_{frac,heat}$, pressure in the main fractionator p_{frac}, feed flowrate reconciled $F_{f,\,rec}$, feed flowrate F_{feed} and content of impurities in the feed $x_{f,N2}$, $x_{f,S}$, T_f.

The pressure-compensated temperature is calculated according to Clausius-Clapeyron equation (King, 2011):

$$PCT = \frac{1}{\frac{R}{H_v} \ln \frac{P}{P_{ref}} + \frac{1}{T}}, \tag{5}$$

where R is the universal gas constant, H_v is a heat of vaporization, P is an absolute pressure, P_{ref} is a reference pressure and T is the absolute temperature.

Current inferential sensor (Ref) used in the refinery is of the following linear structure:

$$T_{95\%,HGO} = a_0 + a_1 PCT_{HGO}. \tag{6}$$

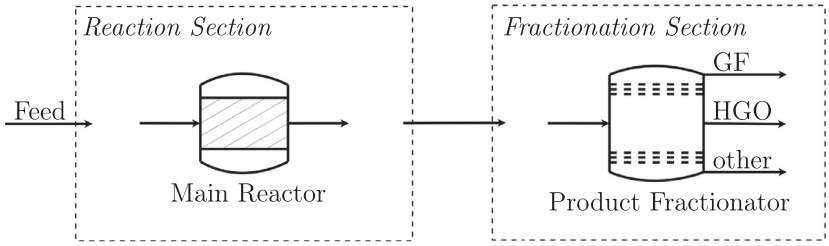


Fig. 2. A schematic diagram of the VGH unit.

This seemingly simple inferential sensor is actually a nonlinear soft-sensor. The operators in the refinery have a good past experience with its performance. However, some recent operating conditions and changes to feedstock in the VGH unit caused significant deviations between estimated values from the reference inferential sensor and the values obtained by the lab analysis. The plant management is unsure about the cause and so this study looks at the whole unit and its operation within up- and down-stream sections.

3. Preliminaries

This section introduces raw data pre-processing methods, methods for multivariate data treatment, and selected methods of soft sensor design. The analyzed dataset includes n measurement points, therefore:

$$M := \begin{pmatrix} m_1^\intercal \\ m_2^\intercal \\ \vdots \\ m_n^\intercal \end{pmatrix}, \quad y := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \tag{7}$$

where M is a matrix of input dataset and y is a vector of output variable measurements.

3.1. Data pre-processing

The multivariate dataset usually contains data on different scales, e.g., due to standards applied in the company for data units. This can inhibit a proper analysis of the dataset covariance and of the impact of variables in the analyzed system. A step to reduce the discrepancy between the variables is the data pre-processing involving the centering and normalization of the data.

The mean-centered data can be obtained by:

$$M_C = M - 1\bar{m}^\intercal, \quad y_C = y - \bar{y}, \tag{8}$$

where M_C is a matrix of centered input dataset, 1 is a vector of ones in $\mathbb{R}^n$, $\overline{m} \in \mathbb{R}^{n_p}$ is a mean vector of M taken column-wise, y_C is a vector of centered output variable measurements and $\bar{y}$ is a mean of the output variable measurements. Subsequently, data normalization ($M_C \to M_N$, $y_C \to y_N$) can be performed such that the data values lie in a desired interval. A commonly used interval is $[-1, 1]$. Another option is the standardization of the data (e.g., required for PCA) to have a zero mean and a unit variance (Kadlec et al., 2009).

3.2. Data treatment methods

Industrial data contains systematic and random errors (Su et al., 2009). The presence of systematic errors in measurements is caused by the non-standard and infrequent situations in the industrial unit, which can be expected (e.g., maintenance) or unexpected (shutdown or plant tripping). Another significant source of systematic errors is failures and inaccuracies (measurement bias) of the sensors.

The detection of a certain class of systematic errors can be carried out through visual inspection in time series plots (Alves and Nascimento, 2007). If the same interval of significantly deviated measurements is distinct in all variables, it suggests a potential source of systematic errors and it needs to be omitted before the design of the inferential sensor. Unlike other errors, this situation is easy to detect using a bare eye.

On the other hand, there are situations when one (or some) of the online sensors is suddenly broken or malfunctions. Such systematic errors can be difficult or impossible to indicate by visual inspection in time series plots. This section refers to several multivariate data treatment methods to reduce the number of systematic errors remaining after the visual detection.

3.2.1. Hotelling's T^2 distance

The Hotelling's T^2 distance (T^2 distance) is based on the distribution developed by H. Hotelling (Hotelling, 1931). This distribution is a generalization of Student's t-distribution. The T^2 distance allows us to analyze multivariate datasets and to detect outliers within the set of different variables, e.g., temperatures, pressures, or flow rates. The key aspect of this distance is a covariance among variables involved in the analyzed data expressed by variance-covariance matrix S:

$$S = \frac{1}{n-1} M_N M_N^\intercal. \tag{9}$$

The covariance matrix is used to determine Hotelling's T^2 distance for each data point:

$$d_{T^2} = \begin{bmatrix} (m_{N,1} - \mu)^\intercal S^{-1} (m_{N,1} - \mu) \\ \vdots \\ (m_{N,n} - \mu)^\intercal S^{-1} (m_{N,n} - \mu) \end{bmatrix}, \tag{10}$$

where $m_{N,i}$ is a vector of measured variables for i^{th} sample point and μ is a mean of the sample.

If the data is normalized (see previous section), $\mu = 0$. According to the values of d_{T^2}, it is possible to determine the most deviated measurements (outliers) from the center. The condition to separate admissible and inadmissible measurements by T^2 distance is usually set as the empirical 3σ rule of thumb (probability to include 99.7% measurements) or using χ^2 test but some tuning might be needed based on the data quality.

3.2.2. Minimum covariance determinant

The Minimum Covariance Determinant (MCD) method (Rousseeuw, 1984) is one of the first tools for the outliers detection with high robustness. The distance metric of MCD is the so-called Mahalanobis distance given by the following equation:

$$d_{MCD} = \begin{bmatrix} \sqrt{(m_{N,1} - \mu)^\intercal S^{-1} (m_{N,1} - \mu)} \\ \vdots \\ \sqrt{(m_{N,n} - \mu)^\intercal S^{-1} (m_{N,n} - \mu)} \end{bmatrix}, \tag{11}$$

which is closely related to Hotteling's T^2 distance (10). Despite the similarity of the distance metrics of these methods, the principle of MCD is quite different from Hotteling's T^2 method. MCD looks for the subset of measurements with the minimum determinant of the corresponding covariance matrix. In other words, the resulting subset of measurements should occupy the smallest volume possible (determinant of the covariance matrix). The algorithm can be viewed as an enhancement of Hotteling's T^2 distance method.

The iterative algorithm of MCD starts with a random guess of the initial subset. Subsequently, the mean μ and covariance matrix S of the initial subset are calculated. According to the calculated μ and S, it is possible to evaluate d_{MCD} from (11) for each measurement (not only for the selected subset). Subsequently, the new subset of h measurements with the smallest distances d_{MCD} is selected from the whole set. If the covariance determinant of the new subset is decreased compared to the covariance determinant of the previous subset, the new subset is used in the next iteration of the MCD algorithm. Otherwise, the sought subset has been found (as the previously selected subset) and the MCD algorithm is terminated. The tuning parameter of this scheme is represented by the least number of the retained measurements h from the treated dataset. This parameter is usually adjusted according to the interval $\frac{n+n_{\text{p}}+1}{2} \leq h \leq n$ (Hubert and Debruyne, 2010) or it can be adjusted by the user e.g., based on the visual inspection of the time series of some crucial variables.

Due to the random character of this method, it is desired to perform several runs with different initial guesses to avoid local minima. According to the results from different runs of the MCD method, it is possible to derive a final subset. The vector of distances d_{MCD} is evaluated in each iteration. The measurements with the smallest distances create a new subset for the next iteration of MCD. This process is terminated when the determinant of the covariance matrix does not decrease anymore.

The mean μ and the covariance matrix S of the final subset are subsequently used to evaluate d_{MCD} from (11) for the whole set. According to the values of d_{MCD}, it is possible to determine the most deviated measurements (outliers) from the center. The condition to separate admissible and inadmissible measurements by MCD is established by appropriate distribution (Hardin and Rocke, 2005) considering the desired confidence level.

3.2.3. k-means clustering

The k-means clustering method (Forgy, 1965) separates measurements from the multivariate dataset into different groups (clusters). Each measurement is assigned to the cluster according to the closest center of a cluster (i.e., mean of the cluster data points). The area of clusters should be as small as possible yet the data points in the different clusters should be as far from each other as possible.

The k-means clustering is able to adjust the performance by using different distance metrics d_{CL}. The frequently used distance metric is the squared Euclidean distance in the form:

$$d_{\text{CL},i,j} = (m_{\text{N},i,j} - \mu_j)^{\intercal}(m_{\text{N},i,j} - \mu_j), \tag{12}$$

where $\mu_j,\ \ j \in \{1, 2, \ldots, k\}$ is the center of the j^{th} cluster representing the mean of the corresponding data points and the index i characterizes the ordinal number of the data points within the j^{th} cluster.

The selection of the desired number of clusters (k) is highly related to the data quality. In trivial cases, it is possible to determine the value of k by visual inspection, where the data points create visible groups representing different operating conditions of the unit. In a non-trivial case, it is possible to determine the value of k according to the elbow method or using various goodness of fit criteria (Kodinariya and Makwana, 2013).

The algorithm of k-means clustering is initiated by a random guess of the locations of the desired centers. Therefore several runs of the algorithm should be performed with different initial guesses. The measurements should be assigned to the clusters with the highest frequency of the assignment from the different runs of the algorithm, similar to the final subset of the MCD method. Once the final clusters are created, it is possible to determine the outliers as measurements in the particular cluster. It can be seen that the clusters predominantly constituted by outliers contain a smaller amount of data compared to the rest of the clusters.

According to the nature of the aforementioned data treatment methods, one could expect the performance of MCD at least as good as the performance of the T^2 distance method. The k-means clustering can outperform the rest of the methods if measurements involve several clearly distinct operating points (steady states) of a particular unit.

3.3. Methods for inferential-sensor design

We study data-driven methods for soft-sensor design. Each method solves two sub-problems: the structure selection of the soft sensor and the calculation of soft sensor parameters. The investigated methods are based on different principles: on the analysis of the variance-covariance matrix of the dataset (PCA and PLS) and on sparsity enforcement (LASSO and subset selection).

An effective design procedure usually requires splitting the available dataset (input matrix M_{N}, output vector y_{N}) into the following subsets: dataset for sensor design that contains training data $(M_{\text{N}}(\mathcal{T}), y_{\text{N}}(\mathcal{T}))$ and dataset used for the performance evaluation of designed soft sensors that contains testing data $(M_{\text{N}}(\mathcal{S}), y_{\text{N}}(\mathcal{S}))$. Here $\mathcal{T}$ and $\mathcal{S}$ denote the corresponding row-selection operators.

3.3.1. Ordinary least squares regression

The basic method of soft-sensor design is Ordinary Least-Squares Regression (OLSR). This method estimates the parameters of an inferential sensor according to

$$\min_a \frac{1}{2} \sum_{\forall i \in \mathcal{T}} (y_{\text{N},i} - m_{\text{N},i}^{\intercal} a)^2 \equiv \min_a \frac{1}{2} \|y_{\text{N}}(\mathcal{T}) - M_{\text{N}}(\mathcal{T})a\|_2^2, \tag{13}$$

which minimizes the sum of squared errors between measurements and sensor predictions.

The method can potentially result in a sparse sensor structure, e.g., when strong linear dependencies exist among some variables. One can thus talk about OLSR being able to select (sparsify) the sensor structure. Generally, OLSR cannot effectively (or actively) strive against overfitting. Its performance has to be usually enhanced in combination with other methods that consider not only the accuracy but also the complexity of the resulting model.

3.3.2. Principal component analysis

Principal Component Analysis (PCA) (Pearson, 1901) is a method of identifying an $\tilde{n}_{\text{p}}$-dimensional subspace ($\tilde{n}_{\text{p}} \leq n_{\text{p}}$) of orthogonal coordinates that exhibit a maximum variance in a given dataset.

The principal components are identified by the eigendecomposition of the covariance matrix S of the mean-centered, unit-variance data by taking the eigenvectors (for subset definition) and the associated eigenvalues (for measure of variance). Each eigenvector represents one principal component that explains a certain amount of the data variance. The desired amount of total variance can be captured by selecting several ($\tilde{n}_{\text{p}}$) principal components within the eigenvectors with maximum explained variance. The regression is then carried out over the selected (principal components) subspace using (13) with $\tilde{n}_{\text{p}}$ parameters.

The usage of PCA regression represents an advantage mainly in the case of an insufficient amount of the output data. In fact, this

is the usual situation in the industry, where the measurement of the desired output variable is too expensive or rare. Such situation leads to performance deterioration of many data-driven methods for soft-sensor design as they usually require large number of measurements. The PCA regression method has gained its popularity because of being able to learn from the measurements of the online sensors and thus being able to outperform other data-driven methods in certain cases.

3.3.3. Partial least squares

Partial Least Squares (PLS) regression (Wold et al., 1984) is a statistical method searching for a linear regression model of predicting the output (predicted) variable using input variables by the projection into a new space of principal components. Although PLS regression is not an unsupervised learning approach (as PCA), the principle of these methods is essentially the same (Wold et al., 1984) and both methods are intended to reduce the dimensionality of the problem.

The most common approaches for PLS regression are nonlinear iterative partial least squares (NIPALS) and SIMPLS (de Jong, 1993). Both approaches iteratively calculate the principal components. The principal components are calculated by SVD decomposition of the following cross-covariance matrix S_{cross}:

$$S_{\text{cross}} = \frac{1}{n-1} M_{\text{N}} y_{\text{N}}^{\intercal}. \tag{14}$$

The selection of the desired number of principal components is then performed in the same way as in the case of PCA. Subsequently, the principal components are used to design the soft sensor much like in the case of PCA.

3.3.4. Least absolute shrinkage and selection operator

Least Absolute Shrinkage and Selection Operator (LASSO) (Santosa and Symes, 1986) is a method that simultaneously identifies the structure of the model and its parameters by solving the following optimization problem:

$$\min_{a} \ \frac{1}{2} \sum_{\forall i \in \mathcal{T}} (y_{\text{N},i} - m_{\text{N},i}^{\intercal} a)^2 + \lambda \|a\|_1, \tag{15}$$

where λ is a weight between the accuracy of the model training and the model overfitting. The magnitude of the ℓ_1-penalization element results in certain parameters being equal to zero at the optimum of (15). The resulting model is then less complicated and usually more interpretable.

The LASSO technique belongs to the regularised regressions family. Beside the LASSO regression, this family involves many other methods, but the most important ones are ridge regression (Hoerl and Kennard, 1970) and elastic net (Zou and Hastie, 2005). The ridge regression has a similar objective function compared to LASSO. However, the ridge regression uses the ℓ_2-penalization element to reduce the value of all parameters. The usage of ridge regression is preferred when the input variables are highly correlated. The elastic net technique effectively combines LASSO and ridge regression. It weighs between ℓ_1-penalization element and ℓ_2-penalization element within the objective function.

3.3.5. Optimal subset selection with model-overfitting criteria

Subset selection denotes a class of methods that explicitly seek for the simplest possible sensor structure such that some model-overfitting criterion $J(a, z)$ is minimized (Miyashiro and Takano, 2015). Here the variable z denotes a vector with binary entries $z \in \{0,1\}^{n_{\text{p}}}$ signifying the selection of j^{th} input into the sensor structure. Correspondingly, the sum of the vector entries $\sum_{j=1}^{n_{\text{p}}} z_j = 1^{\intercal} z$ denotes the sensor complexity.

Optimal subset selection solves the following bi-level program (Bertsimas et al., 2016):

$$\min_{a,\, z \in \{0,1\}^{n_{\text{p}}}} \ J(a, z) \tag{16a}$$

$$\text{s.t. } a \in \arg\min_{\tilde{a}} \frac{1}{2} \|y_{\text{N}}(\mathcal{T}) - M_{\text{N}}(\mathcal{T})\tilde{a}\|_2^2 \tag{16b}$$

$$-\bar{a} z_j \le \tilde{a}_j \le \bar{a} z_j, \forall j \in \{1, \dots, n_{\text{p}}\}, \tag{16c}$$

where $\bar{a}$ represents an upper bound on $\|a\|_\infty$ to be tuned and the optimization criterion $J(\cdot)$ might take the form ($\text{RSS} := \|y_{\text{N}}(\mathcal{T}) - M_{\text{N}}(\mathcal{T})a\|_2^2$):

$$J_{R^2_{\text{adj}}} = \frac{\text{RSS}}{n - 1^{\intercal} z - 1}, \tag{17}$$

or

$$J_{\text{AIC}_{\text{C}}} = n \log \frac{\text{RSS}}{n} + 2(1^{\intercal} z), \tag{18}$$

or

$$J_{\text{BIC}} = n \log \frac{\text{RSS}}{n} + \log(n)(1^{\intercal} z). \tag{19}$$

The bi-level program (Eq. (16)) can be effectively resolved by standard MIQP solvers using big-M reformulation as shown in Takano and Miyashiro (2020).

3.3.6. Optimal subset selection with cross-validation criterion

The principle of Subset Selection with Cross-Validation Criterion (SS-CV) is to mimic a standard cross-validation procedure within the training dataset. Let us divide the training data into K smaller subsets $\mathcal{N}_k$, such that:

$$\mathcal{T} = \bigcup_{k \in K} \mathcal{N}_k, \quad \mathcal{N}_k \cap \mathcal{N}_{k'} = \emptyset, \quad \forall k \neq k', \quad K \ge 2. \tag{20}$$

The data is distributed into training ($\mathcal{T}_k$) and validation ($\mathcal{V}_k$) sets as follows:

$$\mathcal{V}_k := \mathcal{N}_k, \quad \mathcal{T}_k := \mathcal{T} \setminus \mathcal{N}_k, \quad \text{card}(\mathcal{T}_k) \ge n_{\text{p}}, \quad \forall k \in K. \tag{21}$$

where $\mathcal{V}_k$ sets contain unique data, while the different $\mathcal{T}_k$ sets involve recurring measurements. The optimal SS with cross-validation solves (Takano and Miyashiro, 2020):

$$\min_{a^{(k)}, \forall k \in K,\ z \in \{0,1\}^{n_{\text{p}}}} \ \frac{1}{2} \sum_{k=1}^{K} \|y_{\text{N}}(\mathcal{V}_k) - M_{\text{N}}(\mathcal{V}_k) a^{(k)}\|_2^2 \tag{22a}$$

$$\text{s.t. } \forall k \in K: \ a^{(k)} \in \arg\min_{\tilde{a}} \frac{1}{2} \|y_{\text{N}}(\mathcal{T}_k) - M_{\text{N}}(\mathcal{T}_k)\tilde{a}\|_2^2 \tag{22b}$$

$$\text{s.t. } -\bar{a} z_j \le \tilde{a} \le \bar{a} z_j, \ \forall j \in \{1, \dots, n_{\text{p}}\}. \tag{22c}$$

The problem (22) can be solved for several values of K—considering constraints on parameter identifiability, i.e., the cardinality condition in Eq. (21)—and for different randomly generated distributions of data into $\mathcal{T}_k$ and $\mathcal{V}_k$ sets. The structure of the resulting sensor is then given by the most frequent inputs occurring in the calculated sensors. Once the optimal sensor structure is calculated, a least-squares fitting of such a model is used with the entire training dataset to determine the parameters of the designed soft sensor. Similarly to problem (16), the problem (22) can be effectively resolved by standard MIQP solvers.

4. Results

We present the results for both the presented use cases. We compare the performance of the presented data treatment methods and methods for soft-sensor design. Due to data confidentiality, the graphical representations of the results use the normalization of variables in the interval $[0, 1]$.

4.1. Implementation details

The implementation of all the presented methods is performed in MATLAB. For the initial data treatment, we use the Hotelling's T^2 distance considering χ^2-distribution with the probability of including 99.7% measurements. For the MCD method, we select the value of parameter h as a midpoint of the interval $\frac{n+n_p+1}{2} \leq h \leq n$ (Hubert, Debruyne, 2010). The outliers are determined by MCD considering an approximation of F-distribution (Hardin, Rocke, 2005) with the same probability as in the T^2 distance method. As a preliminary analysis suggested, the industrial data seem to be not normally distributed. Therefore the T^2 distance method considering χ^2-distribution tends to remove larger portions of data than MCD with F-distribution. The number of the desired clusters for the k-means clustering is determined using the elbow method. The results of the MCD method and the k-means clustering are gathered and averaged over 100 different runs of the respective algorithms. This is because of the inherent randomness of these methods, as mentioned above.

For the soft-sensor design, we set the variance-covariance methods (PCA and PLS) to select the amount of the variance explained by the principal components to at least 98%. The PLS method uses SIMPLS approach from MATLAB. We use Yalmip (Löfberg, 2004) and Gurobi (Gurobi Optimization LLC, 2020) to solve various instances of the problems (13), (15), (16), and (22).

We will study two different scenarios of soft-sensor design for each use case w.r.t. splitting the data into training and testing subsets. In both scenarios, the training set is used for the design of the reference as well as the rest of the studied inferential sensors. In the first scenario, the available dataset is divided into subsets based on the time series. The data from an earlier time period is used for training and the data from a later time period is used for testing. This situation simulates soft-sensor design at a certain point in time using historical (training) data. The testing phase then mimics the future sensor performance, where the sensor is employed without any adaptation of its structure despite possible variations in plant operating conditions.

The second scenario groups the available data among training/testing subsets randomly. The results thus reveal the potential of the studied sensor-design methods for adaptation of the sensor structure to the changing operating conditions. In this scenario, the final results are gathered from 50 runs with different training/testing dataset distributions.

In order to tune the value of λ in (15) we use the goodness-of-fit criteria (17)–(19) and cross-validation on the training set. We first obtain the candidate values of λ that minimize one of the goodness-of-fit criteria by training the sensors on the whole training set. Subsequently, we generate twenty different distributions of the training data into two subsets (similar to the SS-CV method). The candidate values of λ are used for regression and cross-validation on the generated subsets and the best performing value is used for the final sensor training.

When determining the final design of the soft sensor according to SS with cross-validation, we take a median of $1^\intercal z$ from the results of the different runs (different validation data distribution and different values of K: $K \leq 6$ for the FCC unit, $K \leq 4$ for the VGH unit) to obtain the $n_p^* \leq n_p$, i.e., the number of inputs of the final sensor. Subsequently, we select the n_p^* most frequent inputs from the results of the different runs to finalize the sensor structure.

The complexity of each designed soft-sensor structure is determined according to the number of input variables n_p^*. We measure the impact of a particular input on the soft-sensor performance by the value of $|a_i|$. If the impact of a particular term is less than 0.1% of the maximum value of the desired inferred variable, we neglect the corresponding part of the soft sensor.

The accuracy of the soft sensors is evaluated and compared by the root mean square error (RMSE) of the sensor prediction on the testing dataset. The performance of the industrial soft sensors can be adjusted during the operation by an adaptive bias correction, also called bias update. Many industrial software solutions offer this form of soft-sensor maintenance against the change of operating conditions. The purpose of the bias correction is to improve the accuracy of the succeeding predictions of the soft sensor by adjusting the constant (bias) term a_0. The bias is updated when a measurement from a lab analysis is available and when it differs significantly from the sensor prediction (Quelhas, 2009). The frequency of bias correction is thus, on the one hand, a measure of frequency of the change in the plant operating conditions. On the other hand, it reflects the ability of the sensor itself to react to the changes in the operating conditions. The plant operators trust more a soft sensor with less frequent bias updates. Therefore, in addition to the soft-sensor complexity (n_p^*) and accuracy (RMSE), we evaluate the effort of the bias correction (BC) by simulating a bias correction procedure in parallel, i.e., without affecting the prediction error of the sensor evaluated by RMSE. The measure of the bias-correction effort is expressed as the percentage of measurement-based sensor corrections occurrences in the testing dataset.

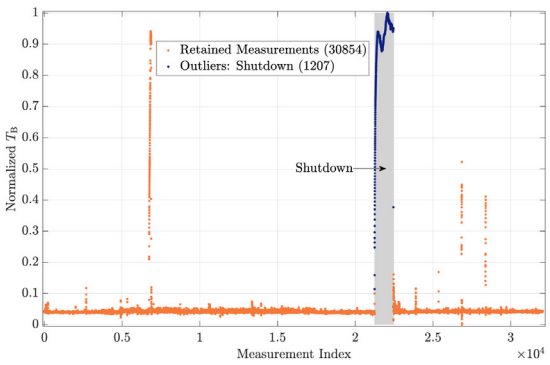
4.2. Inferential sensors for the FCC unit

The available historical data involving 32,061 measurement points from online sensors (candidate input variables) represents more than two years of production in the period 2016–2019. This time span contains 181 lab measurements of the bottom product concentration x_B (output variable).

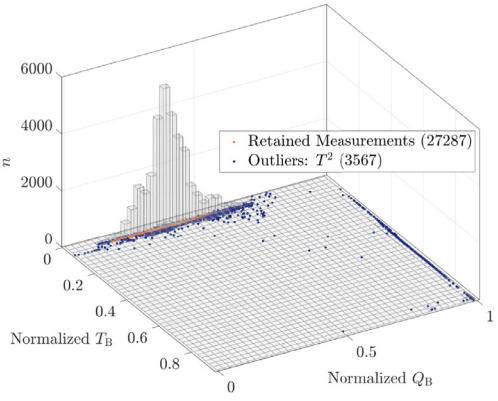
We first perform the data treatment to reduce the amount of systematic and gross errors. Fig. 3(a) shows visualization of the data treatment results on the normalized temperature of the bottom product T_B. The visual inspection of the time series of the available data (data pre-treatment) reveals the initial set of systematic errors with significantly deviated data, which corresponds to the shutdown period of the unit. This is marked as a thick gray bar in Fig. 3(a). The unit operators confirmed in consultation the correctness of omission of the corresponding 1207 data points from the further processing.

Subsequently, we applied the T^2 distance, MCD, and k-means clustering methods to detect outliers in the dataset. The performance of these methods is individually visualized and compared in Fig. 3(b)–(d) for lucidity. Each figure shows a histogram of data points of bottom product temperature vs. reboiler heat duty. All the methods clearly identify the most distinct outliers. The results further show that k-means clustering (Fig. 3(d)) might be overly conservative as it selected significantly fewer outliers than the other two methods. The low performance of this method is caused by the complex tuning (e.g., number of clusters). The k-means clustering method detects only five data clusters, which results in the low number of indicated outliers by this method. The number of outliers indicated by the MCD method (Fig. 3(c)) is almost twice higher compared to the T^2 distance method (Fig. 3(b)). The MCD method thus appears as a reasonable choice here as it removes a significant amount of outliers, yet retains reasonable number of data points, of which it guarantees better quality than the T^2 distance approach.

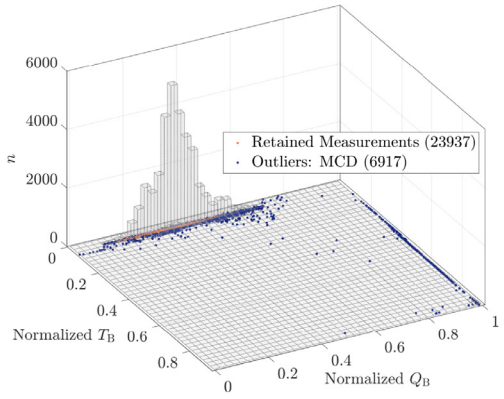
It is obvious that the majority of identified outliers (blue points in Fig. 3(c)) by the MCD method deviates from the area with the highest density of the online measurements. On the same line, the approved measurements (green points in Fig. 3(c)) are located inside or are very close to this area. This also indicates the good performance of the MCD method. The final set of the retained measurements by this method for the soft-sensor design is shown in



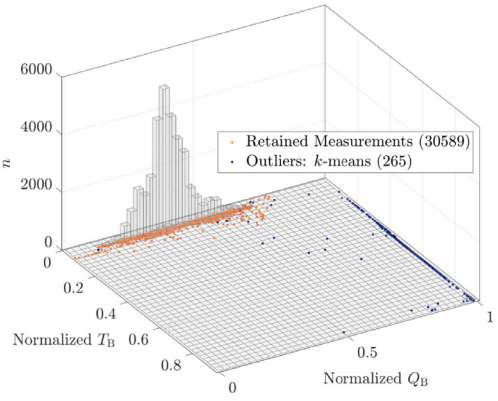
(a) Data pre-treatment by visual inspection detecting plant shutdowns.



(b) Data treatment by the T^2 distance method (3567 outliers)



(c) Data treatment by the MCD method (6917 outliers)



(d) Data treatment by the k-means clustering method (265 outliers)

Fig. 3. **(a)** Normalized bottom product temperature of the FCC unit vs. measurement index. **(b), (c), (d)** Histogram of the bottom product temperature vs. reboiler heat duty of the FCC unit and retained measurement vs. outliers as detected by data treatment methods.

Fig. 4. It is evident that the MCD method provides well-poised data set, which appears to be close to normal distribution. We can conclude that the available industrial data are of good quality and that the conducted data treatment was able to reveal the high-quality data.

4.3. Design of inferential sensors for the FCC unit using time series data

We first study a scenario where the (chronologically) first 50% of the available data is assigned to the training set and the last 50% of data is assigned to the testing set.

Soft-sensors designed by PCA and PLS require six and seven principal components, respectively, to explain 98% of the variance in the data. This relatively high number of principal components suggests, on the one hand, to use a more complex structure of soft sensor than the reference soft sensor. On the other hand, sensors designed by these methods might be overfitted.

When designing a soft sensor by the SS methods, we compared the performance of the presented overfitting criteria (R^2_{adj}, AIC_C, BIC). We used the principle of parsimony. The simplest sensor yet the best performing one is designed by SS with R^2_{adj} criterion. This sensor is the same as suggested by SS with cross-validation in this case and it is selected for further performance analysis.

Table 1
Comparison of the number of inputs n_p^* (number of principal components for PCA and PLS shown in brackets), sensor accuracy (RMSE) and bias correction relative frequency (BC) using time series data for the FCC unit.

	OLSR	PCA	PLS	LASSO	SS	SS-CV	Ref
n_p^*	11	11 (6)	11 (7)	5	4	4	3
RMSE	0.120	0.096	0.104	0.099	0.099	0.099	0.117
BC [%]	29.7	21.6	24.3	20.3	23.0	23.0	28.4

A comparison of the designed sensors in terms of their complexity (n_p^*), accuracy (RMSE), and the effort of the bias correction (BC) is shown in Table 1. The results clearly suggest to enrich the structure of reference soft sensor to include at least one extra variable in order to improve its performance (see n_p^* in Table 1). The least complex sensors are suggested by the LASSO and SS methods. These methods suggest replacing bottom pressure p_B by temperatures $T_{C,D}$ and $T_{C,B}$ (LASSO selected also the ratio R/F). These sen-

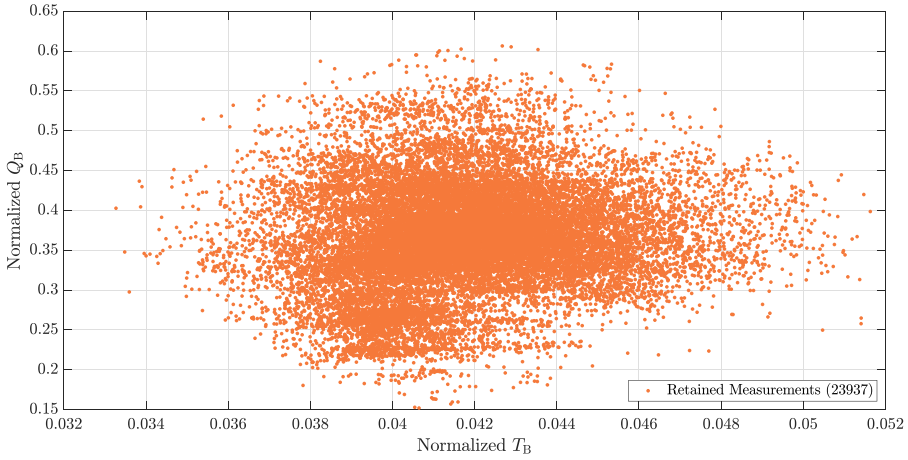


Fig. 4. The retained online measurements (by the MCD method) of the bottom product temperature and reboiler heat duty of the FCC unit.

sors (including PCA) exhibit a reduced amount of bias correction compared to all others.

Overall, the accuracy of the reference soft sensor (see RMSE in Table 1) shows almost the worst performance. Only the (most likely overfitted) soft sensor designed by OLSR is worse in this comparison, despite using all the possible eleven inputs. The overfitting by OLSR can be documented by worsened accuracy and also by a high effort of the bias correction.

The highest sensor accuracy is achieved for the PCA-based soft sensor. The improvement compared to the reference soft sensor is approximately 18%. Other proposed advanced sensors show similar performance (improvements of at least 15%).

Looking at the amount of bias correction, we can see that the most frequently corrected soft sensor is designed by OLSR, while the soft sensor designed by LASSO requires the bias correction less frequently than others. The best sensor would be selected as a compromise between accuracy, complexity, and maintenance (BC) effort. In this respect, all the advanced designed soft sensors represent good candidates.

In order to provide a more comprehensive comparison of the soft sensors, Fig. 5 visualizes their predictive performance on the output variable. The lab-analysis data is shown as black squares (training dataset) and black stars (testing dataset), respectively. The data show significant variability indicating several changes of the operating conditions within the studied time window, in both training and testing datasets. This means that the trained sensors face a rich portfolio of situations and thus a trained sensor can be expectedly valid for a long time after its commissioning. This is confirmed by the aforementioned good performance of the designed sensors and by the relatively low effort of the bias-update mechanism.

Fig. 5 further presents the training and testing (predictions) performance of the designed advanced soft sensors, by PCA and PLS (Fig. 5(a); green solid line and red dashed line, respectively) and by LASSO and SS-CV (Fig. 5(b); magenta solid line and green dashed line, respectively), compared in both figures to the reference soft sensor (blue dotted line).

When looking at the performance of the reference sensor in both plots, one can clearly identify several points, where the reference sensor is not able to explain the measurements yet the advanced sensors are. This is present throughout the whole studied time window but it is most evident in the testing phase (around the measurements 80–120).

We can see that despite the behavior of the soft sensors designed by PCA and by PLS being similar in the training phase, the evolution of the predictions of these sensors on the testing data is quite different. This also explains differences in the accuracy and frequency of the bias correction. It also further supports our earlier conjecture of possible overfitting present in these sensors. This observation is in contrast with the bottom plot (LASSO and SS-CV), where the outputs of the visualized advanced sensors are almost identical.

Table 2
Comparison of the number of inputs n_p^* (number of principal components of PCA and PLS), sensor accuracy (RMSE) and bias correction relative frequency (BC) over 50 random training/testing data distributions for the FCC unit.

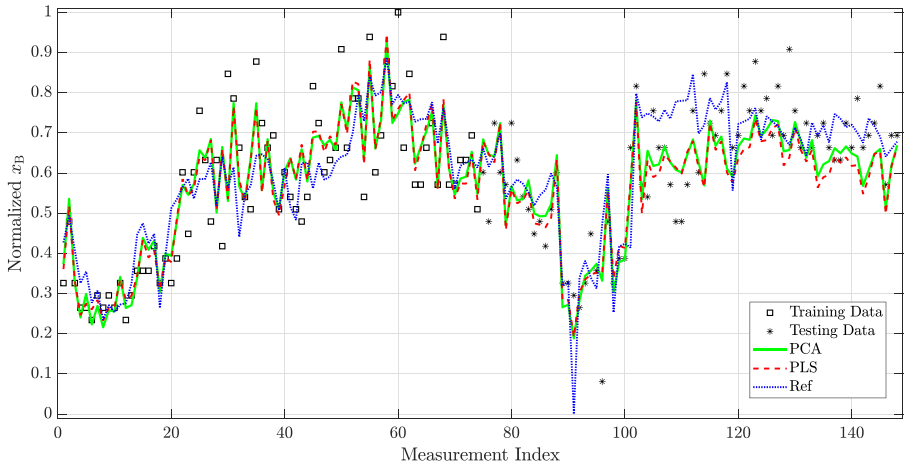
	OLSR	PCA	PLS	LASSO	SS	SS-CV	Ref
n_p^*	11	11 (7)	11 (8)	7	6	5	3
RMSE	0.105	0.104	0.106	0.106	0.106	0.110	0.121
BC [%]	23.0	24.3	23.0	24.3	27.0	24.3	28.4

A noticeable part of the testing phase is the last period (around the measurements 130–148), where it seems that the operating conditions in the FCC unit change considerably. There exist corresponding significant discrepancies between the measurements and values inferred by all the advanced soft sensors. The reference soft sensor, however, performs well here, which suggests good robustness properties of this sensor. All the advanced sensors exhibit a slower or faster drift from the measurements. This situation calls for sensor maintenance or complete structural change. It appears that a practical solution of performing bias update would be sufficient. We will revisit and analyze this issue in the following section in order to confirm whether the operating conditions change so dramatically that one would need to change the soft sensor structure.

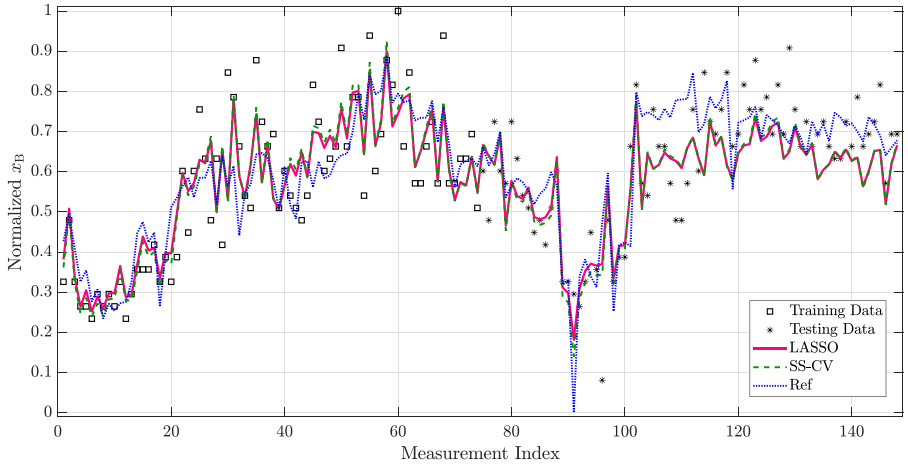
4.4. Design of inferential sensors for the FCC unit using randomly distributed data

We randomly distribute 50% of the available data to the training set and the remaining data to the testing set. We generate 50 such distributions to increase the interpretabilily of the results. We then use the same workflow to design the soft sensors as outlined above.

We report averages of n_p^*, RMSE and BC for each soft sensor over the 50 data distributions in Table 2. According to the sensor complexity criterion (n_p^*), we can see that the designed soft sensors suggest more complex structure (at least two extra input variables) compared to the reference structure and also compared to the previous scenario with chronological training/testing data assignment. This suggests that varying operating conditions in the plant would require frequent revision of the sensor structure for better perfor-



(a) Training and prediction performance of the sensors designed by PCA and PLS methods and of the reference sensor.



(b) Training and prediction performance of the sensors designed by LASSO and SS-CV methods and of the reference sensor.

Fig. 5. Comparison of the soft sensors for the FCC unit designed using time series data.

mance. The performance of the designed advanced sensors does not improve compared to the designs using chronological training/testing data distribution, which is a consequence of the overfitting implied by the increased complexity of the sensor. For example, LASSO and both SS methods commonly suggest including T_D and Q_B on top of the inputs suggested in the previous section. However, none of these variables seem to be significantly useful for the sensor overall. While, unlike for distillate temperature T_D, inclusion of Q_B would make sense from process viewpoint, its effect is already present in the input Q_B/F. Only the inferential sensor designed by OLSR exhibits improved accuracy compared to the design with chronologically distributed training/testing data. This is a consequence of providing better training data (more similar to testing ones) to the sensors, which reduces the overfitting effect. Designed advanced soft sensor (including PCA) shows the increased frequency of the bias correction, which can be attributed to the large noise magnitude in the lab data and overfitting.

The performance features of the particular sensors remain practically the same as in the case of chronological training/testing data distribution. The soft sensor designed by PCA is slightly more accurate than other soft sensors and it improves the accuracy of the

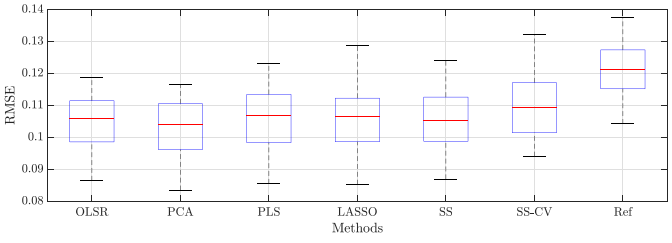
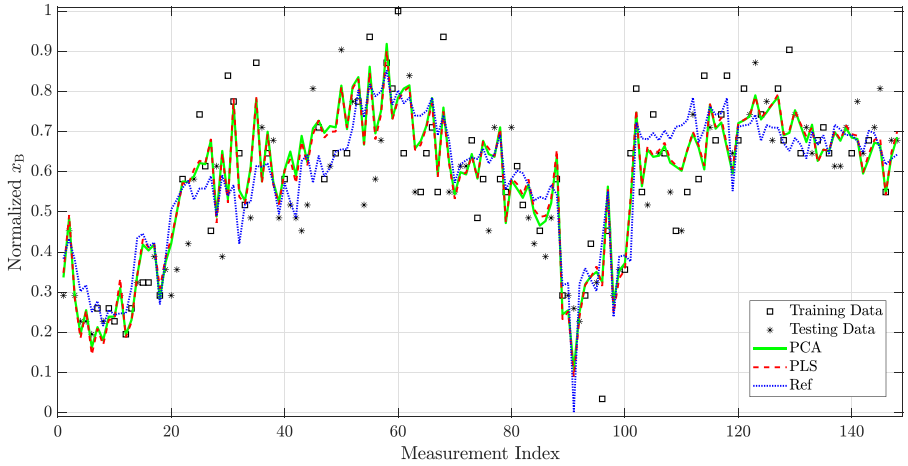


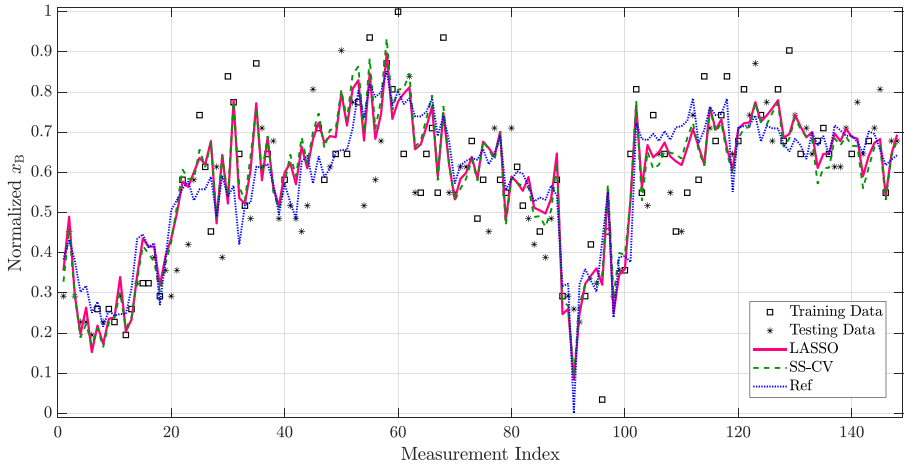
Fig. 6. Comparison of accuracy of the designed inferential sensors over 50 different random training/testing data distributions for the FCC unit.

reference soft sensor by about 14%. Yet the drop in this improvement confirms the overfitting. The structure of the soft sensor designed by SS-CV is less complicated than the structures of other designed soft sensors. As expected, the least complex sensor designed by SS-CV is again followed in terms of performance by design using the SS and LASSO methods, respectively.

Fig. 6 visualizes the accuracy statistics of each soft sensor from the 50 randomly generated training/testing datasets using



(a) Training and prediction performance of the reference sensor and of the sensors designed by PCA and PLS methods.



(b) Training and prediction performance of the reference sensor and of the sensors designed by LASSO and SS-CV methods.

Fig. 7. Comparison of the soft sensors using randomly distributed training/testing data for the FCC unit.

box plots. The central horizontal-line marker indicates the median, the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively, the whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted individually using the '+' symbol. We can see that the median performance mostly copies the average performance of the designed soft sensors outlined in Table 2.

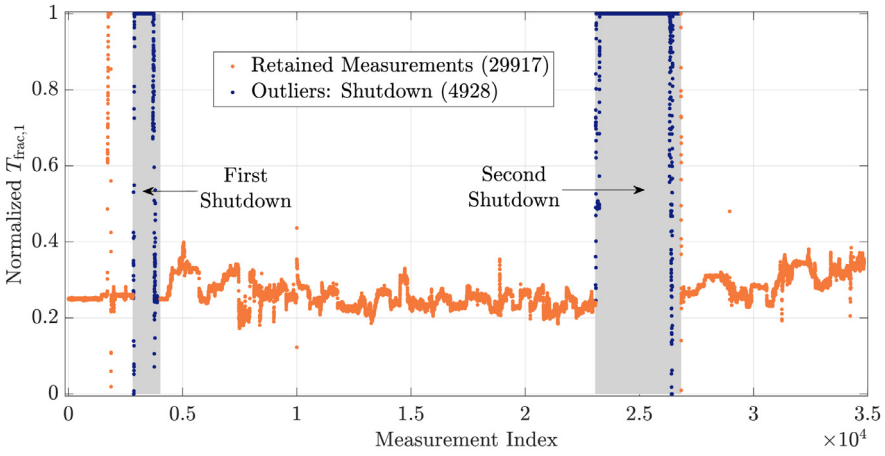
The accuracy variance seems to be considerable for all sensors, which confirms the aforementioned large noise in the samples the possible sensor overfitting. The least variance is present in the reference sensor, which is due to the aforementioned robustness properties.

As in the previous section, we visualize the training and prediction performance of the designed soft sensors in Fig. 7 for one representative random training/testing data distribution. We again show results obtained for the reference soft sensor (both plots; blue dotted line), the soft sensors designed by PCA and PLS (Fig. 7(a)); green solid line and red dashed line, respectively), and the soft sensors designed by LASSO and SS-CV (Fig. 7(b); magenta solid line and green dashed lines, respectively).

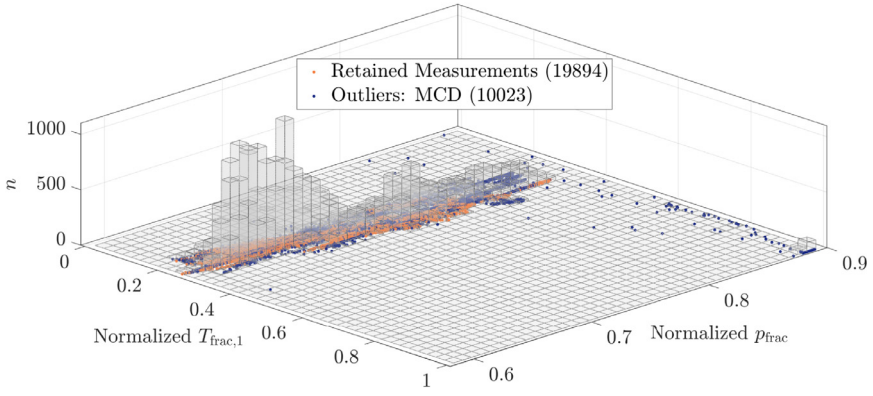
As can be expected, the performance of the soft sensors is similar to the performance of the soft sensors designed by using time series data. The previously discussed discrepancy between the sensors and the measurements (around the measurements 130–148) is decreased. This, together with the increased complexity of the sensors designed using randomly distributed data, leads us to the conclusion that the performance of an advanced sensor can only be maintained if the sensor structure changes frequently or if the sensor parameters are frequently updated. Of course, in this particular case, the problem would be practically resolved by bias update.

4.5. Inferential sensors for the VGH unit

The available historical data encompasses almost two years of production in the period 2018–2019 with 34,845 time points of online measurements. This is a comparable amount of data as in the previous case study. The desired output variable $T_{95\%,HGO}$, which, similarly to the previous use case, indicates the purity of the distillation product, is determined by the lab analysis. The mentioned time span involves 689 measurements of the output variable as it is measured more frequently than in the case of FCC.



(a) Data pre-treatment by visual inspection detecting plant shutdowns.



(b) Data treatment by the MCD method (10023 outliers).

Fig. 8. **(a)** Normalized temperature in the main fractionator of the VGH unit vs. measurement index. **(b)** Histogram of the temperature vs. pressure in the main fractionator of the VGH unit and retained measurements vs. outliers after data treatment.

We first perform the pre-treatment of the available data. Based on the visual inspection of the time series of a temperature in the main fractionator (Fig. 8(a)), we eliminated two intervals with obviously deviated measurements (see gray intervals in Fig. 8(a)). The unit operators confirmed that the omitted 4928 measurement points (black points in Fig. 8(a)) correspond to the unit shutdowns.

Subsequently, the remaining data is processed using the T^2 distance, MCD, and k-means clustering methods. Although the visualized temperature data does not seem to be much qualitatively different in nature than the case of the FCC unit, there are more distinct variations and steady states. This feature causes that the T^2 distance method suggests removing more outliers than MCD and k-means clustering.

For the case of the T^2 distance method, 13,874 outliers is indicated, which represents almost half of the available pre-treated data. This behavior can be attributed to the previously observed distinct variations and steady states, which bias the statistics used in the T^2 distance method. In fact, if we wanted to tune this method to the similar performance as the MCD method, we would require increasing the probability of measurements acceptance from 99.7% to 99.9%. This seemingly small alteration represents a significant increase in the acceptance, by one half of a standard deviation.

The MCD method indicates slightly more outliers (10,023 measurements) as in the case of the FCC unit (6,917 measurements), which may be caused by the worse quality of the data from the VGH unit. The k-means clustering method indicates much more outliers (11,229 measurements) than in the FCC unit (265 measurements). It uses 21 clusters (compared to five clusters detected for the FCC unit), which seems to be a consequence of the distinct variations and steady states. Nonetheless, the data distribution among the clusters exhibits certain uniformity, which further demonstrates the sensitivity of the k-means clustering method to tuning (e.g., number of clusters).

As in the case of the FCC unit, we again choose to remove the outliers labeled by the MCD method as it retains reasonable amount of data points. Even though the data quality (e.g., number of shutdowns, variations of the operating conditions) of the VGH unit is worse than the FCC unit, we can see more minor differences among the applied data-treatment methods. Therefore, only the performance of the MCD method is further shown via the histogram of data points of temperature vs. pressure in the main fractionator in Fig. 8(b). The blue points represent indicated outliers and the rest of the data (green points) is retained for the design of soft sensors (Fig. 9). We can conclude that the marked outliers are mostly measurements deviated from the area with the highest

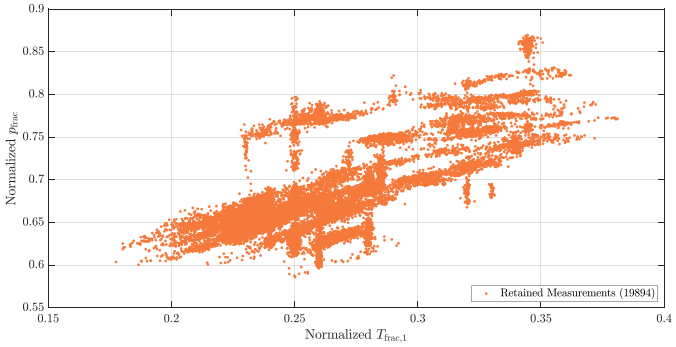


Fig. 9. The retained online measurements (by the MCD method) of the temperature and pressure in the main factionator of the VGH unit.

density of the measurements. This proves the effectiveness of the MCD method to indicate deviated and undesirable measurements.

4.6. Design of inferential sensors for the VGH unit using time series data

We design inferential sensors in the same way as in Section 4.3. Therefore, we distribute (chronologically) first 50% of the available time series data to training set and last 50% of available time series data to the testing set.

Soft-sensors designed by PCA and PLS require thirteen and fifteen principal components, respectively, to explain 98% of the variance in the data. This, on the one hand, suggests possible overfitting yet, on the other hand, there seems to be a good agreement between the advanced design methods on the number of important variables (or their combinations), i.e., 13–15. When designing a soft sensor by the SS methods, similar to the previous use case, we found that the combination with the R^2_{adj} criterion gave the best results. Unlike in the case of the FCC unit, the SS-CV method proposes different sensor structure as the SS method (with R^2_{adj} criterion).

A comparison of the designed sensors in terms of their complexity (n^*_p), accuracy (RMSE), and the amount of bias correction (BC) is shown in Table 3. As we can see, the suggested structure (n^*_p) of the designed soft sensors is much more complicated than the structure of the reference soft sensor. Out of 30 candidate inputs, the designed sensors suggest to include at least eleven more inputs. All the design methods (even OLSR) are able to sparsify to a certain extent the structure of the full sensor (4). Beside PCT_{HGO} included in the reference sensor, LASSO suggests involving $T_{frac,2}$, $T_{frac,1}$, $T_{f,50p}$ and x_{H2} among the most influential variables. On contrary, the SS methods suggest including $T_{frac,2}$, $T_{frac,1}$, PCT_{GF} and $T_{wabt,1}$. Despite the disagreement on the added variables, it appears that certain variables from the reaction section of the plant could play a role in explaining the bad performance of the reference sensor when qualitatively different feedstock is used.

Overall, the accuracy of the designed soft sensors (see RMSE in Table 3) shows the best performance for the soft sensor designed by PCA, good performance of the soft sensors designed by the PLS and LASSO methods and the worst performance of the soft sensors designed by OLSR and SS methods. Apparently, the reference sensor shows high robustness. The poor accuracy of the soft sensor designed by SS methods can be explained by the highly varying operating conditions of the plant. This can also be documented by the much increased amount of bias correction compared to the case of the FCC unit (see in Table 1).

Table 3
Comparison of the number of inputs n^*_p (number of principal components of PCA and PLS), sensor accuracy (RMSE) and bias correction relative frequency (BC) using time series data for the VGH unit.

	OLSR	PCA	PLS	LASSO	SS	SS-CV	Ref
n^*_p	19	24 (13)	22 (15)	14	15	12	1
RMSE	0.184	0.103	0.158	0.145	0.190	0.182	0.114
BC [%]	82.8	85.3	80.2	74.6	82.3	79.3	75.4

We can see that the soft sensor designed by OLSR is much more complicated, less accurate and more frequently corrected than the reference soft sensor. The results show that PCA and PLS methods are not able to reduce the dimensionality of the soft sensor compared to OLSR. The high number of principal components of these methods also suggests that a complex structure is required to express the behavior of the desired variable. The soft sensors designed by PCA, PLS and LASSO are more accurate than other designed soft sensors. Nevertheless, only the PCA sensor is more accurate (by about 10%) than the reference soft sensor. According to the values of the BC criterion in Table 3, the soft sensor designed by PLS is more appropriate than the PCA soft sensor, although both sensors are more frequently corrected than the reference soft sensor. The further values of BC indicate that soft sensors designed by LASSO and SS-CV are corrected less frequently than other designed soft sensors, which results from their simple structure (and implied robustness).

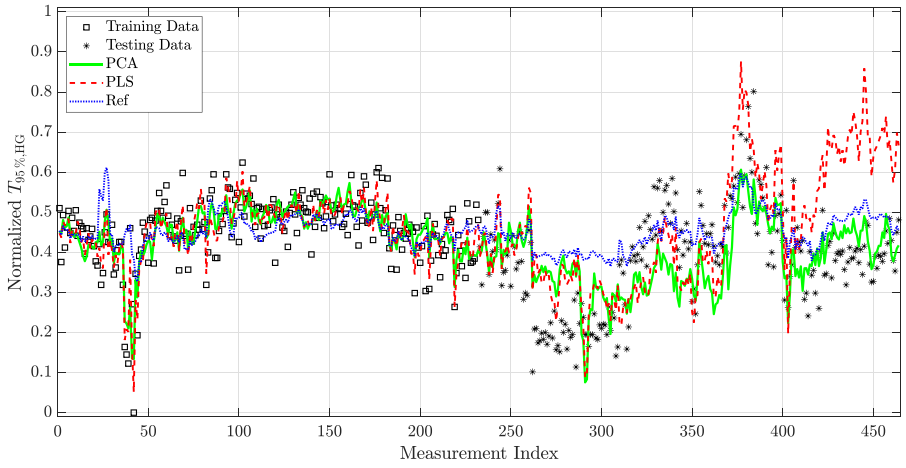
In order to provide more comprehensive comparison of the soft sensors, we visualize their predictive performance on the output variable in Fig. 10 using the same color coding as in the previous case study. The data shows high variability indicating several changes of the operating conditions within the studied time window, in both training and testing datasets. Nonetheless, the variability within the testing set appears to be higher. This might explain the poor performance of the designed advanced sensors and it is confirmed by the high effort of bias correction.

Fig. 10 (a) further presents the training and testing (predictions) performance of the designed advanced soft sensors, by PCA and PLS (Fig. 10(a)) and by LASSO and SS-CV (Fig. 5(b)), compared to the reference soft sensor. We can directly see the training performance of the designed advanced inferential sensors being much better than the reference soft sensor. However, there are several sections in the testing dataset, where these soft sensors are not able to explain the behavior of the output variable. This is most prominent around the measurements 260–320 and 420–464. Interestingly, PCA-based soft-sensor performs relatively well in both the designated periods, which suggests that some process features were successfully caught in the sensor. On the other hand, it exhibits a relatively poor performance around measurement index 350, where it is outperformed by other sensors (even the reference sensor). These observations suggest that the training set is poor and should be expanded.

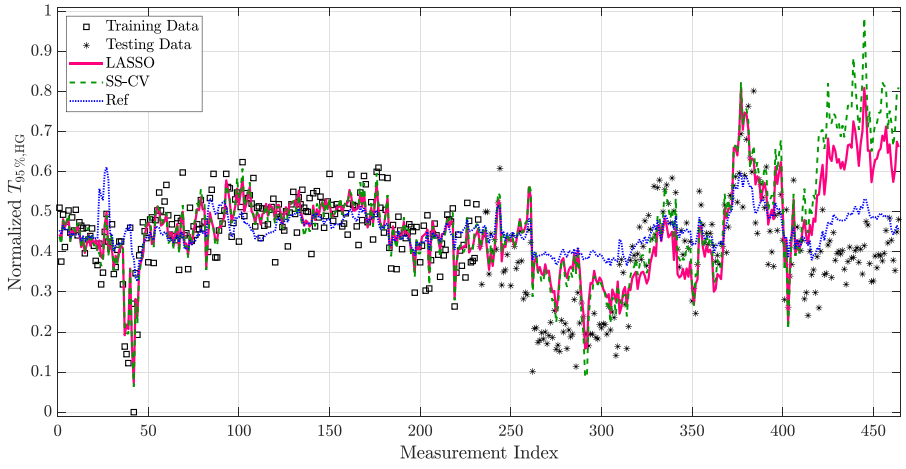
It appears that a practical solution of performing bias update would be sufficient in this situation. We will revisit and analyze this situation in the following section in order to confirm whether the operating conditions change so dramatically that one would need to vary the soft-sensor structure often.

4.7. Design of inferential sensors for the VGH unit using randomly distributed data

Next, we design soft sensors using randomly distributed data. We assign 50% of the available randomly distributed data to the training set and the remaining data to the testing set. We generate 50 such distributions and we use the same training/testing workflow as above. We finally present the average performance measures from the different runs of the corresponding soft-sensor design.



(a) Training and prediction performance of the sensors designed by PCA and PLS methods and reference sensor.



(b) Training and prediction performance of the sensors designed by LASSO and SS-CV methods and reference sensor.

Fig. 10. Comparison of the soft sensors for the VGH unit designed using time series data.

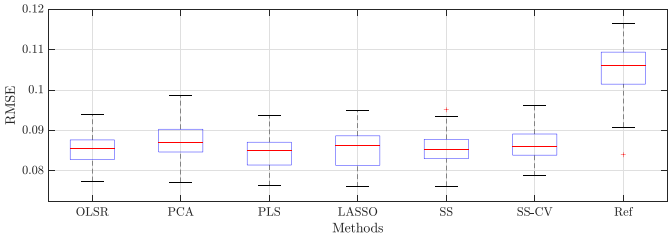
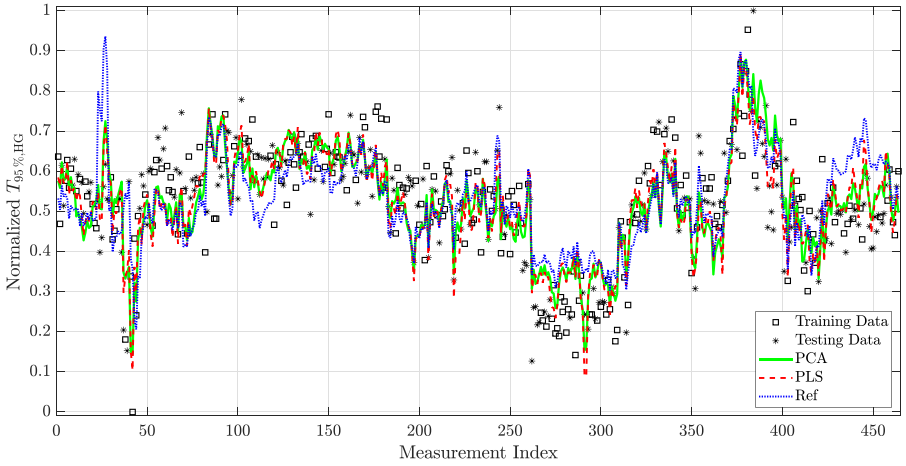


Fig. 11. Comparison of the designed inferential sensor accuracy (RMSE) over 50 different randomly generated training/testing data distributions for the VGH unit.

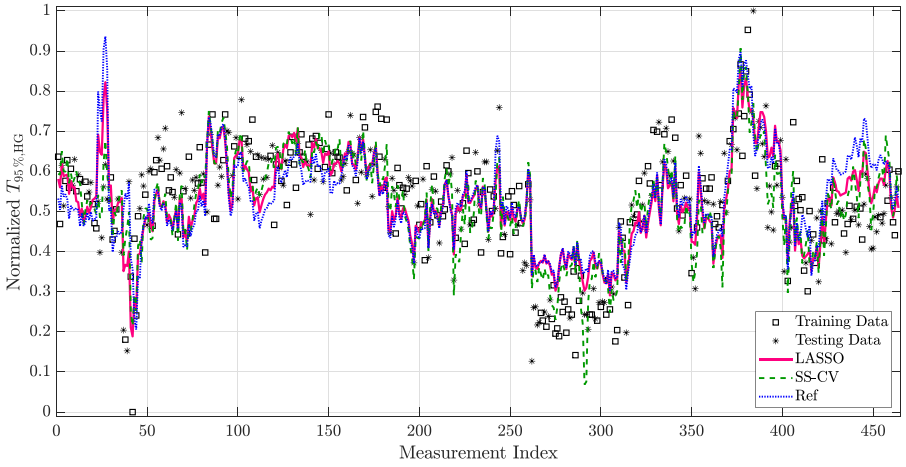
Table 4
Comparison of the number of inputs n_p^* (number of principal components of PCA and PLS), sensor accuracy (RMSE) and bias correction relative frequency (BC) over 50 random training/testing data distributions for the VGH unit.

	OLSR	PCA	PLS	LASSO	SS	SS-CV	Ref
n_p^*	24	25 (15)	25 (17)	15	16	12	1
RMSE	0.086	0.087	0.085	0.086	0.086	0.087	0.105
BC [%]	88.8	86.6	86.6	90.1	89.7	87.1	91.0

The comparison of soft sensors in Table 4 involves the same criteria (n_p^*, RMSE, BC) as in the previous section. In terms of complexity of the designed sensors, we see similar trend as in the FCC use case. The overall complexity of the designed soft sensors is mostly higher compared to the soft sensors designed on chronologically distributed data (see Table 3). This is a recurring observation (from the first case study) and points at the need of enriching the number of explaining variables to adapt for varying plant operating conditions. Only the soft sensor designed by SS-CV is an exception and it even maintains exactly the same sensor structure. These observations reveal that despite SS-CV found good sensor structure in case of chronologically distributed data, the variation in the operating conditions would require to adapt sensor parameters. This also definitely proves high variation of operating conditions and its strong influence on the sensor performance. Similar to SS-CV, also for the rest of the designed sensors the most influential inputs selected by the design methods remain unchanged compared to the case of chronological training/testing data distribution. Each designed soft sensor shows the increased frequency of the bias correction, which can be attributed to the large noise



(a) Training and prediction performance of the sensors designed by PCA and PLS methods and reference sensor.



(b) Training and prediction performance of the sensors designed by LASSO and SS-CV methods and reference sensor.

Fig. 12. Comparison of the soft sensors using randomly generated training/testing data distribution for the VGH unit.

magnitude in the lab data and to the need for adapting the sensor frequently due to operating conditions.

The accuracy of the designed inferential sensors is essentially the same and each sensor is more accurate than the reference sensor. The most accurate sensor is designed by PLS and it improves the accuracy of reference sensor by about 19%. A drop in this performance by PCA-based sensor can be attributed to significant changes in the operating conditions in combination with changes in the sensitivity of the output variable to different inputs (online measurements). The latter claim is supported by the comparatively better performance of the sensor designed by PLS.

Fig. 11 visualizes the accuracy statistics using box plots of each soft sensor from the 50 randomly distributed training/testing datasets. We can see that the performance statistics of all the designed soft sensors mostly copies the conclusions reached in the discussion on the average performance (see Table 4). The results show similar accuracy variance of each soft sensor, which means that the variance is caused mainly by the particular noise realizations in the data. The smallest variance is though achieved for the sensors found by SS methods.

As in the previous section, Fig. 12 visualizes the training and prediction performance of the designed soft sensors for one representative random training/testing data distribution. Results are shown for the reference soft sensor (both plots), the soft sensors designed by PCA and PLS (Fig. 12(a)), and the soft sensors designed by LASSO and SS-CV (Fig. 12(b)). The performance improvement of the soft-sensors with randomly distributed data compared to chronological data is evident. We can observe this on previously mismatched measurements around markers 420–464. Yet, we can clearly identify the period of measurements 260–320 that still exhibits unsatisfactory sensor performance. This calls for another investigation at the plant and revision of the set of candidate sensor inputs.

In conclusion, the advanced design methods show great potential for improving the sensor accuracy beside the good robustness properties of the reference sensor. Yet due to the complexity of the use case, the price to pay for the improved performance is paid in terms of higher sensor complexity. Moreover, due to varying operating conditions, the advanced sensors would need to be often updated or trained on a carefully selected training set.

5. Discussion

Overall, we can say that each studied data treatment method is able to a certain extent indicate outliers in the multivariate data. The advantage of the T^2 distance method is a simple principle. However, this method is strongly affected by the number of treated variables or by the data distribution. The T^2 distance method selects fewer outliers in the FCC unit data (3,567 outliers) than the VGH unit (13,874 outliers). The best results were achieved using the MCD method, which guarantees higher quality of the retained data than the T^2 distance method. The performance of this method seems to be consistent in both case studies. The MCD method indicates 6917 outliers in the FCC unit and 10,023 outliers in the case of the VGH unit. The higher number of indicated outliers in the case of the VGH unit is caused mainly by the worse quality of the data. The treatment of the industrial data pointed out that k-means clustering is quite sensitive to tuning (e.g., number of clusters) that might lead to inferior-quality data treatment. We can see an even more significant discrepancy between the number of indicated outliers in the FCC unit and VGH unit (265 outliers and 11,229 outliers, respectively) by k-means clustering as in the T^2 distance method. It seems that the performance of this method should be adjusted to select more outliers in measurements in the case study on the FCC unit.

The performance of the inferential sensors designed by the studied data-based method (OLSR, PCA, PLS, LASSO, SS and SS-CV) is compared against the reference (current) sensor in both case studies. The reference sensor has a relatively simple structure (three input variables) in the FCC unit and a simple structure (one input variable) in the VGH unit. The low structural complexity provides higher robustness of the inferential sensors. We could see this robustness when the inferential sensors were designed according to the chronological training/testing dataset of the VGH unit. In this case, the designed advanced inferential sensors are more complex yet less accurate than the reference sensor in the final section of the testing dataset. It is most likely that the process deviates from the operating conditions present during the training phase and the advanced sensors would require frequent parameter adaptation to maintain the designed performance.

The results from chronological distribution of training/testing dataset indicate that the inferential sensor designed by PCA achieved the highest accuracy. It outperforms the reference sensor by about 18% in the FCC unit and by about 10% in the VGH unit. Such sensor could be used for plant monitoring. On the other hand, if we also consider the sensor complexity, then the SS-CV method outperforms the rest of the approaches. A low-complexity sensor would be more suitable for optimization or advanced control.

The design of inferential sensors considering both chronologically and randomly distributed training/testing datasets seems to be an effective way to determine the impact of changing operating conditions in the process. The results suggest that the inferential sensors designed over the chronologically distributed training/testing dataset are less sensitive to overfitting than the randomly distributed training/testing dataset. This phenomenon supports the hypothesis of the occurrence of varying operating conditions since the trained sensors tend to involve more inputs to model the changing conditions.

Our investigation has also found that inferential sensors commonly used in the petrochemical industry show high robustness and can give solid performance even long after their commissioning. On the other hand, the relative simplicity of the structure can be easily enhanced in simple cases (the FCC unit use case) by extension of the structure without much maintenance effort. Such sensors can also improve the trust of the operators in the sensors and the automation technology. For this purpose, advanced methods of soft-sensor design (LASSO and SS methods) show a good promise and even the associated computational burden is justified. In more complex cases, the studied design methods can be a promising technology for root-cause analysis.

6. Conclusions

This paper studied soft (inferential) sensors design to monitor unmeasurable variables in the petrochemical industry. Due to the presence of systematic errors and outliers in the industrial measurements, some well-known data pre-treatment methods (T^2 distance, MCD, k-means clustering) were used and compared. The results suggest that MCD is more versatile than T^2 distance or k-means clustering and it performs well overall. Furthermore, the data quality seems to be very well reflected by the number of indicated outliers by the MCD method.

The data retained after the treatment by MCD was subsequently used to design inferential sensors using several data-based methods (OLSR, PCA, PLS, LASSO, SS and SS-CV). The results indicate that PCA tends to design more accurate yet more complex inferential sensors than other methods. On the other hand, the SS-CV method provides well-performing yet structurally less complex sensors than other methods. Therefore, this method is recommended for the design of as simple inferential sensor as possible. A good compromise between the accuracy and complexity is represented LASSO and SS (with overfitting criteria).

The results also indicate that the designed inferential sensors cannot predict the desired variable behaviour over a long time span. There often occur data sections where the designed soft sensors significantly deviate from the measurements of the desired variable. The bias correction seems to be an effective and simple remedy for these discrepancies. In our future work, we will concentrate on finding effective methods of sensor adaptation and/or an efficient way of combining several inferential sensors to cover more operating conditions in the industrial unit.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Martin Mojto: Software, Investigation, Writing – original draft, Visualization. **Karol Ľubušký:** Validation, Investigation, Resources, Formal analysis. **Miroslav Fikar:** Methodology, Validation, Writing – original draft, Writing – review & editing. **Radoslav Paulen:** Conceptualization, Methodology, Validation, Writing – review & editing, Supervision.

Acknowledgments

This research is funded by the Slovak Research and Development Agency under the projects 15-0007, 20-0261, and SK-FR-2019-0004, and by the Scientific Grant Agency of the Slovak Republic under the grant 1/0691/21. MM, MF, and RP would like to thank Dr. András Buti from Slovnaft, a.s in Bratislava for several inspiring discussions and for his fruitful comments on the presented work.

References

Alameddine, I., Kenney, M.A., Gosnell, R.J., Reckhow, K.H., 2010. Robust multivariate outlier detection methods for environmental data. J. Environ. Eng. 136 (11), 1299–1304. doi:10.1061/(ASCE)EE.1943-7870.0000271.

Alves, R.M.B., Nascimento, C.A.O., 2007. Analysis and detection of outliers and systematic errors in industrial plant data. Chem. Eng. Commun. 194 (3), 382–397. doi:10.1080/00986440600899955.

Azzaoui, H., Mansouri, I., Elkihel, B., 2019. Methylcyclohexane continuous distillation column fault detection using stationary wavelet transform and *k*-means. In: Hajji, B., Tina, G.M., Ghoumid, K., Rabhi, A., Mellit, A. (Eds.), Proceedings of the 1st International Conference on Electronic Engineering and Renewable Energy. Springer Singapore, Singapore, pp. 399–411.

Bertsimas, D., King, A., Mazumder, R., 2016. Best subset selection via a modern optimization lens. Ann. Stat. 44 (2), 813–852. doi:10.1214/15-AOS1388.

Chen, L., Bernard, O., Bastin, G., Angelov, P., 2000. Hybrid modelling of biotechnological processes using neural networks. Control Eng. Pract. 8 (7), 821–827. doi:10.1016/S0967-0661(00)00036-8.

Curreri, F., Graziani, S., Xibilia, M.G., 2020. Input selection methods for data-driven soft sensors design: application to an industrial process. Inf. Sci. 537, 1–17.

Dunn, W.J., Scott, D.R., Glen, W.G., 1989. Principal components analysis and partial least squares regression. Tetrahedron Comput. Methodol. 2 (6), 349–376.

Efroymson, M., 1960. Mathematical Methods for Digital Computers. Wiley, New York, NY.

Fontes, C.H., Santos, I.C., Embiruçu, M., Aragão, P., 2021. Pattern reconciliation: a new approach involving constrained clustering of time series. Comput. Chem. Eng. 145, 107169. doi:10.1016/j.compchemeng.2020.107169.

Forgy, E., 1965. Cluster analysis of multivariate data: efficiency versus interpretability of classifications. Biometrics 21, 768–769.

Fortuna, L., Graziani, S., Rizzo, A., & Xibilia, M. G. (2007). Soft Sensors for Monitoring and Control of Industrial Processes. 10.1007/978-1-84628-480-9

Frumosu, F.D., Kulahci, M., 2019. Outliers detection using an iterative strategy for semi-supervised learning. Qual. Reliab. Eng. Int. 35 (5), 1408–1423. doi:10.1002/qre.2522.

Gryzlov, A., Schiferli, W., Mudde, R.F., 2013. Soft-sensors: model-based estimation of inflow in horizontal wells using the extended Kalman filter. Flow Meas. Instrum. 34, 91–104. doi:10.1016/j.flowmeasinst.2013.09.002.

Gurobi Optimization LLC, (2020). Gurobi optimizer reference manual. http://www.gurobi.com.

Hardin, J., Rocke, D.M., 2005. The distribution of robust distances. J. Comput. Graph. Stat. 14 (4), 928–946.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.

Hotelling, H., 1931. The generalization of student's ratio. Ann. Math. Stat. 2 (3), 360–378. doi:10.1214/aoms/1177732979.

Hubert, M., Debruyne, M., 2010. Minimum covariance determinant. WIREs Comput. Stat. 2 (1), 36–43. doi:10.1002/wics.61.

Humod, A., Othman, M., Al-Huseiny, M., Aris, I., Bahari, S., 2020. The efficiency of soft sensors modelling in advanced control systems in oil refinery through the application of hybrid intelligent data mining techniques. J. Phys. 1529, 052049. doi:10.1088/1742-6596/1529/5/052049.

de Jong, S., 1993. Simpls: an alternative approach to partial least squares regression. Chemom. Intell. Lab. Syst. 18 (3), 251–263. doi:10.1016/0169-7439(93)85002-X.

Kadlec, P., Gabrys, B., Strandt, S., 2009. Data-driven soft sensors in the process industry. Comput. Chem. Eng. 33 (4), 795–814. doi:10.1016/j.compchemeng.2008.12.012.

Khatibisepehr, S., Huang, B., Khare, S., 2013. Design of inferential sensors in the process industry: a review of Bayesian methods. J. Process Control 23, 1575–1596.

King, M., 2011. Process Control: A Practical Approach. John Wiley & Sons Ltd..

Kodinariya, T., Makwana, P., 2013. Review on determining of cluster in *k*-means clustering. Int. J. Adv. Res. Comput. Sci. Manag. Stud. 1, 90–95.

Kordon, A., Smits, G., Kalos, A.N., Jordaan, E., 2003. Robust soft sensor development using genetic programming. Data Handl. Sci. Technol. 23, 69–108.

Liu, J., 2010. Developing soft sensors based on data-driven approach. In: International Conference on Technologies and Applications of Artificial intelligence, pp. 150–157. doi:10.1109/TAAI.2010.34.

Löfberg, J., 2004. Yalmip: a toolbox for modeling and optimization in MATLAB. In: Proceedings of the CACSD Conference, Taipei, Taiwan.

Luo, L., Xie, L., Su, H., 2020. Robust mixture Bayesian latent variable regression with structural sparsity and application to inferential sensing of quality variables. Ind. Eng. Chem. Res. 59 (50), 21822–21840. doi:10.1021/acs.iecr.0c03620.

Manenti, F., Cieri, S., Restelli, M., Lima, N., Zuniga Linan, L., Durand, G., et al., 2011. Numerical aspects for the dynamic simulation of the fixed-bed methanol synthesis tubular reactor. In: I-CHEAP-10. ITA, pp. 223–232.

Mejdell, T., Skogestad, S., 1991. Composition estimator in a pilot-plant distillation column using multiple temperatures. Ind. Eng. Chem. Res. 30 (12), 2555–2564.

Mencarelli, L., Pagot, A., Duchêne, P., 2020. Surrogate-based modeling techniques with application to catalytic reforming and isomerization processes. Comput. Chem. Eng. 135, 106772. doi:10.1016/j.compchemeng.2020.106772.

Miyashiro, R., Takano, Y., 2015. Mixed integer second-order cone programming formulations for variable selection in linear regression. Eur. J. Oper. Res. 247, 721–731.

de Morais, G.A.P., Barbosa, B.H.G., Ferreira, D.D., Paiva, L.S., 2019. Soft sensors design in a petrochemical process using an evolutionary algorithm. Measurement 148, 106920. doi:10.1016/j.measurement.2019.106920.

Pearson, K., 1901. LIII. On lines and planes of closest fit to systems of points in space. Lond. Edinb. Dublin Philos. Mag. J. Sci. 2 (11), 559–572.

Qin, S.J., Badgwell, T.A., 2003. A survey of industrial model predictive control technology. Control Eng. Pract. 11, 733–764.

Quelhas, A., 2009. Soft sensor models: bias updating revisited. IFAC Proc. Vol. 42. doi:10.3182/20090712-4-TR-2008.00110.

Rousseeuw, P.J., 1984. Least median of squares regression. J. Am. Stat. Assoc. 79 (388), 871–880. doi:10.1080/01621459.1984.10477105.

Santosa, F., Symes, W.W., 1986. Linear inversion of band-limited reflection seismograms. SIAM J. Sci. Stat. Comput. 7 (4), 1307–1330.

Smith, G., 2018. Step away from stepwise. J. Big Data 5, 32. doi:10.1186/s40537-018-0143-6.

Su, H., Rong, G., Chu, J., 2009. Industrial processes: data reconciliation and gross error detection. Meas. Control 42, 209–215. doi:10.1177/002029400904200704.

Sun, W., Braatz, R.D., 2021. Smart process analytics for predictive modeling. Comput. Chem. Eng. 144, 107–134. doi:10.1016/j.compchemeng.2020.107134.

Takano, Y., Miyashiro, R., 2020. Best subset selection via cross-validation criterion. TOP 28, 475–488.

Tibshirani, R., 2011. Regression shrinkage and selection via the lasso: a retrospective. J. R. Stat. Soc. 73 (3), 273–282. doi:10.1111/j.1467-9868.2011.00771.x.

Torgashov, A., Skogestad, S., 2019. The use of first principles model for evaluation of adaptive soft sensor for multicomponent distillation unit. Chem. Eng. Res. Des. 151, 70–78. doi:10.1016/j.cherd.2019.08.017.

Wold, S., Ruhe, A., Wold, H., Dunn III, W.J., 1984. The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. SIAM J. Sci. Stat. Comput. 5 (3), 735–743.

Wold, S., Sjöström, M., Eriksson, L., 2001. PLS-regression: a basic tool of chemometrics. Chemom. Intell. Lab. Syst. 58 (2), 109–130. doi:10.1016/S0169-7439(01)00155-1.

Xenos, D., Cicciotti, M., Bouaswaig, A., Ricardo, M.-B., Manenti, F., & Thornhill, N. (2014). Simultaneous nonlinear reconciliation and update of parameters for online use of first-principles models: an industrial case-study on compressors. (vol. 33). 10.1016/B978-0-444-63456-6.50077-6

Xu, S., Lu, B., Bell, N., Nixon, M., 2017. Outlier detection in dynamic systems with multiple operating points and application to improve industrial flare monitoring. Processes 5 (2). doi:10.3390/pr5020028.

Yu, Y., Peng, M.-j., Wang, H., Ma, Z.-g., Li, W., 2020. Improved PCA model for multiple fault detection, isolation and reconstruction of sensors in nuclear power plant. Ann. Nucl. Energy 148, 107662. doi:10.1016/j.anucene.2020.107662.

Yuan, X., Ye, L., Bao, L., Ge, Z., Song, Z., 2015. Nonlinear feature extraction for soft sensor modeling based on weighted probabilistic PCA. Chemom. Intell. Lab. Syst. 147, 167–175. doi:10.1016/j.chemolab.2015.08.014.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. R. Stat. Soc. 67 (2), 301–320. doi:10.1111/j.1467-9868.2005.00503.x.

IMA Journal of Mathematical Control and Information (2016) **33**, 563–587
doi:10.1093/imamci/dnu055
Advance Access publication on January 27, 2015

Guaranteed parameter estimation of non-linear dynamic systems using high-order bounding techniques with domain and CPU-time reduction strategies

Radoslav Paulen
Process Dynamics and Operations Group, Department of Biochemical and Chemical Engineering, Technische Universität Dortmund, Emil-Figge-Str. 70, 44221 Dortmund, Germany

and

Mario E. Villanueva and Benoît Chachuat*
Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK
*Corresponding author: b.chachuat@imperial.ac.uk

[Received on 16 March 2014; revised on 19 July 2014; accepted on 6 December 2014]

This paper is concerned with guaranteed parameter estimation of non-linear dynamic systems in a context of bounded measurement error. The problem consists of finding—or approximating as closely as possible—the set of all possible parameter values such that the predicted values of certain outputs match their corresponding measurements within prescribed error bounds. A set-inversion algorithm is applied, whereby the parameter set is successively partitioned into smaller boxes and exclusion tests are performed to eliminate some of these boxes, until a given threshold on the approximation level is met. Such exclusion tests rely on the ability to bound the solution set of the dynamic system for a finite parameter subset, and the tightness of these bounds is therefore paramount; equally important in practice is the time required to compute the bounds, thereby defining a trade-off. In this paper, we investigate such a trade-off by comparing various bounding techniques based on Taylor models with either interval or ellipsoidal bounds as their remainder terms. We also investigate the use of optimization-based domain reduction techniques in order to enhance the convergence speed of the set-inversion algorithm, and we implement simple strategies that avoid recomputing Taylor models or reduce their expansion orders wherever possible. Case studies of various complexities are presented, which show that these improvements using Taylor-based bounding techniques can significantly reduce the computational burden, both in terms of iteration count and CPU time.

Keywords: parameter estimation; dynamic systems; bounded-error estimation; measurement noise; Taylor models; polyhedral relaxations; domain reduction.

1. Introduction

Mathematical modelling has become an integral part of modern process design methodologies as well as in control system design and operations optimization. A typical model development procedure is divided into two main phases, namely specification of the model structure and estimation of the unknown/uncertain model parameters. The latter phase, often referred to as model fitting, normally proceeds by determining parameter values for which the model predictions closely match the available

process measurements. Failure to find an acceptable agreement calls for a revision of the model structure, before repeating the parameter estimation.

Most commonly, the parameter estimation problem is posed as an optimization problem that determines the parameter values minimizing the gap between the measurements and the model predictions, for instance in the least-square sense. Nonetheless, several factors can impair a successful and reliable estimation procedure. First of all, structural model mismatch is inherent to the modelling exercise, and it is illusive to look for the 'true' parameter values in this context. Even in the absence of model mismatch, fitting a set of experimental data exactly is generally not possible due to various sources of uncertainty. A measurement's accuracy is always tied to the resolution of the corresponding apparatus. Moreover, measured data are typically corrupted with noise, for instance Gaussian white noise or more generally coloured noise.

Among the available approaches to account for uncertainty in parameter estimation, the focus in this paper is on *guaranteed* parameter estimation (Walter, 1990), namely the determination of *all* parameter values—referred to as the solution set subsequently—that are consistent with the measurements under given uncertainty scenarios. Specifically, we consider the case that the uncertainty enters the estimation problem in the form of bounded measurement errors. An inherent advantage of this approach over more traditional parameter estimation is that no (consistent) solution to the problem will be lost, and this can help detect problems arising due to lack of identifiability. Moreover, the estimation process does not rely on a particular statistical description of the uncertainty, as is typically the case when applying maximum likelihood or Bayesian techniques. On the downside nonetheless, performing guaranteed parameter estimation turns out to be a very challenging and demanding task from a computational standpoint.

In non-linear algebraic models, the problem of approximating the solution set by a box partition, at an arbitrary precision, has been shown to be tractable using exhaustive search and interval analysis (Moore, 1992), for instance based on the Set Inversion Via Interval Analysis (SIVIA) algorithm (Jaulin & Walter, 1993). This approach has been later extended to dynamic systems using ODE bounding techniques (e.g., Jaulin, 2002; Raissi *et al.*, 2004). In a recent paper, Kieffer & Walter (2011) have identified the main computational bottlenecks of set-inversion algorithms for guaranteed parameter estimation in dynamic systems to be: (i) the need for tight bounds on the solutions of the dynamic system; and (ii) the need for efficient domain-reduction strategies as part of the exclusion tests. It is the objective of this paper to investigate strategies that can enhance the convergence speed of these algorithms, with special emphasis on higher-order ODE bounding techniques and domain-reduction techniques.

The computation of exact bounds on the solution set of non-linear parametric ODEs belongs to the class of computationally intensive problems (non-convex optimization). In response to this, approximate methods that overestimate the solution set of parametric ODEs, yet provide sufficiently tight bounds, have been developed over the years. These methods can be classified as discrete or continuous methods according to the way the enclosures are propagated through time. *Discrete-time methods* proceed by discretizing the integration horizon into a finite number of steps, whereby each step consists of two phases. Phase I is concerned with the computation of a coarse enclosure and a step-size for which existence and uniqueness of the solutions can be established; then, the enclosure is refined at the end of each time-step during Phase II. Recently, Houska *et al.* (2013) also proposed an algorithm reverting the order of the two phases. The types of enclosures that can be propagated with discrete-time methods include intervals (Lohner, 1992; Nedialkov *et al.*, 1999; Rauh *et al.*, 2006), Taylor models with interval remainders (Berz & Makino, 1998; Neher *et al.*, 2007; Lin & Stadtherr, 2007b; Sahlodin & Chachuat, 2011) and Taylor models with ellipsoidal remainders (Houska *et al.*, 2013). *Continuous-time methods*, on

the other hand, involve formulating a set of (parameter-independent) auxiliary ODEs, whose solutions enclose those of the original dynamic model. Similar to discrete-time approaches, the types of enclosures that can be propagated in continuous time include intervals based on the classical theory of differential inequalities (Walter, 1970) as well as ellipsoids using ellipsoidal calculus (Kurzhanski & Varaiya, 2002; Houska *et al.*, 2012). Extensions of these approaches have recently been proposed to enable the propagation of Taylor models with either interval remainders (Chachuat & Villanueva, 2012) or ellipsoidal remainders (Villanueva *et al.*, 2013). See also (Villanueva *et al.*, 2014) for a unified framework and convergence analysis of continuous-time bounding techniques for non-linear parametric ODEs.

In the context of guaranteed parameter estimation, the use of ODE bounding techniques based on Taylor models has been investigated by Lin & Stadtherr (2007a) and Kletting *et al.* (2011) using discrete-time bounding techniques and, more recently, by Paulen *et al.* (2013) using a continuous-time approach. These authors have reported significant improvements in the convergence speed of the set-inversion algorithm compared with classical approaches based on interval enclosures. In principle, the higher the Taylor expansion order of the ODE solutions with respect to the uncertain parameters, the smaller the number of iterations required by the set-inversion algorithm to converge. Nonetheless, a higher-order expansion can incur a significant computational overhead, thereby defining a trade-off in terms of the overall computational burden with regards to the expansion order. This trade-off is investigated further in this paper through the comparison of continuous-time methods propagating Taylor models (with either interval or ellipsoidal remainders) against simple interval box propagation.

Another approach to enhancing the convergence involves applying contractors to the parameter boxes in order to reduce their width. Contractors based on optimality tests were derived by Jaulin *et al.* (2001) using interval analysis, and later applied to dynamic system, e.g., by Kieffer & Walter (2011). Besides enabling higher-order convergence, the use of Taylor models to enclose the ODE solutions provides an explicit representation of parameter dependencies via the multivariate polynomial part. Lin & Stadtherr (2007a) and Kletting *et al.* (2011) took advantage of this representation and used a constraint-propagation strategy in order to contract the parameter boxes. Inspired by developments in the field of global optimization (Zamora & Grossmann, 1999; Neumaier, 2004; Tawarmalani & Sahinidis, 2004; Belotti *et al.*, 2009) and their recent extension to global dynamic optimization (Sahlodin, 2012), this paper investigates a domain-reduction technique that solves linear programs (LPs) constructed from the polyhedral relaxation of Taylor models of the predicted outputs as a means to exclude those parameter subsets whose corresponding response does not intersect with the measurement bounds. In order to further reduce the computational burden, we also investigate new strategies that avoid recomputing Taylor models of the predicted outputs or reduce their order as soon as the corresponding overestimation is within a given threshold.

The rest of the paper is organized as follows. In Section 2, the problem of guaranteed parameter estimation is defined mathematically and the set-inversion algorithm is concisely stated. Section 3 presents a new methodology for enhancing the convergence of set inversion in a guaranteed parameter estimation context, which relies on Taylor-model bounding in combination with domain-reduction and CPU-time-reduction strategies; a simple case study is carried out through this section to illustrate the developments. Then, Section 4 presents the case study of a more challenging model of anaerobic digestion with complex dynamics and multiple time-scales, demonstrating that the proposed improvements allow tackling guaranteed parameter estimation in up to seven parameters within reasonable computational times. Finally, Section 5 concludes the paper.

2. Guaranteed parameter estimation

2.1 *Problem statement*

Consider a dynamic process described by parametric ODEs of the form

$$\dot{x}(t,p) = f(x(t,p),p) \quad \text{with } x(0,p) = h(p), \tag{2.1a}$$

$$y(t,p) = g(x(t,p),p), \tag{2.1b}$$

where $x : [0,t_N] \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_x}$ denotes the vector of process states, $p \in R^{n_p}$ stands for the vector of (unknown) process parameters and $y : [0,t_N] \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_y}$ denotes the n_y-dimensional vector of model outputs (predictions). Note the parametric dependencies in the right-hand side function f, the output function g and the initial value function h. In particular, this latter dependency can be used to handle dynamic systems with uncertain initial conditions. Given a bounded subset $P \subseteq \mathbb{R}^{n_p}$ for the parameters, we introduce the point-wise-in-time reachable sets of (2.1a) and (2.1b) as

$$\forall t \in [0,t_N], \quad X(t,P) := \{x(t,p) \mid p \in P \subseteq \mathbb{R}^{n_p}\} \quad \text{and} \quad Y(t,P) := \{y(t,p) \mid p \in P \subseteq \mathbb{R}^{n_p}\}. \tag{2.2}$$

For a given set of output measurements $y_{\mathrm{m}}(t_i)$ at N time points $t_1, \ldots, t_i, \ldots, t_N$, *classical* parameter estimation seeks for *one* particular instance p_{e} of the parameter values for which the (possibly weighted) normed difference between these measurements and the corresponding model outputs y is minimized. This optimization problem, for instance in the least-square sense, is given by:

$$p_{\mathrm{e}} \in \arg\min_{p \in P_0} \quad \sum_{i=1}^{N} \|y_{\mathrm{m}}(t_i) - y(t_i,p)\|_2^2, \tag{2.3a}$$

$$\text{s.t.} \quad \dot{x}(t,p) = f(x(t,p),p) \quad \text{with } x(0,p) = h(p), \tag{2.3b}$$

$$y(t,p) = g(x(t,p),p), \tag{2.3c}$$

where the interval box $P_0 := [p_0^{\mathrm{L}}, p_0^{\mathrm{U}}]$ denotes the a priori set of admissible values for the parameters. The superscripts $^{\mathrm{L}}$ and $^{\mathrm{U}}$ representing the lower and upper bounds of an interval box are understood component-wise throughout.

In contrast, *guaranteed* (bounded-error) parameter estimation accounts for the fact that the actual process outputs, y_{p}, are only known within some bounded measurement error $e \in E := [e^{\mathrm{L}}, e^{\mathrm{U}}]$, so that

$$y_{\mathrm{p}}(t_i) \in y_{\mathrm{m}}(t_i) + [e^{\mathrm{L}}, e^{\mathrm{U}}] =: Y_{\mathrm{p}}(t_i). \tag{2.4}$$

Then, the main objective is to estimate the set P_{e} of *all* possible parameter values p such that $y(t_i;p) \in Y_{\mathrm{p}}(t_i)$ for every $i = 1, \ldots, N$; that is,

$$P_{\mathrm{e}} := \left\{ p \in P_0 \;\middle|\; \begin{array}{l} \exists x \text{ such that:} \\ \dot{x}(t,p) = f(x(t,p),p) \quad \text{with } x(0,p) = h(p), \\ g(x(t_i,p),p) \in Y_{\mathrm{p}}(t_i), \quad i = 1, \ldots, N \end{array} \right\}. \tag{2.5}$$

Depicted in red on the left plot in Fig. 1 is the set of all output trajectories satisfying $y(t_i,p) \in Y_{\mathrm{p}}(t_i)$ with $i = 1, \ldots, N$, and on the right plot the corresponding set P_{e} projected onto the (p_1, p_2) space. Obtaining an exact characterization of the set P_{e} is not possible in general, and one has to resort to

approximation techniques that make the problem computationally tractable. The focus in the remainder of the paper is on algorithms approximating P_e using set-inversion techniques.

2.2 *Set-inversion algorithm*

We consider a variant of the SIVIA algorithm by Jaulin & Walter (1993) in order to approximate the solution set P_e to a desired accuracy. This algorithm has already been exploited in a number of papers in the context of dynamic parameter estimation (e.g., Jaulin, 2002; Raissi *et al.*, 2004; Lin & Stadtherr, 2007a; Kieffer & Walter, 2011; Kletting *et al.*, 2011; Paulen *et al.*, 2013).

Let $Y^{-1}(t,\cdot)$, with $Y^{-1}:[0,t_N]\times\mathbb{IR}^{n_y}\to\mathbb{IR}^{n_p}$, denote the inverse of the reachable set mapping $Y(t,\cdot)$ defined earlier in (2.2). It follows that characterizing P_e via (2.5) is equivalent to intersecting the inverse image sets $Y^{-1}(t_i,Y_p(t_i))$ for each $i=1,\ldots,N$:

$$P_e=\left(\bigcap_{i=1}^{N}Y^{-1}(t_i,Y_p(t_i))\right)\cap P_0. \tag{2.6}$$

A prototypical set-inversion algorithm based on exhaustive search that uses this property is as follows:

Input: Termination tolerances $\varepsilon_{box}\geqslant 0$ and $\varepsilon_{bnd}\geqslant 0$

Initialization: Set partitions $\mathbb{P}_{bnd}=\{P_0\}$, $\mathbb{P}_{int}=\emptyset$, and $\mathbb{P}_{out}=\emptyset$; Set iteration counter $k=0$

Main Loop:

1. Select a parameter box P in the partition $\mathbb{P}_{bnd}$ and remove it from $\mathbb{P}_{bnd}$
2. Compute enclosures $\overline{Y}(t_i,P)\supseteq Y(t_i,P)$, for each $i=1,\ldots,N$
3. Exclusion Tests:
   (a) **If** $\overline{Y}(t_i,P)\subseteq Y_p(t_i)$ for all $i\in\{1,\ldots,N\}$, insert P into $\mathbb{P}_{int}$
   (b) **Else if** $\overline{Y}(t_i,P)\cap Y_p(t_i)=\emptyset$ for some $i\in\{1,\ldots,N\}$, insert P into $\mathbb{P}_{out}$
   (c) **Else** bisect P and insert subsets back into $\mathbb{P}_{bnd}$
4. Termination Tests:
   (a) **If** $V_{bnd}:=\sum_{P\in\mathbb{P}_{bnd}}\text{volume}(P)\leqslant\varepsilon_{bnd}$, **stop**
   (b) **If** $\text{width}(P)\leqslant\varepsilon_{box}$ for all $P\in\mathbb{P}_{bnd}$, **stop**
5. Increment counter k+=1; **Return** to step 1

Output: Partitions $\mathbb{P}_{int}$, $\mathbb{P}_{bnd}$, and $\mathbb{P}_{out}$; Iteration count k

---

An illustration of a parameter box belonging to the partition $\mathbb{P}_{int}$, $\mathbb{P}_{bnd}$ or $\mathbb{P}_{out}$ is shown on the right plot in Fig. 1, together with the corresponding output trajectories on the left plot using a consistent

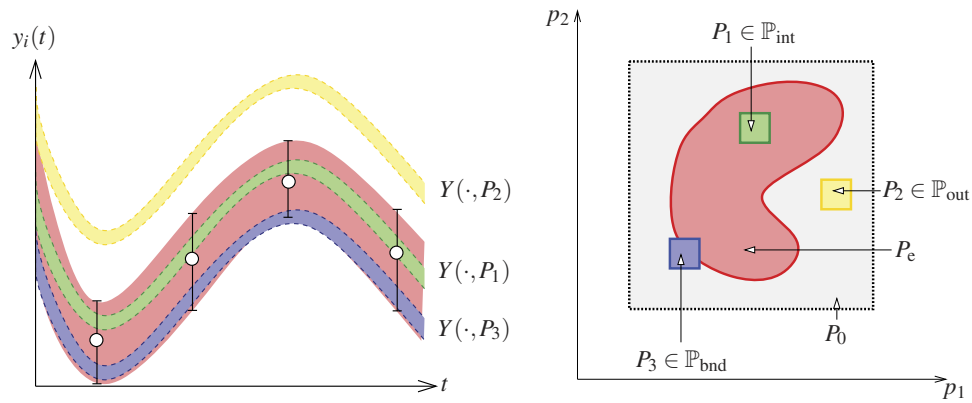


FIG. 1. Illustration of guaranteed parameter estimation concepts in the space of output trajectories (left plot) and in the parameter space (right plot).

colour scheme. Upon termination, this algorithm returns partitions $\mathbb{P}_{\text{int}}$ and $\mathbb{P}_{\text{bnd}}$ such that

$$\bigcup_{P\in\mathbb{P}_{\text{int}}} P \subseteq P_e \subseteq \bigcup_{P\in\mathbb{P}_{\text{int}}\cup\mathbb{P}_{\text{bnd}}} P. \tag{2.7}$$

The following remarks are in order regarding the set-inversion algorithm:

- Multiple heuristics can be used regarding the selection and the bisection of a parameter box in steps 1 and 3c, respectively. In order for the search to be exhaustive, one can select a parameter box that has the largest width in priority and apply bisection at the mid-point along the least reduced axis of a box for instance.
- Step 2 calls for a procedure capable of computing an enclosure of the output reachable set $Y(\cdot,P)$ for the current parameter box P. The main difficulty of this bounding step lies in the computation of a point-wise-in-time enclosure $\overline{X}(t,P)$ of the state reachable set $X(t,P)$, after which an enclosure $\overline{Y}(t,P) \supseteq Y(t,P)$ can be computed readily by using standard interval analysis (Jaulin *et al.*, 2001; Moore *et al.*, 2009). For instance, Jaulin (2002) first used the theory of differential inequalities, which provides a rule for propagating an interval enclosure $\overline{X}(t,P) := [x^{\text{L}}(t), x^{\text{U}}(t)] \in \mathbb{IR}^{n_x}$ of the reachable set $X(t,P)$ in the form of auxiliary ODEs:

$$\dot{x}_i^{\text{L}}(t,P) = \min_{\xi,\rho}\left\{ f_i(\xi,\rho) \;\middle|\; \begin{array}{l} \xi_i = x_i^{\text{L}}(t) \\ \xi \in [x^{\text{L}}(t), x^{\text{U}}(t)] \\ \rho \in P \end{array} \right\} \quad \text{with } x_i^{\text{L}}(0,P) = \min_{\rho}\{h(\rho) \mid \rho \in P\}, \tag{2.8}$$

$$\dot{x}_i^{\text{U}}(t,P) = \max_{\xi,\rho}\left\{ f_i(\xi,\rho) \;\middle|\; \begin{array}{l} \xi_i = x_i^{\text{U}}(t) \\ \xi \in [x^{\text{L}}(t), x^{\text{U}}(t)] \\ \rho \in P \end{array} \right\} \quad \text{with } x_i^{\text{U}}(0,P) = \max_{\rho}\{h(\rho) \mid \rho \in P\}, \tag{2.9}$$

for each $i \in \{1,\ldots,n_x\}$. In principle, any ODE bounding technique can be used for this step as long as the computed output enclosures $\overline{Y}(t_i,P)$ shrink when the diameter of the parameter host set

$\text{diam}(P) := \max_{x,y \in P} |x - y| \to 0$ in order to guarantee finite termination of the algorithm. This is the case in particular for the higher-order bounding techniques considered in Section 3.1.

- Test 4a is an addition to the original SIVIA algorithm (Jaulin & Walter, 1993), which interrupts the iterations when a specified level of approximation of the solution set P_e is reached. The level of approximation is measured here as the total volume V_{bnd} of the boxes in the partition, with corresponding threshold ε_{bnd}. In contrast, stopping the algorithm when a minimum width is reached for all the boxes in $\mathbb{P}_{bnd}$ (Test 2.2) does not give any guarantee on the actual approximation level of the solution set boundary because of the overestimation in step 2. We also note that finite termination of the algorithm requires that either $\varepsilon_{box} > 0$ or $\varepsilon_{bnd} > 0$.

Variants of this basic algorithm exist that improve the convergence speed by introducing additional exclusion tests. One such test involves checking whether (an enclosure of) the gradient of the objective function in (2.3a) for a given box P does not contain 0, in which case $P \in \mathbb{P}_{out}$ (Kieffer & Walter, 2011)—this is because there cannot exist any global optimizer of (2.3a) in P in this case. The downside of these strategies is the need to compute bounds on the first-order sensitivities (or adjoints) of model (2.1), which can cause a significant computational overhead. Reduction of the parameter boxes was also investigated, e.g., by Lin & Stadtherr (2007a) and Kletting *et al.* (2011) using constraint propagation on Taylor models of the model outputs. The following section describes a new methodology to enhancing the convergence of set-inversion in a guaranteed parameter estimation context and be in a position to tackle more challenging, larger-scale problems.

3. Guaranteed parameter estimation methodology using Taylor models

Kieffer & Walter (2011) have argued that the main computational bottlenecks of the set-inversion algorithm in Section 2.2 for guaranteed parameter estimation are the need for tight bounds on the solutions of the dynamic system as well as efficient domain-reduction strategies supporting exclusion tests. The methodology developed through this section aims precisely at addressing these needs. It relies on higher-order techniques based on Taylor models to bound the dynamics (Section 3.1), and then takes full advantage of the resulting Taylor model estimators for driving optimization-based domain reduction (Section 3.2). Moreover, special strategies are developed that avoid recomputing Taylor models or reduce their expansion orders wherever possible.

3.1 *Higher-order bounding strategy*

The reachable set of a non-linear ODE is a non-convex set in general, and enclosing it within an interval box can lead to significant overestimation due to both the wrapping effect and the dependency problem. One way of propagation non-convex enclosures $\overline{X}(t,P)$ of the reachable set $X(t,P)$ involves using a qth-order multivariate polynomial $\mathscr{P}_x^q(t,\cdot)$, whose image set on P approximates $X(t,P)$, as

$$\overline{X}(t,P) := \{\mathscr{P}_x^q(t,p) \mid p \in P\} \oplus \mathscr{R}_x^q(t,P), \tag{3.1}$$

where $\mathscr{R}_x^q(t,P) \supseteq \{x(t,p) - \mathscr{P}_x^q(t,p \mid p \in P\}$ is the so-called remainder term bounding the approximation error on P. In turn, an interval enclosure can be derived from (3.1) by bounding the polynomial part, for which multiple approaches have been proposed in the literature (Lin & Rokne, 1995; Neumaier, 2002). For simplicity, the approach used in this work considers exact bounding of the linear and diagonal quadratic terms, while overestimating the remaining terms using natural interval extensions (Lin & Stadtherr, 2007b).

The focus hereafter is on Taylor models (Makino & Berz, 1999; Neumaier, 2002; Bompadre *et al.*, 2013), although alternative types of polynomial approximation can be used in principle as long as these constructions can be automated for general factorable functions. In this approach, the polynomial approximant $\mathscr{P}_x^q(t,\cdot)$ matches the qth-order Taylor expansion of $x(t,\cdot)$ on P at a given reference point $\hat{p} \in P$:

$$\forall p \in P, \quad \mathscr{P}_x^q(t,p) := \sum_{\substack{\gamma \in \mathbb{N}^{n_p}, \\ |\gamma| \leqslant q}} \frac{\partial^\gamma x(t,\hat{p})}{\gamma!}(p-\hat{p})^\gamma, \tag{3.2}$$

where multi-index notation is used and $\partial^\gamma x_i(\cdot,\hat{p})$ denotes the state sensitivities $(\partial^{|\gamma|}x_i/\partial p_1^{\gamma_1}\ldots\partial p_n^{\gamma_n})(\cdot,\hat{p})$ at $\hat{p}$. Note that this construction requires that the right-hand side function f and initial-value function h are at least $(q+1)$-times continuously differentiable in all their arguments. Moreover, it requires that a system of state-sensitivity equations of size $O(n_x n_p^q)$ is integrated on the time horizon.

Of the alternatives to compute the point-wise-in-time convex remainder enclosure $\mathscr{R}_x^q(t,P)$, our focus in this paper is on interval and ellipsoidal enclosures, which are summarized below.

*Taylor models with interval remainder bounds.*The method of differential inequalities can be applied to propagate an interval enclosure $\mathscr{R}_x^q(t,P) := [r_x^{\mathrm{L}}(t), r_x^{\mathrm{U}}(t)]$ of the qth-order remainder term by integrating, together with the sensitivity equations for $\partial^\gamma x_i(\cdot,\hat{p})$ with $|\gamma| \leqslant q$, the following auxiliary ODEs:

$$\dot{r}_i^{\mathrm{L}}(t) = \min_{\xi,\rho} \left\{ f_i(\mathscr{P}_x^q(t,\rho)+\xi,\rho) - \dot{\mathscr{P}}_x^q(t,\rho) \;\middle|\; \begin{array}{l} \xi_i = r_{x_i}^{q,\mathrm{L}}(t) \\ \xi \in [r_x^{\mathrm{L}}(t), r_x^{\mathrm{U}}(t)] \\ \rho \in P \end{array} \right\} \tag{3.3}$$

$$\text{with } r_i^{\mathrm{L}}(0) = \min_p \{h(p) - \mathscr{P}_h^q(p) \mid p \in P\},$$

$$\dot{r}_i^{\mathrm{U}}(t) = \max_{\xi,\rho} \left\{ f_i(\mathscr{P}_x^q(t,\rho)+\xi,\rho) - \dot{\mathscr{P}}_x^q(t,\rho) \;\middle|\; \begin{array}{l} \xi_i = r_{x_i}^{q,\mathrm{U}}(t) \\ \xi \in [r_x^{\mathrm{L}}(t), r_x^{\mathrm{U}}(t)] \\ \rho \in P \end{array} \right\} \tag{3.4}$$

$$\text{with } r_i^{\mathrm{U}}(0) = \max_p \{h(p) - \mathscr{P}_h^q(p) \mid p \in P\},$$

for each $i \in \{1,\ldots,n_x\}$, with $\mathscr{P}_h^q$ denoting the multivariate polynomial in the Taylor expansion of the initial value function h on P at $\hat{p}$. The resulting enclosures (3.1) enjoy $(q+1)$th-order convergence to the actual reachable set $X(t,P)$, but the size of the auxiliary bounding system scales as $O(n_x n_p^q)$.

*Taylor models with ellipsoidal remainder bounds.*Likewise, ellipsoidal calculus provides a means of propagating an ellipsoidal enclosure $\mathscr{E}(Q_x^q(t)) := \{Q_x^q(t)^{\frac{1}{2}}v \mid \forall v \in \mathbb{R}^{n_x} : v^{\mathrm{T}}v \leqslant 1\}$ of the qth-order remainder term by integrating, together with the parametric sensitivity equations up to order q, the following

auxiliary ODEs:

$$\begin{aligned}\dot{Q}_x^q(t) = &\left(\frac{\partial f}{\partial x}(\mathscr{P}_x^q(t,\hat{p}),\hat{p})\right) Q_x^q(t) + Q_x^q(t)\left(\frac{\partial f}{\partial x}(\mathscr{P}_x^q(t,\hat{p}),\hat{p})\right)^{\mathrm{T}} \\ &+ \sum_{i=1}^{n_x} \kappa_i(t) Q_x^q(t) + \operatorname{diag}(\kappa(t))^{-1} \operatorname{diag}\operatorname{rad}(\Omega_f^q[Q_x^q(t),P,\hat{p}])^2 \\ &\text{with } Q_x^q(0) = \operatorname{diag}\operatorname{rad}(\Omega_h^q[P,\hat{p}])^2. \end{aligned} \tag{3.5}$$

The non-linearity bounders $\Omega_f^q[Q(t),P,\hat{p}], \Omega_h^q[P,\hat{p}] \in \mathbb{IR}^{n_x}$ must satisfy

$$\forall (r,\rho) \in \mathscr{E}(Q) \times P, \quad f(\mathscr{P}_x^q(t,\rho)+r,\rho) - \dot{\mathscr{P}}_x^q(t,\rho) - \frac{\partial f}{\partial x}(\mathscr{P}_x^q(t,\rho),\rho) r \in \Omega_f^q[Q,P,\hat{p}], \tag{3.6}$$

and

$$\forall \rho \in P, \quad h(\rho) - \mathscr{P}_h^q(\rho) \in \Omega_h^q[P,\hat{p}], \tag{3.7}$$

and they can be constructed, at a given time t, on application of interval analysis for instance. Moreover, the scaling function κ can be chosen in such a way as to minimize $\operatorname{tr}(Q_x^q(t))$. The resulting enclosures $\{\mathscr{P}_x^q(t,p) \mid p \in P\} \oplus \mathscr{E}(Q_x^q(t))$ enjoy $(q+1)$th-order convergence to the actual reachable set $X(t,P)$, now at the price of solving an auxiliary bounding system of size $O(n_x n_p^q + n_x^2)$.

The reader is referred to Villanueva *et al.* (2014) for more details about the theory and implementation of these methods. A comparison in the context of guaranteed parameter estimation is presented next for a simple case study.

Case study. Consider the following dynamic model involving two state variables $x = (x_1, x_2)^{\mathrm{T}}$ and three uncertain parameters $p = (p_1, p_2, p_3)^{\mathrm{T}} \in [0.01, 1]^3$ (Kieffer & Walter, 2011):

$$\dot{x}_1(t) = -(p_1 + p_3) x_1(t) + p_2 x_2(t) \quad \text{with } x_1(0) = 1, \tag{3.8a}$$

$$\dot{x}_2(t) = p_1 x_1(t) - p_2 x_2(t) \quad \text{with } x_2(0) = 0. \tag{3.8b}$$

This system has a single output variable y, which corresponds to the state variable x_2, $y(t,p) := x_2(t,p)$, with $N = 15$ measurements corresponding to the time instants $t_i = 1, \ldots, 15$. Synthetic experimental data are generated by simulating the model (3.8) with parameter values $p^* = (0.6, 0.15, 0.35)^{\mathrm{T}}$, and then rounding the output $y(t_i)$ up or down to the nearest value by retaining two significant digits only; then, measurement error ranges of $\pm 5 \times 10^{-3}$ are added around these values.

The guaranteed parameter estimation algorithm (Section 2.2) is implemented in a C++ program that uses the library MC++ (http://projects.coin-or.org/MCpp) for computations involving Taylor models. Moreover, the code calls the ODE integration methods in the GNU Scientific Library to bound the parametric ODEs based on the techniques outlined in Section 3.1. All the numerical results presented subsequently use the explicit embedded Runge–Kutta–Fehlberg (4,5) method, with both relative and absolute tolerances set to 10^{-7}, and are obtained on a workstation with Intel Core i7-3770 processors at 3.40 GHz and running 64-bit Linux.

The performance of guaranteed parameter estimation is investigated for continuous-time ODE bounding techniques propagating Taylor models of orders $q = 1, \ldots, 4$ with interval or ellipsoidal remainders (Section 3.1) and compared with standard differential inequalities. To allow for fair

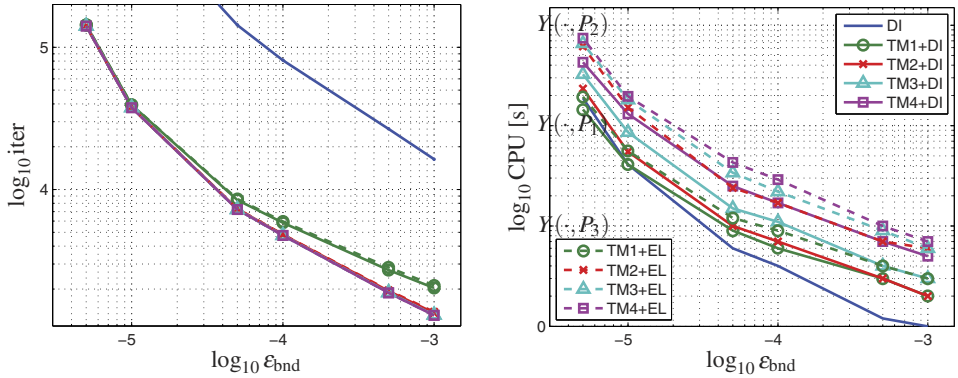


FIG. 2. Performance of guaranteed parameter estimation using various ODE bounding techniques in the set-inversion algorithm. Left: number of iterations vs. convergence threshold. Right: CPU time vs. convergence threshold.

comparisons, the termination criterion is defined in terms of the level of accuracy ε_{bnd} of the solution set (Test 2.2) in the range $10^{-3} \to 5 \times 10^{-6}$—the termination criterion in terms of the minimum box size ε_{box} (Test 2.2) is set to zero, on the other hand. The results are shown in Fig. 2 in terms of the number of iterations (left plot) and CPU time (right plot).

It is evident that classical differential inequalities require by far the largest number of iterations, at any accuracy level. For accuracies of $\varepsilon_{\text{bnd}} = 10^{-5}$ and $\varepsilon_{\text{bnd}} = 5 \times 10^{-6}$, respectively, 932, 454 and 3,612,968 iterations are needed. Memory storage of such a high number of parameter boxes during the course of the algorithm can become a serious issue with an increasing number of uncertain parameters, calling for less conservative bounding techniques. Despite the large number of iterations, however, this approach allows for the fastest computations for accuracies down to $\varepsilon_{\text{bnd}} \approx 10^{-5}$ due to its simplicity. At higher accuracy levels, bounding techniques based on Taylor models are seen to exhibit faster convergence as the extra computational burden of these higher-order bounding techniques is overpowered by a dramatic reduction in overall number of iterations (more than an order of magnitude). The shortest run-time is obtained with first-order Taylor model with interval remainder bounds (labelled TM1+DI in Fig. 2) for $\varepsilon_{\text{bnd}} < 10^{-5}$ here.

In terms of overall number of iterations, the performance of the set-inversion algorithm between first-order Taylor models (both variants TM1+DI and TM1+EL), on the one hand, and between all Taylor models of second-, third- and fourth-order (both variants TM2+DI, TM3+DI, TM4+DI and TM2+EL, TM3+EL, TM4+EL), on the other hand, is about the same. The lower performance of first-order Taylor models compared with higher-order Taylor models can be attributed to the fact that first-order Taylor models compute convex enclosures and are thus limited for the approximation of (potentially) non-convex reachable sets. In terms of the overall run-time though, first-order Taylor models with interval remainders are found to outperform the other bounding techniques based on higher-order Taylor models in this case study. Bearing in mind the trade-off between a smaller number of iterations and a larger processing time needed for a single iteration, it is expected that higher-order bounding techniques will become advantageous for dynamic models of higher complexity or with more uncertain parameters nonetheless.

Finally, the left and right plots in Fig. 3 show projections of the approximate solution sets—$\mathbb{P}_{\text{int}}$ and $\mathbb{P}_{\text{bnd}}$ are shown using the same colour scheme as in Fig. 1 above—onto the (p_1, p_2) and (p_2, p_3)

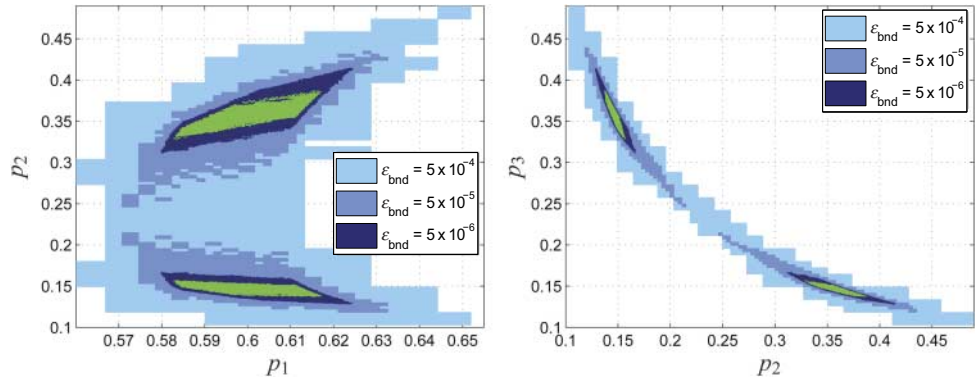


FIG. 3. Outer approximations of the sets of guaranteed parameter estimates for different levels of accuracy ε_{bnd}. Inner approximation of the set of guaranteed parameter estimates for $\varepsilon_{\text{bnd}} = 5 \times 10^{-6}$ plotted in green. Left: projections onto (p_1, p_2) space. Right: projections onto (p_2, p_3) space.

subspaces, respectively, for different levels of accuracy ε_{bnd}. Observe first that the 'true' parameter values p^* used to generate the pseudo-experimental data are part of the solution set, for each reported level ε_{bnd}. Moreover, the solution set for this problem turns out to be disconnected, thus suggesting a possible structural identifiability problem. Interestingly, this non-connectedness of the solution set can only be detected when the accuracy level ε_{bnd} is already lower than 5×10^{-5}. This clearly supports the need for developing strategies that can accelerate the convergence of the set-inverse algorithm, such as box-reduction and other CPU-time-reduction approaches.

3.2 *Domain reduction and CPU-time reduction strategies*

Optimization-based domain reduction. An advantage inherent to using Taylor models for bounding the reachable set of a dynamic system is that the multivariate polynomial part captures the parametric dependencies in the ODE solutions. For given restrictions (constraints) on the state or output variables, it becomes possible in turn to exclude part of the parameter set for which these restrictions cannot be met—the so-called constraint propagation approach. Lin & Stadtherr (2007a) and Kletting *et al.* (2011) used this idea in order to contract the parameter boxes at each iteration of the set-inversion algorithm. Notwithstanding its effectiveness, this approach only exploits the dependencies contained in the linear parts of the Taylor models. In contrast, this paper proposes an optimization-based domain-reduction approach that fully exploits the dependencies in the Taylor models of the predicted outputs.

Given a parameter box $P := [p^{\text{L}}, p^{\text{U}}]$ as well as qth-order Taylor model enclosures $\overline{Y}(t_i, P) := \{\mathscr{P}_y^q(t_i, p) \mid p \in P\} \oplus \mathscr{R}_y^q(t, P) \supseteq Y(t_i, P)$, the lower and upper parameter bounds p_j^{L} and p_j^{U} for each $j = 1, \ldots, n_p$ can be tightened by solving optimization problems of the form:

$$p_j^{\text{L}} = \max \left\{ p_j \mid \{\mathscr{P}_y^q(t_i, p)\} \oplus \mathscr{R}_y^q(t, P) \supseteq Y_{\text{p}}(t_i),\ \forall p \in P,\ \forall i = 1, \ldots, N \right\}, \tag{3.9a}$$

$$p_j^{\text{U}} = \min \left\{ p_j \mid \{\mathscr{P}_y^q(t_i, p)\} \oplus \mathscr{R}_y^q(t, P) \supseteq Y_{\text{p}}(t_i),\ \forall p \in P,\ \forall i = 1, \ldots, N \right\}. \tag{3.9b}$$

This way, a reduced box P is obtained after solving $2 \times n_p$ optimization problems—one problem for the lower bound and one for the upper bound of each parameter. In the case of Taylor models with interval

remainder bounds for instance, $\mathscr{R}_y^q(t,P) := [r_y^{q,\mathrm{L}}(t,P), r_y^{q,\mathrm{U}}(t,P)]$, the inclusion constraints in (3.9) can be equivalently rewritten in the form

$$\mathscr{P}_y^q(t_i,p) + r_y^{q,\mathrm{L}}(t,P) \leqslant \mathrm{Sup}\{Y_\mathrm{p}(t_i)\}, \tag{3.10a}$$

$$\mathscr{P}_y^q(t_i,p) + r_y^{q,\mathrm{U}}(t,P) \geqslant \mathrm{Inf}\{Y_\mathrm{p}(t_i)\}, \tag{3.10b}$$

for each $i = 1,\ldots,N$.

Since the range of the polynomial part of a qth-order Taylor model, $\{\mathscr{P}_y^q(t_i,p) \,|\, p \in P\}$, turns out to be a non-convex set for $q \geqslant 2$ in general, the bound-reduction problems (3.9) themselves be non-convex. Instead of trying to solve these problems directly to global optimality, we construct polyhedral relaxations in the form of LPs, similar to the approach used for bound contraction in branch-and-bound search (see, e.g., Zamora & Grossmann, 1999; Neumaier, 2004; Tawarmalani & Sahinidis, 2004). This relaxation procedure follows three steps:

1. *Decomposition*. The multivariate polynomials $\mathscr{P}_y(t_i,\cdot)$, $i = 1,\ldots,N$, are decomposed into factored form, comprised of binary sums, binary products and univariate composition terms only, via the introduction of auxiliary variables (Smith & Pantelides, 1999; Tawarmalani & Sahinidis, 2004). Here, the constraints in the reformulated optimization problem are either linear or contain a single bilinear term $p_j p_k$ or integer power term $(p_j)^k$.

2. *Relaxation*. The non-convex terms in the reformulated problem are relaxed so as to obtain a convex optimization problem. Here, this relaxation involves replacing both the bilinear and integer power terms with their convex/concave envelopes, e.g., based on McCormick relaxations (McCormick, 1976).

3. *Polyhedral outer approximation*. Since the convex/concave envelopes of power terms are nonlinear in general, polyhedral outer-approximations are constructed via linearization at a number of points. These are usually so chosen as to meet a given level of accuracy (Tawarmalani & Sahinidis, 2004).

By construction, the relaxed optimization problems are fully linear, making it possible to exploit the robustness, efficiency and speed of state-of-the-art LP solvers such as GUROBI or CPLEX. We also note that further improvements could be obtained by tightening the relaxations, for instance using reformulation–linearization technique (Sherali, 2002; Sherali *et al.*, 2012) or exploiting intermediate substructures in the factored optimization problem (Misener & Floudas, 2014; Zorn & Sahinidis, 2014).

In practice, the domain-reduction procedure can be performed as an extra step in the set-inversion algorithm, between Steps 2 and 3. Moreover, in case the reduction of a parameter box P is larger than a given threshold, for instance $\geqslant 20\%$ in volume, it can be repeated multiple times. It is important to bear in mind that repeating the reduction several times requires recomputing the enclosures $\overline{Y}(t_i,P)$ of the model outputs on the reduced box P though. This defines a clear trade-off between the extra computational burden and the reduction in the size of the partition $\mathbb{P}_\mathrm{bnd}$, which is of course problem-dependent. An illustration of the effectiveness of this approach is presented below.

*CPU-time reduction.*When combined with domain-reduction techniques, Taylor models can improve the convergence speed of the set-inversion algorithm significantly. But because Taylor models can also cause a large computational overhead, this benefit is mostly noticeable at an early stage of the

set-inversion procedure, when many boxes can be fathomed or greatly reduced. This calls for further CPU-time-reduction strategies in order to make guaranteed parameter estimation more competitive for high-order Taylor models.

In the basic set-inversion algorithm of Section 2.2, the enclosures $\overline{Y}(t_i, P)$ are recomputed at every iteration because of the overestimation inherent to ODE bounding techniques. In this context, a simple CPU-time-reduction strategy involves reusing the enclosures computed at a parent node (i.e., for a larger parameter box P) as soon as the overestimation at all sampling times and for all output variables has become smaller than a given threshold $\varepsilon_{\text{cvg}} > 0$. For Taylor models, such overestimation is directly measured by the remainder term $\mathscr{R}_x^q$, and a possible re-usability condition thus reads

$$\forall i \in \{1, \ldots, N\}, \quad \operatorname{diam}(\mathscr{R}_x^q(t_i, P)) \leqslant \varepsilon_{\text{cvg}}. \tag{3.11}$$

As soon as this condition is met, the corresponding Taylor models $(\mathscr{P}_x^q(t_i, \cdot), \mathscr{R}_x^q(t_i, P))$ can indeed be stored and used later on in any child node $P' \subseteq P$, effectively by-passing the ODE bounding step 2. Variants of this approach can of course be used that consider relative convergence criteria and scaling for instance.

In addition to reusing Taylor models at children nodes, a further CPU-time-reduction strategy involves reducing the order of the Taylor models, which can lead to significant savings in connection to the relaxation and solution of the optimization-based domain-reduction problems (3.9). A simple order-reduction procedure is as follows:

Input: Convergence threshold $\varepsilon_{\text{cvg}} > 0$; parameter box P; qth-order Taylor models $(\mathscr{P}_x^q(t_i, \cdot), \mathscr{R}_x^q(t_i, P))$ of $x(t_i, \cdot)$ on P satisfying (3.11)

Initialization: Set reduced order $\varrho = q$

Main Loop:

1. Compute enclosures $B^\varrho(t_i, P) \supseteq \{\sum_{\gamma \in \mathbb{N}^{n_p},\ |\gamma| = \varrho} \frac{\partial^\gamma x(t_i, \hat{p})}{\gamma!} \mid p \in P\}$ for all $i \in \{1, \ldots, N\}$
2. **If** $\operatorname{diam}(B^\varrho(t_i, P)) > \varepsilon_{\text{cvg}}$ for some $i \in \{1, \ldots, N\}$, **stop**
3. Reduce order ϱ-=1; **Return** to step 1

Output: Reduced Taylor model order ϱ

In particular, bounding of all the monomials of a given order ϱ in step 1 can be achieved using interval analysis or other less conservative strategies (Lin & Rokne, 1995; Neumaier, 2002). Regarding the convergence threshold ε_{cvg} finally, we like to note that a larger threshold will lead to reusing Taylor models from parent nodes earlier as well as reducing their order faster, but too large a threshold can prevent convergence of the set-inversion algorithm if the stopping criterion is based solely on the total volume threshold ε_{bnd} (Step 4a).

Case study (*Continued*).We continue the case study of the dynamic system (3.8) in order to investigate the effect of optimization-based domain reduction and CPU-time reduction. Guaranteed parameter estimation is applied with and without the use of domain reduction as an extra step in the set-inversion

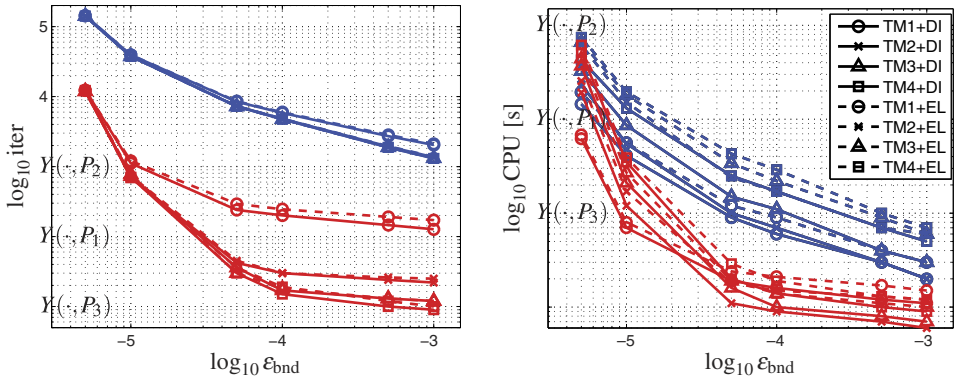


FIG. 4. Performance of guaranteed parameter estimation with (red lines) and without (blue lines) the use of domain reduction and with ODE bounding techniques based on Taylor models of orders $q = 1, \ldots, 4$. Left: number of iterations vs. convergence threshold. Right: CPU time vs. convergence threshold.

algorithm (reduction threshold of 20% and maximum of 10 reduction loops at each iteration). Taylor models of orders $q = 1, \ldots, 4$ are considered for enclosing the output reachable set $\overline{Y}(\cdot, P)$, and the termination criteria remain the same as defined previously.

The number of iterations and the CPU time required by the set-inversion algorithm to terminate with different Taylor model orders and with or without the use of domain reduction are reported on the left and right plots of Fig. 4, respectively, as a function of the termination tolerance ε_{bnd}. It is evident that the number of iterations decreases significantly when domain reduction is used—here by at least one order of magnitude for all considered tolerance levels ε_{bnd}. Moreover, the higher the order of the Taylor model, the smaller the number of iterations required by the algorithm to converge for a given accuracy level. In terms of overall CPU time, the use of domain reduction is found to be mostly beneficial at an early stage of the set-inversion procedure, where many boxes can be significantly reduced or even eliminated using optimization-based domain reduction.

A certain trade-off is observed in terms of CPU time on the right plot of Fig. 4, whereby higher-order Taylor models can cause a significant computational overhead. On the whole, first- or second-order Taylor models with interval remainder bounds are found to enable the fastest computations in this case study. Although higher-order Taylor models reduce the overestimation, lower-order Taylor models eventually become computationally advantageous as the parameter boxes shrink. Another trade-off is observed in terms of the overhead caused by the application of domain reduction (construction and solution of relaxed LP problems). These trends show a clear need for CPU-time-reduction strategies in connection to Taylor model-based ODE bounding. Nonetheless, when used in combination with domain reduction, Taylor model-based ODE bounders now greatly outperforms classical differential inequalities (see Fig. 2).

The plots in Fig. 5 show the projections of the solution set outer-approximation onto the (p_1, p_2) and (p_2, p_3) subspaces, for increasing accuracy levels of $\varepsilon_{\text{bnd}} = 5 \times 10^{-4}, 5 \times 10^{-5}$ and 5×10^{-6}, using optimization-based domain reduction and second-order Taylor models with interval remainder terms for ODE bounding. In comparing outer-approximations of the guaranteed parameter set P_{e} for various accuracy levels, it is found that setting $\varepsilon_{\text{bnd}} = 5 \times 10^{-5}$ already provides a tight approximation of P_{e}, with only 34 boxes and a run-time of $\sim$2 s.

As expected, a much tighter approximation is obtained by setting $\varepsilon_{\text{bnd}} = 5 \times 10^{-6}$, yet this is at the price of a much finer box partition comprising 11,250 boxes here and a corresponding run-time of

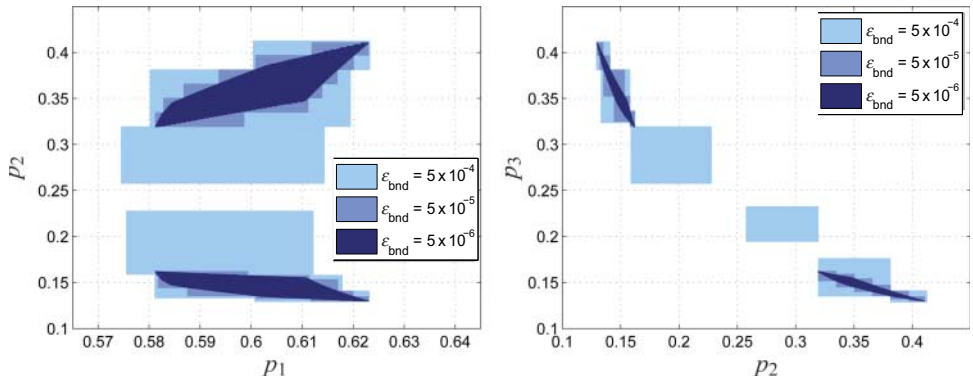


FIG. 5. Outer approximations of the sets of guaranteed parameter estimates for different levels of accuracy ε_{bnd} using domain reduction and second-order Taylor models. Left: projections onto (p_1, p_2) space. Right: projections onto (p_2, p_3) space.

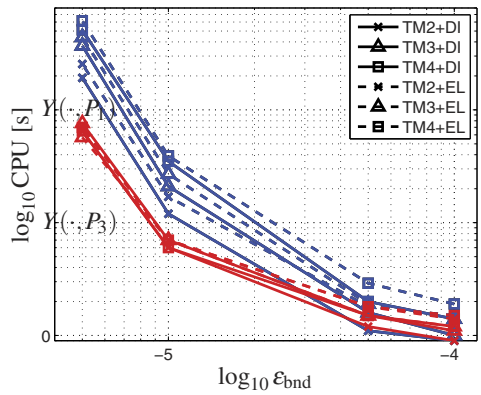


FIG. 6. Performance of guaranteed parameter estimation with (red lines) and without (blue lines) CPU-time-reduction strategies, in combination with domain-reduction strategy and ODE bounding techniques based on Taylor models of orders $q = 2, \ldots, 4$: CPU time vs. convergence threshold.

over 60 s. For the sake of comparison we also note that, when no domain reduction is used, the partition comprises over 2,200 boxes with $\varepsilon_{\text{bnd}} = 5 \times 10^{-5}$ and over 70,000 boxes with $\varepsilon_{\text{bnd}} = 5 \times 10^{-6}$. These results also suggest that the efficiency of the set-inversion algorithm in computing highly accurate set approximations could be improved significantly if affine cuts were enabled in addition to simple bounds contraction during the domain-reduction procedure. Such cuts would provide the extra flexibility needed to closely approximate the actual parameter set and will be the topic of future research.

Finally, we investigate the effect of CPU-time reduction, by considering both strategies of reusing and reducing the order of Taylor models computed at parent nodes. A convergence threshold of $\varepsilon_{\text{cvg}} = 10^{-4}$ (determined heuristically) is used here.

Computational time requirements for the set-inversion algorithm to converge are shown in Fig. 6 for various termination tolerances ε_{bnd}. Not reported on this plot are the CPU times for first-order Taylor models since the corresponding improvement is marginal—convergence of first-order Taylor models within $\varepsilon_{\text{cvg}} = 10^{-4}$ is only achieved for very small parameter boxes in this case. For higher-order Taylor

models, it is evident that the CPU-time-reduction strategies are effective. The best performance is achieved when second-order Taylor models with interval remainder bounds are used, but third- and fourth-order Taylor models lead to comparable run-times nonetheless.

With all the proposed improvements used together, guaranteed parameter estimation of the dynamic system (3.8) can be solved to within $\varepsilon_{\text{bnd}} = 5 \times 10^{-6}$ in <60 s. This is a three-fold reduction compared with the classical method of differential inequalities.

4. Guaranteed parameter estimation for an anaerobic digestion process

This section illustrates the benefits of using high-order ODE bounding, optimization-based domain reduction, and CPU-time reduction in the context of guaranteed parameter estimation for a case study in anaerobic digestion. We consider a six-state model representing the dynamics of an anaerobic digester, as originally proposed by Bernard *et al.* (2001). Enclosing the solutions of this model in the presence of parametric uncertainty is challenging due to the presence of complex and liquid–gas transfer and pH self-regulation mechanisms. Moreover, the system exhibits both fast dynamics acting on a time-scale of minutes/hours, and slow dynamics acting on a time-scale of days.

$$\dot{X}_1 = (\mu_1(S_1) - \alpha D)X_1, \tag{4.1a}$$

$$\dot{X}_2 = (\mu_2(S_2) - \alpha D)X_2, \tag{4.1b}$$

$$\dot{S}_1 = D(S_1^{\text{in}} - S_1) - k_1\mu_1(S_1)X_1, \tag{4.1c}$$

$$\dot{S}_2 = D(S_2^{\text{in}} - S_2) + k_2\mu_1(S_1)X_1 - k_3\mu_2(S_2)X_2, \tag{4.1d}$$

$$\dot{Z} = D(Z^{\text{in}} - Z), \tag{4.1e}$$

$$\dot{C} = D(C^{\text{in}} - C) - q_{\text{CO}_2} + k_4\mu_1(S_1)X_1 + k_5\mu_2(S_2)X_2. \tag{4.1f}$$

The states X_1 and X_2 stand for the concentrations of acidogenic and methanogenic biomass, respectively; S_1, the organic substrate concentration (COD other than volatile fatty acids (VFA)); S_2, the VFA concentration; Z, the total alkalinity concentration (TALK); and C, the total inorganic carbon concentration (TIC). Moreover, D represents the dilution rate; S_1^{in}, S_2^{in}, Z^{in} and C^{in} are the inlet concentrations of organic substrate, VFA, TALK and TIC, respectively; α is the fraction of biomass in the liquid phase (i.e., not attached to a support); and $k_1, \ldots, k_6$ are pseudo-stoichiometric yield coefficients.

The specific growth rates of acidogenic bacteria, μ_1, and methanogenic bacteria, μ_2, are assumed to follow Michaelis–Menten and Haldane kinetics,

$$\mu_1(S_1) := \bar{\mu}_1 \frac{S_1}{S_1 + K_{S_1}}, \tag{4.1g}$$

$$\mu_2(S_2) := \bar{\mu}_2 \frac{S_2}{S_2 + K_{S_2} + S_2^2/K_{I_2}}, \tag{4.1h}$$

with maximum growth rates $\bar{\mu}_1$ and $\bar{\mu}_2$, half-saturation constants K_{S_1} and K_{S_2}, and inhibition constant K_{I_2} (methanogenic bacteria only). Finally, the molar flow rate of CO_2, q_{CO_2}, is given by

$$q_{\text{CO}_2} := k_{\text{L}}a(C + S_2 - Z - K_{\text{H}}P_{\text{CO}_2}), \tag{4.1i}$$

TABLE 1 *Estimated parameters of the anaerobic digestion model* (4.1)

Parameter	Nominal value	Range	Unit
$\bar{\mu}_1$	1.2	[1.15, 1.25]	/day
K_{S_1}	7.1	[6.7, 7.3]	g(COD)/L
$\bar{\mu}_2$	0.74	[0.735, 0.75]	/day
K_{S_2}	9.28	[9.2, 9.5]	mmol/L
K_{I_2}	256	[235.0, 265.0]	mmol/L

TABLE 2 *Constant parameters and initial states of the anaerobic digestion model* (4.1)

Parameter	Value	Unit	Parameter	Value	Unit
k_1	42.14	g(COD)/g(cell)	$X_1(0)$	0.5	g(VSS)/L
k_2	116.5	mmol/g(cell)	$X_2(0)$	1.0	g(VSS)/L
k_3	268.0	mmol/g(cell)	$S_1(0)$	1.0	g(COD)/L
k_4	50.6	mmol/g(cell)	$S_2(0)$	5.0	mmol/L
k_5	343.6	mmol/g(cell)	$C(0)$	40.0	mmol/L
k_6	453.0	mmol/g(cell)	$Z(0)$	50.0	mmol/L
k_La	19.8	/day	P_t	1	atm
K_H	16	mmol/L/atm	α	0.5	–

$$\text{with} \quad P_{CO_2} := \frac{\phi_{CO_2} - \sqrt{\phi_{CO_2}^2 - 4K_H P_t (C + S_2 - Z)}}{2K_H} \tag{4.1j}$$

$$\phi_{CO_2} := C + S_2 - Z + K_H P_t + \frac{k_6}{k_L a} \mu_2(S_2) X_2, \tag{4.1k}$$

where k_La denotes the liquid–gas transfer constant, K_H is Henry's constant and P_t is the total pressure.

Nominal values for all the parameters are taken from Bernard *et al.* (2001). We apply guaranteed parameter estimation to estimate the kinetic parameters describing biomass growth. These parameters are listed in Table 1 with their nominal values and the considered variation ranges. The rest of the parameters as well as the initial conditions used are reported in Table 2 for the sake of completeness.

In order to apply guaranteed parameter estimation, pseudo-experimental data are generated by simulating the model (4.1) with nominal parameter values from Tables 1 and 2 over a four-day period. The profiles used for the dilution rate and for the influent concentrations are those reported in Table 3. Moreover, three outputs are considered to carry out the estimation, namely S_1, S_2, and C, with measurements every 4 h. In order to simulate the effect of measurement noise, the simulated values are rounded up or down to the nearest values by retaining, respectively, 2, 1 and 1 significant digits only; then, measurement error ranges of, respectively, ± 0.01, ± 0.1 and ± 0.1 are added around these values.

In the remainder of this section, we investigate guaranteed parameter estimation with different bounding techniques and with both optimization-based domain-reduction and CPU-time-reduction strategies in order to demonstrate the proposed improvements on a real-life problem. As previously in the simple case study, a 20% threshold and a maximum of 10 reduction loops are defined for the optimization-based domain-reduction strategy, and an absolute convergence threshold of $\varepsilon_{cvg} = 10^{-4}$ is

TABLE 3 *Dilution rate and inlet concentration profiles corresponding to the pseudo-experimental data*

Input	Day 1	Day 2	Day 3	Day 4
D [/day]	0.25	1.00	1.00	0.25
S_1^{in} [g(COD)/L]	2.38	2.38	4.76	2.38
S_2^{in} [mmol/L]	80.0	80.0	160.0	80.0
Z^{in} [mmol/L]	50.0	50.0	100.0	50.0
C^{in} [mmol/L]	5.0	5.0	10.0	5.0

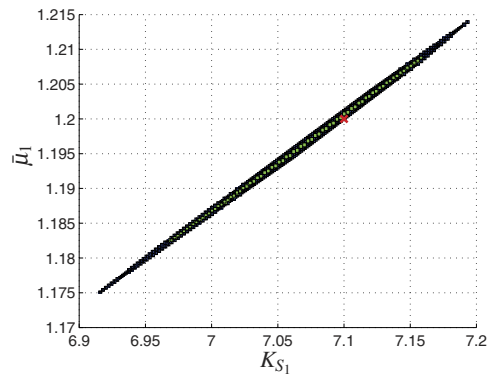


FIG. 7. Guaranteed parameter set approximation ($\mathbb{P}_{\text{int}}$ in green, $\mathbb{P}_{\text{bnd}}$ one in blue) for Case Study 1. The red cross indicates the 'true' (nominal) parameter values.

defined in connection to the CPU-time-reduction strategy. Problems of increasing complexity with 2, 3, 5 and 7 estimated parameters are addressed in Sections 4.1–4.3.

4.1 *Case study* 1 – *two-parameter guaranteed parameter estimation*

We consider the estimation of the parameters $\bar{\mu}_1$ and K_{S_1}, while the rest of the parameters from Tables 1 and 2 are fixed at their nominal values. The set-inversion algorithm is used with continuous-time ODE bounding techniques propagating Taylor models of orders $q = 1, \ldots, 4$ with interval or ellipsoidal remainders. The termination criterion is defined as $\varepsilon_{\text{bnd}} = 10^{-4}$, whereas ε_{box} is set to zero, and the maximum number of iterations and maximal computational time are set to 1,000,000 iterations and 10 h, respectively.

Figure 7 shows both the inner- and outer-approximation of the set of guaranteed parameter estimates for the selected termination criteria. We start by noting that the true parameter values lie inside the approximation of the set P_e and that the selected termination criteria appear to be appropriate in view of the approximation level. Such a shape of the guaranteed parameter set is characteristic of the large correlations between the parameters $\bar{\mu}_1$ and K_{S_1}, according to (4.1g), and shows that $\mu_1(S_1) \approx (\bar{\mu}_1/K_{S_1})X_1$ in this case.

When the method of differential inequalities is used to bound the reachable set, the algorithm stops after 1,000,000 iterations, without reaching the desired level of approximation—the volume of

TABLE 4 *Iteration counts and run-times of the set-inversion algorithm for Case Study* 1

Bounding method	Domain reduction	CPU-time reduction	Number of iterations	CPU time [s]
TM1 + EL	✗	✗	8,738	801
TM1 + DI	✓	✗	4,280	1,682
TM4 + EL	✓	✗	3,690	5,748
TM4 + EL	✓	✓	5,229	49

the partition $\mathbb{P}_{\text{bnd}}$ is $\sim 2.5 \times 10^{-4}$ then. This behaviour is attributed to the inability of the method of differential inequalities to generate tight bounds for the anaerobic digestion model, even for very small parameter uncertainty. In contrast, higher-order ODE bounding techniques enable convergence of the set-inversion algorithm, as summarized in Table 4. Using Taylor models in combination with domain reduction, the algorithm is found to converge within a few thousand iterations (2nd and 3rd row), yet this remains insufficient to override the extra computational burden associated with domain reduction (1st row). The use of domain reduction becomes advantageous only when combined with CPU-time reduction (4th row), then leading to dramatic reduction of the run-time down to 49 s. Note that the number of iterations increases in the latter case compared with a run with the same settings but without CPU-time-reduction strategies, a behaviour that is indeed expected and attributed to the approximation introduced by the finite convergence threshold ε_{cvg}.

4.2 *Case study* 2*—three-parameter guaranteed parameter estimation*

Next, we consider the estimation of the parameters $\bar{\mu}_2$, K_{S_2} and K_{I_2}, while the rest of the parameters from Tables 1 and 2 are fixed at their nominal values. The set-inversion algorithm is run with the exact same settings as previously in Section 4.1, to the exception of the termination criterion ε_{bnd} that is now set to 5×10^{-5}.

Figure 8 shows the outer-approximation of the set of guaranteed parameter estimates for the selected termination criteria. The true parameter values lie inside the approximation of the set P_{e} and the selected termination criteria is deemed appropriate by visual inspection of the approximation level. Here again, the shape of the guaranteed parameter set is expected given the large correlations between the parameters $\bar{\mu}_2$, K_{S_2} and K_{I_2} according to (4.1h).

When the method of differential inequalities is used to bound the reachable set, the algorithm stops after 1,000,000 iterations, without reaching the desired level of approximation—the volume of the partition $\mathbb{P}_{\text{bnd}}$ is $\sim 1.7 \times 10^{-3}$ then. This behaviour is again due to the inability of the method of differential inequalities to generate tight bounds for the anaerobic digestion model, even for very small parametric uncertainty. In contrast, higher-order ODE bounding techniques enable convergence of the set-inversion algorithm, as summarized in Table 5. Using first-order Taylor models with ellipsoidal remainders but no other improvement, the set-inversion algorithm takes ∼24,000 iterations to converge (1st row). This is to be compared with a few thousand iterations when domain domain reduction is used (2nd, 3rd and 4th rows), similar to the previous 2-parameter case despite the extra parameter. This suggests that the domain reduction might become more and more advantageous as the number of uncertain parameters increases, a trend that will confirm later on in Section 4.3. An expected behaviour here is the reduction in the number of iterations as higher-order Taylor models are used. Finally, the effect of the CPU-time-reduction strategy is rather dramatic, with a run-time reduction about two orders of magnitude lower in the case of fourth-order Taylor models with ellipsoidal remainder bounds. It is noteworthy that the

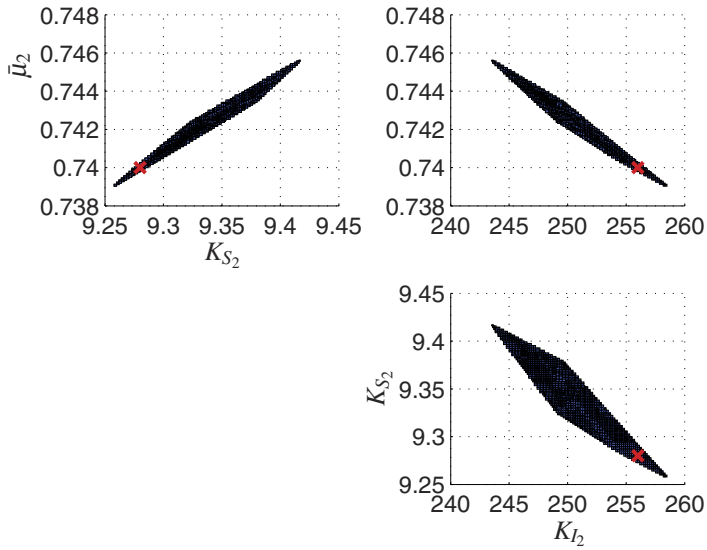


FIG. 8. Outer approximation of the set of guaranteed parameter estimates for Case Study 2. Projections onto the subspaces $(K_{S_2},\bar{\mu}_2)$, $(K_{I_2},\bar{\mu}_2)$ and (K_{I_2},K_{S_2}). The red crosses indicate the true (nominal) parameter values.

TABLE 5 *Iteration counts and run-times of the set-inversion algorithm for Case Study* 2

Bounding method	Domain reduction	CPU-time reduction	Number of iterations	CPU time [s]
TM1 + EL	✗	✗	23,838	3,584
TM1 + DI	✓	✗	4,095	1,154
TM4 + EL	✓	✗	3,423	9,593
TM4 + EL	✓	✓	3,488	111

shortest runtime in this case is even lower, down to 41 s when fourth-order Taylor models with interval remainder bounds are used.

4.3 *Case study* 3—*five- and seven-parameter guaranteed parameter estimation*

We now consider the estimation of the parameters $\bar{\mu}_1$, K_{S_1}, $\bar{\mu}_2$, K_{S_2} and K_{I_2} simultaneously, leaving the other parameters at their nominal values in Table 2. The set-inversion algorithm is run with the exact same settings as previously in Sections 4.1 and 4.2, apart from the termination criterion ε_{bnd} that is now set to 5×10^{-7}.

Figure 9 shows the outer-approximation of the set of guaranteed parameter estimates for the selected termination criterion. The true parameter values lie inside the approximation of the set P_{e} and the approximation level, although coarse, validates the chosen termination criterion. Large correlations between the parameters $\bar{\mu}_1$ and K_{S_1}, on the one hand, and between $\bar{\mu}_2$, K_{S_2} and K_{I_2}, on the other hand, are observed, which is in complete agreement with the results shown earlier in Figs 7 and 8. In contrast, rather small cross-correlations are observed between these two parameter subsets, as illustrated for instance for the parameters K_{S_1} and $\bar{\mu}_2$ in the top-right plot of Fig. 9.

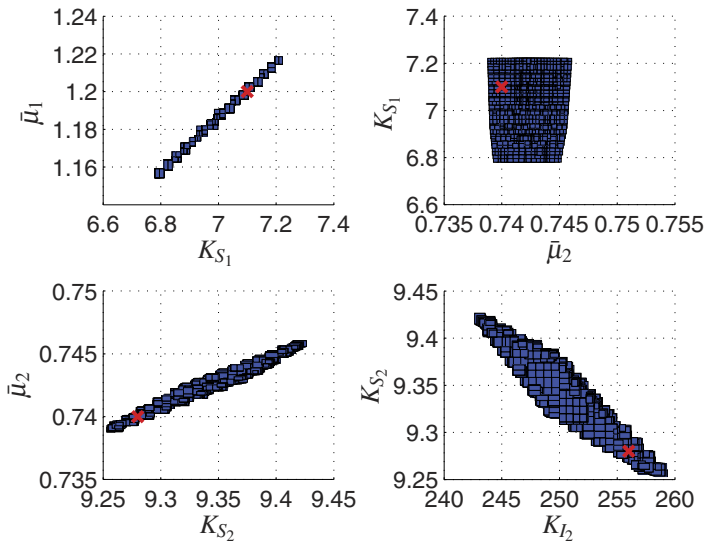


FIG. 9. Outer approximation of the set of guaranteed parameter estimates for Case Study 3. Projections onto the subspaces $(K_{S_1},\bar{\mu}_1)$, $(\bar{\mu}_2,K_{S_1})$, $(K_{S_2},\bar{\mu}_2)$ and $(K_{I_2},\bar{\mu}_2)$ only. The red crosses indicate the true (nominal) parameter values.

TABLE 6 *Iteration counts and run-times of the set-inversion algorithm for Case Study* 3

Bounding method	Domain reduction	CPU-time reduction	Number of iterations	CPU time [s]
TM1 + EL	✗	✗	219,178	36,000†
TM1 + DI	✓	✗	41,148	8,295
TM4 + EL	✓	✗	3,010	18,346
TM4 + EL	✓	✓	3,130	2,118

†Did not converge to the specified termination criterion within the maximum allowable time.

Table 6 presents a comparison of the performance of various ODE bounding techniques and other improvement strategies. As previously, early termination is obtained with the method of differential inequalities after 1,000,000 iterations (without a single parameter box being fathomed here), and the algorithm now fails to converge after 10 h with Taylor models as well when domain reduction is not applied. With respect to Taylor models combined with domain reduction, the benefit of higher-order ODE bounds in terms of the number of iterations is becoming more obvious in this 5-parameter problem—for instance, 10 times more iterations are needed with a first-order Taylor model compared to a fourth-order one. Yet, this large reduction is still not enough to overpower the extra computational burden of a single iteration with a higher-order Taylor model. Only when used in combination with CPU-time-reduction strategies are fourth-order Taylor models found to become competitive, with a runtime down to about 2,100 s. Finally, it is noteworthy that the shortest runtime in this case is close to 1,400 s, which is obtained for fourth-order Taylor model with interval remainder bounds and with all the developed reduction strategies.

Concerning the anaerobic digestion application, a more realistic parameter estimation problem should of course consider the initial biomass concentrations to be uncertain as well. Adding both initial concentrations $X_1(0)$ and $X_2(0)$ to the five uncertain kinetic parameters in Table 1 yields a

total of seven parameters. In order to carry out the computations, only a small level of uncertainty of ± 0.001 g(VSS)/L is considered for $X_1(0)$ and $X_2(0)$ here, and the termination criterion ε_{bnd} is decreased to 5×10^{-11}. The set-inversion algorithm appears to be tractable only with fourth-order Taylor models (or higher) and only when combined with domain-reduction and CPU-time-reduction strategies—convergence is achieved after 4,225 iterations and a corresponding runtime just above 19,000 s (5.3 h) in this case. These results confirm the advantage of high-order ODE bounding techniques and improved domain- and CPU-time-reduction strategies in addressing real-life problems possessing complex dynamics and more than a handful of uncertain parameters.

4.4 *Discussion*

In order to bring guaranteed parameter estimation to the next level and allow for problems with more than 10 parameters, the case studies in this paper suggest that improving the performance of existing ODE bounding techniques remains key, i.e., to enable tight bounds on larger parameter ranges while reducing the computational burden at the same time. It is also found that both domain-reduction and CPU-time-reduction strategies can have a dramatic effect on the performance, and it would thus appear important to develop new 'smart' heuristics to further enhance the search; see, e.g., the recent work by Caprara & Locatelli (2010) and Locatelli (2014). The use of tight termination tolerances in the set-inversion algorithm typically results in a very large accumulation of parameter boxes, similar in essence to the cluster effect in global optimization (Du & Kearfott, 1994; Neumaier, 2004). Strategies to help mitigate this behaviour are also clearly warranted.

A practical limitation for the guaranteed parameter estimation problem as formulated in (2.5) is the need for consistent measurement data and bounds throughout the entire time series; otherwise, there may not be any model response matching the output measurements within the specified error bounds, in which case the parameter set $\mathscr{P}_e$ is empty. In most applications based on real data, this calls for data preprocessing, for instance using data reconciliation techniques, in order to get rid of the outliers. For instance, these outliers could be due to over-optimistic noise bounds or to sensor failures at given time instants. To handle this situation, it is possible to 'protect' the estimator against at most n outliers, by allowing for a number of output variables to be outside of their prior feasible intervals (see, e.g., Jaulin *et al.*, 2001; Kieffer & Walter, 2005). Another situation whereby the parameter set $\mathscr{P}_e$ may be empty is in the presence of significant model mismatch. Taking guaranteed parameter estimation to the next level in order to address complex, large-scale problems of practical applicability, calls for the development of further robustification strategies as well as alternative guaranteed parameter estimation paradigms (Csáji *et al.*, 2012; Kieffer & Walter, 2014).

5. Conclusions

The focus of this paper has been on the problem of guaranteed parameter estimation, which seeks to determine all parameter values of a dynamic model that are consistent with some experimental data, within specified error bounds. Set-inversion techniques based on exhaustive search are considered, with special emphasis on high-order bounding techniques for uncertain dynamic systems combined with efficient strategies for enhancing convergence speed. Specifically, a methodology is developed which starts by computing state/output bounds in the form of Taylor models that capture the parametric dependencies, and then takes full advantage of these Taylor models to perform optimization-based domain reduction. On top of this, strategies are implemented in order to decrease the overall run-time, whereby the Taylor models computed at a parent node can be reused and their order automatically reduced wherever

possible. The potential of these new developments have been demonstrated both on a simple case study and on a real-life problem in anaerobic digestion with up to seven uncertain parameters, showing clear improvements of the set-inversion algorithm for guaranteed parameter estimation.

Acknowledgements

R.P. gratefully acknowledges the contribution of the European Commission under research project MOBOCON (Grant agreement number 291458). M.V. and B.C. acknowledge financial support from Marie Curie under grant PCIG09-GA-2011-293953. M.V. is also grateful to CONACYT for a doctoral scholarship.

References

Belotti, P., Lee, J., Liberti, L., Margot, F. & Wächter, A. (2009) Branching and bounds tightening techniques for non-convex MINLP. *Optim. Methods Softw.*, **24**, 597–634.

Bernard, O., Hadj-Sadok, Z., Dochain, D., Genovesi, A. & Steyer, J. P. (2001) Dynamical model development and parameter identification for an anaerobic wastewater treatment process. *Biotechnol. Bioeng.*, **75**, 424–438.

Berz, M. & Makino, K. (1998) Verified integration of ODEs and flows using differential algebraic methods on high-order Taylor series. *Reliab. Comput.*, **4**, 361–369.

Bompadre, A., Mitsos, A. & Chachuat, B. (2013) Convergence analysis of Taylor and McCormick–Taylor models. *J. Global Optim.*, **57**, 75–114.

Caprara, A. & Locatelli, M. (2010) Global optimization problems and domain reduction strategies. *Math. Program.*, **1**, 123–137.

Chachuat, B. & Villanueva, M. E. (2012) Bounding the solutions of parametric ODEs: when Taylor models meet differential inequalities. *22nd European Symposium on Computer Aided Process Engineering*, vol. 30 (I. D. L. Bogle & M. Fairweather eds). Amsterdam: Elsevier, pp. 1307–1311.

Csáji, B. C., Campi, M. C. & Weyer, E. (2012) Non-asymptotic confidence regions for the least-squares estimate. *Proceedings of the 16th IFAC Symposium on System Identification (SYSID 2012)*. Brussels, Belgium, pp. 227–232.

Du, K. & Kearfott, R. B. (1994) The cluster problem in multivariate global optimization. *J. Glob. Optim.*, **5**, 253–265.

Houska, B., Logist, F., Van Impe, J. & Diehl, M. (2012) Robust optimization of nonlinear dynamic systems with application to a jacketed tubular reactor. *J. Process Control*, **22**, 1152–1160.

Houska, B., Villanueva, M. E. & Chachuat, B. (2013) A validated integration algorithm for nonlinear ODEs using Taylor models and ellipsoidal calculus. *Proceedings of the 52nd IEEE Conference on Decision and Control (CDC)*. Florence, Italy, pp. 484–489.

Jaulin, L. (2002) Nonlinear bounded-error state estimation of continuous-time systems. *Automatica*, **38**, 1079–1082.

Jaulin, L., Kieffer, M., Didrit, O. & Walter, E. (2001) *Applied Interval Analysis*. London: Springer.

Jaulin, L. & Walter, E. (1993) Set inversion via interval analysis for nonlinear bounded-error estimation. *Automatica*, **29**, 1053–1064.

Kieffer, M. & Walter, E. (2005) Interval analysis for guaranteed nonlinear parameter and state estimation. *Math. Comput. Model. Dyn. Syst.*, **11**, 171–181.

Kieffer, M. & Walter, E. (2011) Guaranteed estimation of the parameters of nonlinear continuous-time models: contributions of interval analysis. *Int. J. Adapt. Control Signal Process.*, **25**, 191–207.

Kieffer, M. & Walter, E. (2014) Guaranteed characterization of exact non-asymptotic confidence regions as defined by LSCR and SPS. *Automatica*, **50**, 507–512.

Kletting, M., Kieffer, M. & Walter, E. (2011) Two approaches for guaranteed state estimation of nonlinear continuous-time models. *Modeling, Design, and Simulation of Systems with Uncertainties, Mathematical Engineering*, vol. 3. Berlin, Heidelberg, New York: Springer, 199–220.

Kurzhanski, A. B. & Varaiya, P. (2002) Reachability analysis for uncertain systems—the ellipsoidal technique. *Dyn. Continuous, Discrete Impulsive Syst. Ser. B*, **9**, 347–368.

Lin, Q. & Rokne, J. G. (1995) Methods for bounding the range of a polynomial. *J. Comput. Appl. Math.*, **58**, 193–199.

Lin, Y. & Stadtherr, M. A. (2007a) Guaranteed state and parameter estimation for nonlinear continuous-time systems with bounded-error measurements. *Ind. Eng. Chem. Res.*, **46**, 7198–7207.

Lin, Y. & Stadtherr, M. A. (2007b) Validated solutions of initial value problems for parametric ODEs. *Appl. Numer. Math.*, **57**, 1145–1162.

Locatelli, M. (2014) Alternative branching rules for some nonconvex problems. *Optim. Methods Softw*. doi:10.1080/10556788.2014.885521.

Lohner, R. J. (1992) Computation of guaranteed enclosures for the solutions of ordinary initial and boundary value problems. *Computational Ordinary Differential Equations*, vol. 1 (J. R. Cash & I. Gladwell eds). Oxford: Clarendon Press, pp. 425–436.

Makino, K. & Berz, M. (1999) Efficient control of the dependency problem based on Taylor model methods. *Reliab. Comput.*, **5**, 3–12.

McCormick, G. P. (1976) Computability of global solutions to factorable nonconvex programs: Part I—Convex underestimating problems. *Math. Program.*, **10**, 147–175.

Misener, R. & Floudas, C. A. (2014) A framework for globally optimizing mixed-integer signomial programs. *J. Optim. Theor. Appl*. doi:10.1007/s10957-013-0396-3.

Moore, R. E. (1992) Parameter sets for bounded-error data. *Math. Comput. Simul.*, **34**, 113–119.

Moore, R. E., Kearfott, R. B. & Cloud, M. J. (2009) *Introduction to Interval Analysis*. Philadelphia, PA: SIAM.

Nedialkov, N. S., Jackson, K. R. & Corliss, G. F. (1999) Validated solutions of initial value problems for ordinary differential equations. *Appl. Math. Comput.*, **105**, 21–68.

Neher, M., Jackson, K. R. & Nedialkov, N. S. (2007) On Taylor model based integration of ODEs. *SIAM J. Numer. Anal.*, **45**, 236–262.

Neumaier, A. (2002) Taylor forms—use and limits. *Reliab. Comput.*, **9**, 43–79.

Neumaier, A. (2004) Complete search in continuous global optimization and constraint satisfaction. *Acta Numer.*, **13**, 271–369.

Paulen, R., Villanueva, M. E., Fikar, M. & Chachuat, B. (2013) Guaranteed parameter estimation in nonlinear dynamic systems using improved bounding techniques. *Proceedings of the 2013 European Control Conference (ECC'13)*. Zürich, Switzerland, pp. 4514–4519.

Raissi, T., Ramdani, N. & Candau, Y. (2004) Set membership state and parameter estimation for systems described by nonlinear differential equations. *Automatica*, **40**, 1771–1777.

Rauh, A., Hofer, E. P. & Auer, E. (2006) VALENCIA-IVP: a comparison with other initial value problem solvers. *Proceedings of the 12th GAMM-IMACS International Symposium on Scientific Computing, Computer Arithmetic and Validated Numerics (SCAN'2006)*. Duisburg, Germany, p. 36.

Sahlodin, A. M. (2012) Global optimization of dynamic process systems using complete search methods. *Ph.D. Thesis*, McMaster University.

Sahlodin, A. M. & Chachuat, B. (2011) Convex/concave relaxations of parametric ODEs using Taylor models. *Comput. Chem. Eng.*, **35**, 844–857.

Sherali, H. D. (2002) Tight relaxations for nonconvex optimization problems using the reformulation-linearization/convexification technique (RLT). *Handbook of Global Optimization*, vol. 2 (P. M. Pardalos & H. E. Romeijn eds). Dordrecht, The Netherlands: Kluwer Academic Publishers, pp. 1–63.

Sherali, H. D., Dalkiran, E. & Liberti, L. (2012) Reduced RLT representations for nonconvex polynomial programming problems. *J. Glob. Optim.*, **52**, 447–469.

Smith, E. M. B. & Pantelides, C. C. (1999) A symbolic reformulation/spatial branch-and-bound algorithm for the global optimisation of nonconvex MINLPs. *Comput. Chem. Eng.*, **23**, 457–478.

Tawarmalani, M. & Sahinidis, N. V. (2004) Global optimization of mixed-integer nonlinear programs: a theoretical and computational study. *Math. Program.*, **99**, 563–591.

VILLANUEVA, M. E., HOUSKA, B. & CHACHUAT, B. (2014) Unified framework for the propagation of continuous-time enclosures for parametric nonlinear ODEs. *J. Glob. Optim*. doi:10.1007/s10898-014-0235-6.

VILLANUEVA, M. E., PAULEN, R., HOUSKA, B. & CHACHUAT, B. (2013) Enclosing the reachable set of parametric ODEs using Taylor models and ellipsoidal calculus. *23rd European Symposium on Computer Aided Process Engineering*, 32 (A. Kraslawski & I. Turunen eds). Amsterdam: Elsevier, pp. 979–984.

WALTER, W. (1970) *Differential and Integral Inequalities*. Berlin: Springer.

WALTER, E. (ed.) (1990) Parameter identifications with error bound. *Mathematics & Computers in Simulation*, vol. 32. Amsterdam: Elsevier.

ZAMORA, J. M. & GROSSMANN, I. E. (1999) A branch and contract algorithm for problems with concave univariate, bilinear and linear fractional terms. *J. Glob. Optim.*, **14**, 217–249.

ZORN, K. & SAHINIDIS, N. V. (2014) Global optimization of general non-convex problems with intermediate bilinear substructures. *Optim. Methods Softw.*, **29**, 442–462.

Journal of Process Control 70 (2018) 80–95

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

Set-membership nonlinear regression approach to parameter estimation

Nikola D. Perić[a], Radoslav Paulen[c], Mario E. Villanueva[b], Benoît Chachuat[a,*]

[a] Centre for Process Systems Engineering, Department of Chemical Engineering, Imperial College London, UK
[b] School of Information Science and Technology, ShanghaiTech University, Shanghai, China
[c] Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Slovakia

ARTICLE INFO

Article history:
Received 5 September 2017
Received in revised form 2 April 2018
Accepted 9 April 2018

Keywords:
Parameter estimation
Nonlinear regression
Set-membership estimation
Statistical inference
Semi-infinite programming
Complete-search methods

ABSTRACT

This paper introduces *set-membership nonlinear regression* (SMR), a new approach to nonlinear regression under uncertainty. The problem is to determine the subregion in parameter space enclosing all (global) solutions to a nonlinear regression problem in the presence of bounded uncertainty on the observed variables. Our focus is on nonlinear algebraic models. We investigate the connections of SMR with (i) the classical statistical inference methods, and (ii) the usual set-membership estimation approach where the model predictions are constrained within bounded measurement errors. We also develop a computational framework to describe tight enclosures of the SMR regions using semi-infinite programming and complete-search methods, in the form of likelihood contour and polyhedral enclosures. The case study of a parameter estimation problem in microbial growth is presented to illustrate various theoretical and computational aspects of the SMR approach.

1. Introduction

Mathematical models capable of accurate prediction of physical phenomena have proved to be invaluable tools for engineers and scientists. In the area of process systems engineering, they routinely support the design, control and optimization of production processes, as a means of improving their economical profitability and reducing their environmental footprint. A majority of these models are nonlinear and contain adjustable parameters that need estimating from available experimental data, or else from other, more fundamental, mathematical descriptions. In this context, parameter estimation turns out to be a key step in the verification, and subsequent use, of the mathematical models.

Most commonly, parameter estimation in nonlinear models is cast as a nonlinear regression exercise, where selected parameter values are adjusted so that the model predictions match the available observations as close as possible, for instance in the least-squares or maximum-likelihood sense [1–4]. In order to avoid for the resulting parameter estimates to be biased, one can account for measurement errors in all of the variables, both independent and dependent variable observations, by following the so-called errors-in-variables approach [5,6]. This problem has been widely studied from a computational standpoint over the past decades, including the development of rigorous global optimization approaches for overcoming convergence to local optima [7,8].

Of course, there is more to model identification than just determining values for the unknown parameters. Systematic procedures have been devised to support the development and statistical verification of process models, which include testing structural identifiability, designing experiments for improved parameter precision, and inferring parameter confidence [9–12]. The focus in this paper is on the latter aspect, namely characterizing subregions in parameter space wherein the parameter values can be expected to lie. Other applications of such parameter confidence regions are in design under uncertainty [13,14], robust model predictive control [15–17], robust monitoring [18,19], and robust optimal design of experiments [20–22], to name but a few. For the scope of this paper, the emphasis is on models described by algebraic equations, but these ideas can be extended to dynamic or distributed models described by differential equations too.

Accounting for model mismatch and uncertain observations within the regression problem has spawned several schools of thought. Statistical approaches can be broadly classified as *frequentist* or *Bayesian*. The former seek to determine confidence regions around the regressed parameter values, typically a maximum-likelihood estimate, considered as the 'true' parameter values [1,2,4]. By construction, a $100(1-\alpha)\%$ frequentist confidence region comprises $100(1-\alpha)\%$ of the parameter values that would

* Corresponding author.
E-mail address: b.chachuat@imperial.ac.uk (B. Chachuat).

https://doi.org/10.1016/j.jprocont.2018.04.002

be obtained upon repetition of the parameter estimation using (hypothetical) new observations, considered as random variables. Approximate confidence regions, for instance based on the Wald test or the likelihood-ratio (LR) test, are known to converge to the exact confidence region in the limit of an infinite number of observations under certain conditions. Process modeling environments such as gPROMS and Aspen Custom Modeler have been relying on linear approximation and the Wald test to determine ellipsoidal confidence regions, a computationally efficient procedure for problems having several dozen unknown parameters, but one which may produce inaccurate results with large measurement errors and model mismatch or few measurement points. Confidence regions based on the LR test have been shown to yield superior approximations, but are computationally more involved since the corresponding parameter regions are complex sets in general (e.g., nonconvex, not simply connected) [23,24].

In practice, the term $100(1-\alpha)\%$ confidence region is often misused to refer to the range of parameter values that include $100(1-\alpha)\%$ of their probability distribution [25]. This description corresponds to so-called $100\,(1-\alpha)\%$ credible regions instead, which are defined in the Bayesian inference approach [26]. Bayesian estimation uses the available observations to construct a probability distribution of the parameters, called posterior distribution, based on a likelihood function and a prior probability distribution of the same parameters. In essence, this approach thus considers the unknown parameter values as random variables. Sampling-based techniques such as Markov-Chain Monte-Carlo (MCMC) [27,28] provide a means of constructing (approximate) credible regions, although the computational effort can become prohibitive for problems having upwards of 10 parameters [29]. A most probable estimate can be determined from the posterior distribution, which also corresponds to a maximum-likelihood estimate for a flat prior. Albeit classical frequentist and Bayesian inference regions can be reconciled in special cases, no equivalence can be drawn in general since Bayesian inference incorporates problem specific contextual information from the prior distribution, whereas frequentist inference is solely based on the data; see, e.g., [30, Chapter 5]. The debate on whether to use frequentist or Bayesian statistical inference continues to this day [25,31], but its intricacies are beyond the scope of this paper.

Regardless of whether a mathematical model's structure is correct or not, a frequentist confidence region will normally converge to the maximum- likelihood estimate as the number of observations increases. Likewise, a Bayesian posterior will normally converge to a point mass that corresponds to a most probable estimate, i.e., a point that maximizes the probability of the data given the (possibly wrong) model. An interesting alternative to these statistical approaches is *set-membership* estimation (SME). The traditional SME setting, also called guaranteed parameter estimation (GPE), seeks to determine the set of all possible parameter values for which a model's predictions are consistent with a set of observations subject to bounded errors [32–34]. The fact that this approach does not require a statistical description of the observation errors, solely bounds, is not only less demanding, but also more realistic in many practical applications, including biological systems where the measurements are often scarce and subject to large errors [21]. Beside parameter estimation, the distinctive yes-or-no answer provided by set-membership techniques can also be used for model inconsistency detection [35,36]. One caveat here is that the set of feasible parameter values may be empty in the presence of measurement outliers or due to an inadequate description of the measurement noise, thus calling for remedial strategies [37,38]. Another key challenge in nonlinear set-membership estimation is describing the feasible parameter set accurately, while remaining computationally tractable. This challenge is in fact similar to the one faced by aforementioned statistical inference methods for describing parameter confidence sets, and it may explain why set-membership estimation has not reached a wider diffusion to this day. Existing computational strategies are limited to problem with downwards of a dozen parameters. They range from approximation using sampling-based methods, including stochastic search [39], support vector machines (SVM) [40] and MCMC [41]; to rigorous complete-search methods based on interval analysis and other set arithmetics [42–44]; and to semidefinite relaxation techniques for semi-algebraic problems [45,46].

This paper introduces *set-membership regression* (SMR), a new approach to nonlinear regression. The SMR problem seeks to determine the subregion in parameter space enclosing all (global) solutions to a nonlinear regression problem in the presence of bounded uncertainty on the observed variables. By contrast with the traditional SME setting seeking for parameter values to satisfy certain feasibility constraints, the SMR approach method seeks for parameter values to satisfy an optimality condition. To the best knowledge of the authors, this problem has not been investigated in the general nonlinear setting so far. Milanese [47] studied optimality and convergence properties of least-squares estimates in the presence of unknown bounded disturbance, but their theoretical work is limited to linear problems. This paper sets out to investigate the connections of SMR with both statistical inference and set-membership estimation approaches for nonlinear algebraic models. Another principal contribution is a computational framework to describe tight enclosures of the SMR regions using complete-search methods.

The rest of the paper is organized as follows. Section 2 starts by reviewing classical results from both areas of statistical and set-membership estimation. Section 3 introduces the SMR approach and analyzes its properties, after which numerical solution strategies are developed in Section 4. A simple case study is used throughout Sections 2–4 to illustrate the main concepts and results. Section 5 presents a more challenging estimation problem in microbial growth to demonstrate the SMR approach. Finally, Section 6 concludes the paper and discusses future research opportunities.

2. Background

Our focus throughout this paper is on explicit models in the form

$$\mathbf{y} = \mathbf{g}(\mathbf{p}, \mathbf{u}),$$

where $\mathbf{p} \in \mathbb{R}^{n_p}$ is the vector of unknown parameters; and $(\mathbf{u}, \mathbf{y}) \in \mathbb{R}^{n_u} \times \mathbb{R}^{n_y}$ is the vector of observed variables, denoted collectively by $\mathbf{x}{:=}(\mathbf{u}, \mathbf{y}) \in \mathbb{R}^{n_x}$ for convenience. Notice that $\mathbf{u}$ and $\mathbf{y}$ often correspond to (either controlled or uncontrolled) input and output variables, respectively, in a practical setup. It is also worth pointing out that many of the concepts and methods presented herein can be applied to models described by implicit equation systems, such as $\mathbf{f}(\mathbf{p}, \mathbf{x}) = \mathbf{0}$, and models comprised of differential equations too.

Suppose that n_m observations $\mathbf{x}_k^{\mathrm{m}}{:=}(\mathbf{u}_k^{\mathrm{m}}, \mathbf{y}_k^{\mathrm{m}})$ of the input–output variables are available, and assume that all of these observation errors are independent and described by the probability density functions $p(\cdot \mid \boldsymbol{\psi})$ parameterized by $\boldsymbol{\psi}$. In the error-in-variables approach [6], the reconciled values $\mathbf{u}_1, \ldots, \mathbf{u}_{n_m}$ for the observations are estimated alongside the unknown model parameters $\mathbf{p}$. The joint probability of the prediction-observation mismatch in all data points for the parameter values $\boldsymbol{\theta}{:=}(\mathbf{p}, \mathbf{u}_1, \ldots, \mathbf{u}_{n_m}) \in \mathbb{R}^{n_\theta}$ is described by the following likelihood function:

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^{\mathrm{m}}){:=}\prod_{k=1}^{n_m} p\left(\boldsymbol{\delta}\mathbf{u}_k \middle| \boldsymbol{\psi}_{\mathbf{u}_k}\right) \prod_{k=1}^{n_m} p\left(\boldsymbol{\delta}\mathbf{y}_k \middle| \boldsymbol{\psi}_{\mathbf{y}_k}\right), \tag{1}$$

with $\boldsymbol{\delta}\mathbf{u}_k{:=}\mathbf{u}_k - \mathbf{u}_k^{\mathrm{m}}$ and $\boldsymbol{\delta}\mathbf{y}_k{:=}\mathbf{g}(\mathbf{p}, \mathbf{u}_k) - \mathbf{y}_k^{\mathrm{m}}$. The error-in-equation approach instead, considers the input measurements $\mathbf{u}_k^{\mathrm{m}}$ to be

error-free; that is, the parameter vector $\boldsymbol{\theta}$ reduces to $\mathbf{p}$, and the likelihood function simplifies to

$$\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^{\mathrm{m}}) := \prod_{k=1}^{n_m} p\left(\boldsymbol{\delta}\mathbf{y}_k \middle| \boldsymbol{\psi}_{\mathbf{y}_k}\right). \tag{2}$$

Nonlinear regression in the maximum-likelihood sense seeks to determine values for $\boldsymbol{\theta}$ in order to maximize $\mathcal{L}$ or, equivalently, maximize $\log \mathcal{L}$. In the error-in-variables approach, this estimation entails the solution of an optimization problem in the form of

$$\hat{\boldsymbol{\theta}} \in \underset{\mathbf{p},\mathbf{u}_1,\ldots,\mathbf{u}_{n_m}}{\operatorname{argmax}} \sum_{k=1}^{n_m} \log p(\boldsymbol{\delta}\mathbf{u}_k|\boldsymbol{\psi}_{\mathbf{u}_k}) + \log p(\boldsymbol{\delta}\mathbf{y}_k|\boldsymbol{\psi}_{\mathbf{y}_k}). \tag{3}$$

If the parameters $\boldsymbol{\psi}$ describing the error distribution are also unknown, one may either approximate their values using an ad hoc estimator, or consider them as additional variables in the problem (3) [1].

In the special case of Gaussian-distributed errors, $p(\delta_{k,i} \mid v_{k,i}) = \frac{1}{\sqrt{2\pi v_{k,i}}} \exp\left(-\frac{\delta_{k,i}^2}{2v_{k,i}}\right)$ with zero mean and variance $v_{k,i}$, the maximum-likelihood problem (3) is equivalent to the following weighted least-squares problem

$$\hat{\boldsymbol{\theta}} \in \underset{\mathbf{p},\mathbf{u}_1,\ldots,\mathbf{u}_{n_m}}{\arg\min} \sum_{k=1}^{n_m} \left(\sum_{i=1}^{n_u} \frac{\left(\delta u_{k,i}\right)^2}{v_{u_k,i}} + \sum_{i=1}^{n_y} \frac{\left(\delta y_{k,i}\right)^2}{v_{y_k,i}} \right). \tag{4}$$

While least-squares (ℓ_2) regression is optimal amongst minimum-variance mean-unbiased estimators for normally distributed observation errors, outliers can greatly distort the least-squares estimates. As an alternative, least-absolute-values (ℓ_1) fitting may be preferable in the presence of outliers or if little is known about the distribution of the errors [48,49]. The ℓ_1 regression problem reads

$$\hat{\boldsymbol{\theta}} \in \underset{\mathbf{p},\mathbf{u}_1,\ldots,\mathbf{u}_{n_m}}{\arg\min} \sum_{k=1}^{n_m} \left(\sum_{i=1}^{n_u} \frac{\left|\delta u_{k,i}\right|}{v_{u_k,i}} + \sum_{i=1}^{n_y} \frac{\left|\delta y_{k,i}\right|}{v_{y_k,i}} \right), \tag{5}$$

where standard tricks can be used to reformulate or approximate the nonsmooth absolute value term in the objective function. The solutions to the ℓ_1 regression problem (5) can also be viewed as maximum-likelihood estimates if the observation errors follow the Laplacian distribution $p(\delta_{k,i} \mid v_{k,i}) = \frac{1}{\sqrt{2v_{k,i}}} \exp\left(-|\delta_{k,i}|\sqrt{\frac{2}{v_{k,i}}}\right)$ with zero mean and variance $v_{k,i}$. An ℓ_∞ regression problem can be constructed in a similar way [49].

2.1. Statistical inference

Classical frequentist confidence inference proceeds in two steps: (i) solve a regression problem, e.g., to determine a most-likely parameter estimate as described above; and (ii) construct confidence regions around this estimate.

Under the assumption that $\hat{\boldsymbol{\theta}}$ matches the (unique) 'true' value of the model parameters, both the *likelihood subset ratio statistic* $-2\log[\mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^{\mathrm{m}})/\mathcal{L}(\hat{\boldsymbol{\theta}} \mid \mathbf{x}^{\mathrm{m}})]$, and the *Wald subset statistic*[1] $(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T \mathbf{V}_{\hat{\boldsymbol{\theta}}}^{-1} (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})$, follow a chi-squared distribution with n_θ degrees of freedom with an increasing sample size $n_m \to \infty$ [4].

These asymptotic confidence results can be used to obtain (approximate) $100(1-\alpha)\%$ confidence regions, with the usual frequentist interpretation that the probability for a random confidence region to cover the true value of $\boldsymbol{\theta}$ is, in large samples, equal to $1-\alpha$ [24]:

- $100(1-\alpha)\%$ likelihood-based confidence region:

$$\Theta_{\mathrm{L}} := \left\{ \theta \in \Theta_0 \;\middle|\; -2\log\left(\frac{\mathcal{L}(\theta \mid \mathbf{x}^m)}{\mathcal{L}(\hat{\theta} \mid \mathbf{x}^m)}\right) \le \chi^2_{n_\theta}(1-\alpha) \right\} \tag{7}$$

- $100(1-\alpha)\%$ normal-theory (Wald) confidence region:

$$\Theta_{\mathrm{W}} := \left\{ \theta \in \Theta_0 \;\middle|\; (\theta-\hat{\theta})^T \mathbf{V}_{\hat{\theta}}^{-1} (\theta-\hat{\theta}) \le \chi^2_{n_\theta}(1-\alpha) \right\} \tag{8}$$

where $\Theta_0 \subseteq \mathbb{R}^{n_\theta}$ denotes the allowable (prior) parameter set; and $\chi^2_{n_\theta}(1-\alpha)$ is the $1-\alpha$ quantile of the chi-squared distribution with n_θ degrees of freedom. At this point, we note that confidence intervals can be inferred from any confidence region by bounding the range of values for each parameter θ_i. In the case of the Wald approximation, explicit confidence bounds are obtained as

$$\theta_i \in \left\{ \hat{\theta}_i \pm \sqrt{[\mathbf{V}_{\hat{\boldsymbol{\theta}}}]_{i,i} \chi^2_{n_\theta}(1-\alpha)} \right\}.$$

A classical result in statistical inference is that the confidence regions (7) and (8) are asymptotically equivalent [51,52], with a convergence rate $\propto n_m^{-1}$. However, unlike the likelihood-based confidence regions, the Wald confidence regions are not invariant to a model reparameterization because of the (approximate) covariance term $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$. Conversely, computing a Wald confidence region is straightforward, whereas describing a likelihood-based confidence region for a nonlinear model is generally a hard task since this region may not be convex or not even simply connected.

Unlike the frequentist view, Bayesian estimation treats the parameters as random variables, whose (posterior) probability distribution, $p(\boldsymbol{\theta} \mid \mathbf{x}^{\mathrm{m}})$ can be inferred from Bayes' theorem,

$$p(\boldsymbol{\theta} \mid \mathbf{x}^{\mathrm{m}}) \propto \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^{\mathrm{m}}) \;\; p(\boldsymbol{\theta}), \tag{9}$$

where $p(\boldsymbol{\theta})$ is the so-called prior density of the parameters. Any subset $\Theta_{\mathrm{B}} \subseteq \mathbb{R}^{n_\theta}$ such that

$$\int_{\Theta_{\mathrm{B}}} p(\boldsymbol{\theta}|\mathbf{x}^{\mathrm{m}}) = 1-\alpha \tag{10}$$

is called a $100(1-\alpha)\%$ credible set. One particular kind of credible sets is the *highest posterior density* (HPD) set, given by

$$\Theta_{\mathrm{B}} := \{\boldsymbol{\theta} \mid p(\boldsymbol{\theta}|\mathbf{x}^{\mathrm{m}}) \ge \pi_\alpha\}, \tag{11}$$

where π_α is the largest value for which (10) holds. When a sampling approach is applied to estimate the posterior, for instance a MCMC sampler, the value of π_α can be estimated from a procedure that examines all available samples of $p(\boldsymbol{\theta} \mid \mathbf{x}^{\mathrm{m}})$ [28]. It is also worth mentioning that complete-search approaches to enclosing credible sets have been proposed as well [53,54].

The connections between Bayesian and non-Bayesian statistical inference have been studied since the 1960s, for instance with regards to matching credible and confidence intervals [55,56]; or, more recently, in order to reconcile Bayesian and frequentist higher-order asymptotic expansions for predictive probability densities [57]. In linear regression problems with normally distributed measurement errors, the Bayesian posterior takes the form of a multivariate Gaussian centered at the maximum-likelihood estimate and with covariance matrix $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$ for non-informative priors, so the HPD credible regions match their frequentist counterparts. More generally, such matching can be made in cases where the Bayesian prior is invariant to model reparameterization, which is the case for Jeffreys or reference priors [58].

[1] The covariance matrix $\mathbf{V}_{\hat{\boldsymbol{\theta}}} \in \mathbb{S}_+^{n_\theta \times n_\theta}$ for the parameters at $\hat{\boldsymbol{\theta}}$ can be approximated in various ways [50], which are asymptotically equivalent; for instance [1, § 7-5],

$$\mathbf{V}_{\hat{\boldsymbol{\theta}}} := \hat{\mathcal{H}}^{-1} \frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{x}^{\mathrm{m}})}{\partial \hat{\mathbf{x}} \partial \hat{\mathbf{x}}} \mathbf{V}_{\mathbf{e}} \frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{x}^{\mathrm{m}})}{\partial \hat{\mathbf{x}} \partial \boldsymbol{\theta}} \hat{\mathcal{H}}^{-1}, \tag{6}$$

where $\mathbf{V}_{\mathbf{e}} \in \mathbb{S}_+^{n_x n_m \times n_x n_m}$ stands for the covariance matrix of the observation noise, and $\hat{\mathcal{H}} := \frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}}|\mathbf{x}^{\mathrm{m}})}{\partial \boldsymbol{\theta}^2}$ is the Hessian matrix at $\hat{\boldsymbol{\theta}}$.

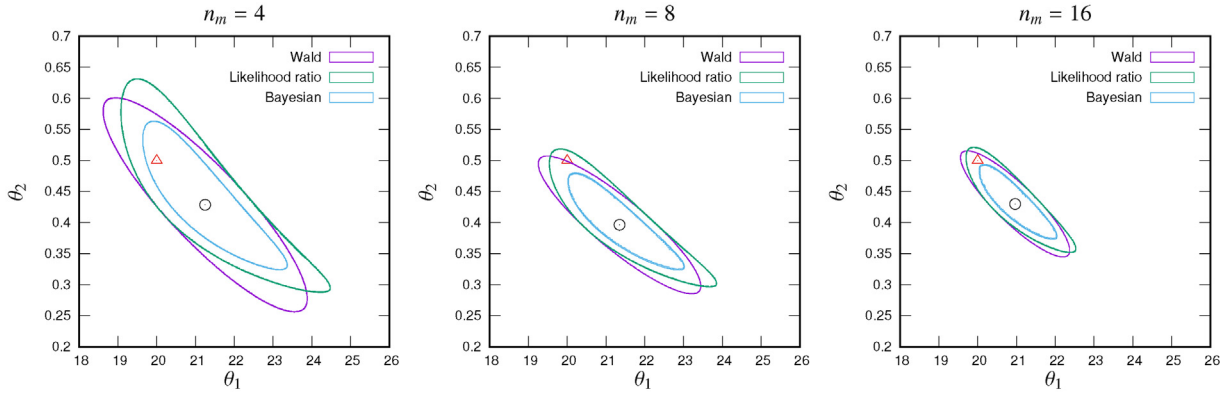


Fig. 1. 90% Wald and log-likelihood confidence regions and 90% HPD credible region for the BOD example with 4 (left), 8 (center), and 16 (right) measurement points. The circles and triangles represent the maximum-likelihood estimates and the real parameter values, respectively.

For simplicity, our focus in this paper is limited to uniform prior distributions with compact supports. Although such priors fail to be invariant under reparameterization, the resulting HPD sets correspond to contour levels of the likelihood function, similar to the likelihood-based confidence regions.

2.2. Set-membership estimation

The usual GPE problem in set-membership estimation seeks to determine a parameter subregion such that the predicted input–output observations are consistent with their matching measurements within given error bounds [32,33],

$$\Theta_G := \left\{ \boldsymbol{\theta} \in \Theta_0 \,\middle|\, \left(\boldsymbol{\delta u}_k, \boldsymbol{\delta y}_k\right)_{1\le k\le n_m} \in \mathbf{E} \right\}. \tag{12}$$

Here the error set $\mathbf{E} \subset \mathbb{R}^{n_x n_m}$ may be any compact set and does not need a statistical description of the uncertainty. In the usual scenario where independent error bounds $\pm\mathbf{e}_{\mathbf{u}_1}, \pm\mathbf{e}_{\mathbf{y}_1}, \ldots, \pm\mathbf{e}_{\mathbf{u}_{n_m}}, \pm\mathbf{e}_{\mathbf{y}_{n_m}}$ are given for each of the measurements, the set-membership estimation problem reads

$$\Theta_G := \left\{ \boldsymbol{\theta} \in \Theta_0 \,\middle|\, \begin{array}{l} \forall k = 1, \ldots, n_m, \\ -\mathbf{e}_{\mathbf{u}_k} \le \boldsymbol{\delta u}_k \le \mathbf{e}_{\mathbf{u}_k}, \quad -\mathbf{e}_{\mathbf{y}_k} \le \boldsymbol{\delta y}_k \le \mathbf{e}_{\mathbf{y}_k} \end{array} \right\}.$$

If statistical information about the observation error is nonetheless available, for instance a uniform or q-Gaussian probability distribution with compact support, one may take **E** directly as this support set. Even when the distribution support is not compact, one could decide to exclude those scenarios having a probability lower than a given threshold and use the corresponding HPD credible region as the error set **E**; see, e.g., [59].

It is not difficult to imagine a situation whereby no parameter value in Θ_0 can be found such that the model predictions are consistent with the observations for a given error set **E**, i.e., the guaranteed parameter region (12) is empty. This may happen in the presence of measurement outliers, or could be caused by a large model mismatch. The former situation is common with experimental data, e.g., due to a failing or drifting sensor. Methods have been developed for robustifying set-membership estimation against outliers [37,38], alongside classical approaches to detecting outliers [60]. Moreover, one can take advantage of the latter situation, for instance to invalidate candidate models that would present a systematic offset with a certain set of observations [35,36], typically after checking for outliers [38]. Another appeal of set-membership estimation lies in its ability to detect a lack of identifiability in parametric models, that is, when model responses corresponding to distinct parameter values are indistinguishable [9].

The vast majority of computational studies in set- membership estimation uses exhaustive-search techniques based on interval analysis or other set arithmetics to describe the parameter regions (12) [42–44,61] . A current bottleneck of these approaches is their applicability to problems having no more than 5–10 parameters. However, if one is ready to abandon guarantees, sampling-based techniques such as SVM or MCMC can be used to approximate the parameter regions, and these remain applicable for black-box models too [40,41].

Illustrative example. We use a simple estimation problem adapted from [3] to illustrate the main approaches described in this background section, and we use the same problem to illustrate the main properties of the SMR framework developed later on in Sections 3 and 4. The model describes the dynamic evolution of biological oxygen demand (BOD), c in a wastewater sample,

$$c = \theta_1(1 - \mathrm{e}^{-\theta_2 t}), \tag{13}$$

with parameters $(\theta_1, \theta_2) \in [0, 50] \times [0, 2]$, and time $t \ge 0$. For this problem, data points $(t_k^{\mathrm{m}}, c_k^{\mathrm{m}})$ have been generated by simulating the model (13) for the parameter values $\theta_1 = 20$ and $\theta_2 = 0.5$, and corrupting these values with a Gaussian white noise with variance $\sigma_c^2 = 1$. These data are reported in Appendix B for the sake of reproducibility.

Both 90% confidence regions and 90% HPD credible regions are compared in Fig. 1, in the case of an ℓ_2-regression problem. Various sets of measurements are considered, namely $n_m = 4$ measurement points (every other day), 8 measurement points (every day), and 16 measurement points (twice a day). The asymptotic convergence of the Wald and likelihood-based confidence regions with an increasing number of measurements is clearly visible. The HPD credible sets shown on these plots are generated from a flat prior, and are consistently smaller than their confidence counterparts; HPD credible sets constructed from a non-informative Jeffreys prior (not shown on the plots) would be identical to the likelihood-based confidence regions.

A comparison between guaranteed parameter regions for the same three sets of measurements, but corresponding to different measurement error sets in (12), is shown in Fig. 2. The first measurement error set corresponds to the usual assumption of independent error bounds on each measurement,

$$\mathbf{E}_1 := \left\{ \mathbf{e}_c \in \mathbb{R}^{n_m} \mid \ \forall k = 1 \ldots n_m, \quad e_{c,k}^2 \le \chi_1^2(0.9)\sigma_c^2 \right\}, \tag{14}$$

here for 90% confidence bounds, so that $\chi_1^2(0.9)\sigma_c^2 \approx 2.706$. Notice how the corresponding guaranteed parameter sets shrink when more measurements are added, as it becomes more challenging for the model predictions to match a larger measurement set in the presence of measurement noise. Such guaranteed parameter

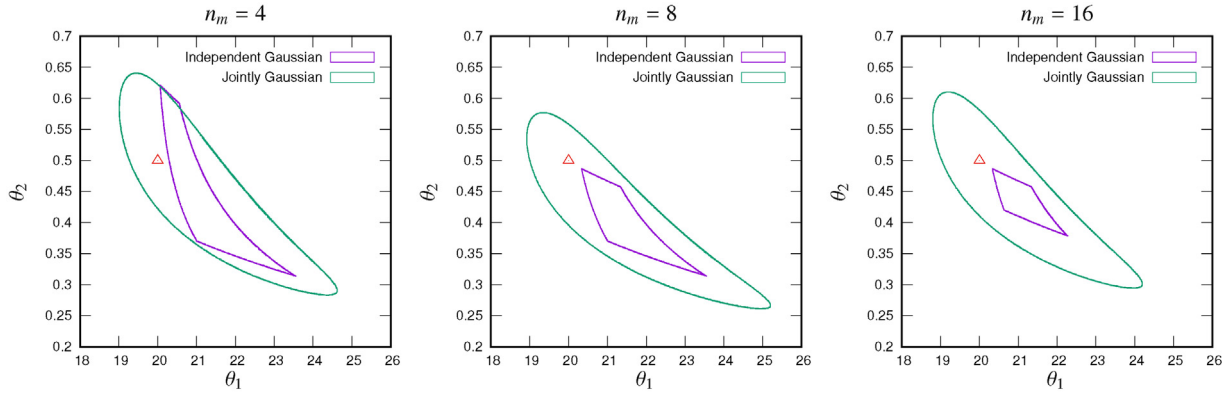


Fig. 2. Guaranteed parameter regions for the BOD example with 4 (left), 8 (center), and 16 (right) measurement points. The regions are for two measurement error sets corresponding to 90% HPD regions on either independent or jointly Gaussian distributions. The triangles represent the real parameter values.

regions could even be empty, which happens for instance with $e_{c_k}^2 \leq 1$ in (14), corresponding to 68% (1-sigma) confidence bounds. Also notice that the real parameter value (20, 0.5) lies outside the guaranteed regions due to the large measurement noise.

The other measurement error set in Fig. 2 is chosen as the HPD set of a joint Gaussian distribution,

$$\mathbf{E}_2 := \left\{ \mathbf{e}_c \in \mathbb{R}^{n_m} \,\middle|\, \mathbf{e}_c^T \mathbf{e}_c \leq \chi^2_{n_m}(0.9)\sigma_c^2 \right\}, \tag{15}$$

again for a 90% confidence limit. Guaranteed parameter sets so constructed do not shrink significantly as more measurements are added into the estimation problem, and they are thus more resilient to measurement noise than their counterpart sets constructed with independent error bounds on each measurement. This higher resilience is essentially due to an enlarged, and hence more flexible, measurement error set $\mathbf{E}_2$ compared to $\mathbf{E}_1$.

3. Set-membership nonlinear regression

The developed *set-membership regression* (SMR) approach seeks to describe the subregion Θ_R in parameter space enclosing *all* (global) solutions to a nonlinear regression problem under *all* possible measurement uncertainty scenarios. Given a bounded uncertainty set $\mathbf{E} \subset \mathbb{R}^{n_x n_m}$ on the observation errors, the SMR region Θ_R is mathematically defined as

$$\Theta_R := \{ \boldsymbol{\theta} \in \Theta_0 \,|\, \boldsymbol{\theta} \in \underset{\boldsymbol{\varpi}}{arg\,\max} \log \mathcal{L}(\boldsymbol{\varpi} \mid \mathbf{x}^m + \mathbf{e}) \;\; \exists \mathbf{e} \in \mathbf{E} : \}. \tag{16}$$

In the context of the ℓ_2-regression problem (4), SMR specializes to

$$\Theta_R^{\ell_2} := \left\{ \boldsymbol{\theta} \in \Theta_0 \;\middle|\; \begin{array}{l} \exists \left(\mathbf{e}_{\mathbf{u}_1}, \mathbf{e}_{\mathbf{y}_1}, \ldots, \mathbf{e}_{\mathbf{u}_{n_m}}, \mathbf{e}_{\mathbf{y}_{n_m}} \right) \in \mathbf{E} : \\ \boldsymbol{\theta} \in \underset{\mathbf{p}, \mathbf{u}_1, \ldots, \mathbf{u}_{n_m}}{\arg\min} \sum_{k=1}^{n_m} \left(\sum_{i=1}^{n_u} \frac{[\delta u_{k,i} - e_{u_k,i}]^2}{v_{u_k,i}} \right. \\ \left. + \sum_{i=1}^{n_y} \frac{[\delta y_{k,i} - e_{y_k,i}]^2}{v_{y_k,i}} \right) \end{array} \right\}, \tag{17}$$

and in the context of the ℓ_1- regression problem (5), to

$$\Theta_R^{\ell_1} := \left\{ \boldsymbol{\theta} \in \Theta_0 \;\middle|\; \begin{array}{l} \exists (\mathbf{e}_{\mathbf{u}_1}, \mathbf{e}_{\mathbf{y}_1}, \ldots, \mathbf{e}_{\mathbf{u}_{n_m}}, \mathbf{e}_{\mathbf{y}_{n_m}}) \in \mathbf{E} : \\ \boldsymbol{\theta} \in \underset{\mathbf{p}, \mathbf{u}_1, \ldots, \mathbf{u}_{n_m}}{\arg\min} \sum_{k=1}^{n_m} \left(\sum_{i=1}^{n_u} \frac{|\delta u_{k,i} - e_{u_k,i}|}{v_{u_k,i}} \right. \\ \left. + \sum_{i=1}^{n_y} \frac{|\delta y_{k,i} - e_{y_k,i}|}{v_{y_k,i}} \right) \end{array} \right\}. \tag{18}$$

Notice that the constraint feasibility condition in the traditional SME formulation (12) is replaced with an optimality condition in the SMR problem (16) , making the parameter regions in SMR expectedly more difficult to characterize. Numerical solution strategies for describing enclosures of an SMR region are presented later on in Section 4. The remainder of this section investigates connections between SMR and the well-established set-membership and statistical inference approaches, respectively in Sections 3.1 and 3.2.

3.1. Set-membership interpretation

By contrast with the usual approach to set-membership estimation (Section 2.2), SMR comes with a guarantee that the set Θ_R is always non-empty, no matter how large the model mismatch or the observation errors might be, since the regression problems in (16) are all feasible by construction. Therefore, the SMR formulation is inherently resilient to the presence of outlying observations, and it does not need for such outliers to be detected or removed from the observation set before computing the parameter regions [38]. In other words, the outlying observations can be dealt with directly into the SMR problem (16) via an appropriate likelihood function.

The following inclusion result holds between SMR and GPE under mild assumptions:

Theorem 1. *Suppose that the probability density functions $p(\cdot|\boldsymbol{\psi})$ participating in the likelihood function (1) are all maximal at 0. Then, for a given error set* $\mathbf{E}$*, the SMR region (16) contains the GPE region (12),* $\Theta_G \subseteq \Theta_R$.

Proof. Let $\bar{\boldsymbol{\theta}} \in \Theta_G$, so that $(\bar{\mathbf{u}}_1 - \mathbf{u}_1^m, \mathbf{g}(\mathbf{p}, \bar{\mathbf{u}}_1) - \mathbf{y}_1^m, \ldots, \bar{\mathbf{u}}_{n_m} - \mathbf{u}_{n_m}^m, \mathbf{g}(\mathbf{p}, \bar{\mathbf{u}}_{n_m}) - \mathbf{y}_{n_m}^m) \in \mathbf{E}$. It follows that $(\bar{\mathbf{u}}_1 - \mathbf{u}_1^m - \mathbf{e}_{\mathbf{u}_1}, \mathbf{g}(\mathbf{p}, \bar{\mathbf{u}}_1) - \mathbf{y}_1^m - \mathbf{e}_{\mathbf{y}_1}, \ldots, \bar{\mathbf{u}}_{n_m} - \mathbf{u}_{n_m}^m - \mathbf{e}_{\mathbf{u}_{n_m}}, \mathbf{g}(\mathbf{p}, \bar{\mathbf{u}}_{n_m}) - \mathbf{y}_{n_m}^m - \mathbf{e}_{\mathbf{u}_{n_m}}) = 0$ for some $\mathbf{e} := (\mathbf{e}_{\mathbf{u}_1}, \mathbf{e}_{\mathbf{y}_1}, \ldots, \mathbf{e}_{\mathbf{u}_{n_m}}, \mathbf{e}_{\mathbf{y}_{n_m}}) \in \mathbf{E}$. Since the probability density functions $p(\cdot|\boldsymbol{\psi})$ in $\mathcal{L}$ are all maximal at 0 by assumption, the log-likelihood function $\log \mathcal{L}(\cdot \mid \mathbf{x}^m + \mathbf{e})$ is (globally) maximal at $\bar{\boldsymbol{\theta}}$, and therefore $\bar{\boldsymbol{\theta}} \in \Theta_R$. □

Remark 1. The assumption on the likelihood function $\mathcal{L}$ in Theorem 1 is not very restrictive in practice. For instance, it is satisfied by both ℓ_2- and ℓ_1-regression problems in (17) and (18), so we have $\Theta_G \subseteq \Theta_R^{\ell_2}$ and $\Theta_G \subseteq \Theta_R^{\ell_1}$. It is also satisfied when the probability density functions are uniform on a compact support, as is the case with ℓ_∞-regression problems [49].

Illustrative example (continued). A comparison between GPE and SMR regions for both ℓ_1- and ℓ_2- regression is presented in Fig. 3, in the case of 8 measurements. The same measurement error sets $\mathbf{E}_1$ and $\mathbf{E}_2$ as introduced earlier in (14) and (15) are used in this comparison. For simplicity, we have applied a simple sampling procedure

to inner-approximate the SMR regions: 20,000 error vectors $\mathbf{e}_c^{(i)}$ are generated within the multi- dimensional error sets $\mathbf{E}_1$ and $\mathbf{E}_2$, here using Sobol quasi-random sampling; then, the following nonlinear regression problem is solved to global optimality to obtain a corresponding point $\boldsymbol{\theta}^{(i)} \in \Theta_R$,

$$\min_{\theta_1,\theta_2} \sum_{k=1}^{n_m} \frac{[c_k^m + e_{c,k}^{(i)} - \theta_1(1 - e^{-\theta_2 t_k^m})]^2}{\sigma_c^2}.$$

We start by noting that the inclusion result in Theorem 1 is indeed satisfied for both measurement error sets and both regression types. Moreover, the SMR regions obtained for either measurement error sets are comparable in size. In the case of independent error bounds on the measurements (set $\mathbf{E}_1$, left plot), the SMR regions do not shrink much when more measurements are added, which is unlike the corresponding GPE regions; compare Fig. 2. This also illustrates the higher resilience of SMR to noisy or outlying measurements than GPE. For both measurement error sets, the SMR-ℓ_2 regions are consistently smaller than their SMR-ℓ_1 counterparts. Interestingly, this observation is consistent with the classical Gauss-Markov theorem stating that the least-squares estimator provides the estimator with lowest variance in linear regression.

3.2. Statistical interpretation

Whenever statistical information is available for the observation errors, for instance in the form of a joint probability distribution, one may choose the error set $\mathbf{E}$ as the corresponding HPD region for a given credibility level $1-\alpha$. In the case of independent and Gaussian-distributed observation errors, such as those leading to the ℓ_2- regression problem (4), the $100(1-\alpha)$% HPD region is given by

$$\mathbf{E} := \left\{ \mathbf{e} \in \mathbb{R}^{n_x n_m} \;\middle|\; \mathbf{e}^T \mathbf{V}_\mathbf{e}^{-1} \mathbf{e} = \left\| \mathbf{V}_\mathbf{e}^{-1/2} \mathbf{e} \right\|_2^2 \leq \chi^2_{n_x n_m}(1-\alpha) \right\}, \tag{19}$$

with the diagonal error covariance matrix $\mathbf{V}_\mathbf{e} := \mathrm{diag}(\mathbf{v}_{\mathbf{u}_1}, \mathbf{v}_{\mathbf{y}_1}, \ldots, \mathbf{v}_{\mathbf{u}_{n_m}}, \mathbf{v}_{\mathbf{y}_{n_m}})$. Likewise, for Laplacian distributed errors as in the ℓ_1-regression problem (5), the $100(1-\alpha)$% HPD region comes in the form

$$\mathbf{E} := \left\{ \mathbf{e} \in \mathbb{R}^{n_x n_m} \;\middle|\; \left\| \mathbf{V}_\mathbf{e}^{-1/2} \mathbf{e} \right\|_1^2 \leq \Gamma_{n_x n_m}(1-\alpha) \right\}, \tag{20}$$

where $\Gamma_{n_x n_m}(1-\alpha)$ is the counterpart of the chi-squared value for a joint Laplacian distribution.

Notice that with the error sets in (19) and (20), the SMR regions Θ_R may not converge to a singleton (or a finite set) as more observations are added into the regression problem, since the HPD limits $\chi^2_{n_x n_m}(1-\alpha)$ and $\Gamma_{n_x n_m}(1-\alpha)$ are themselves increasing with n_m for a given confidence level $1-\alpha$. The SMR regions derived from such error sets are thus unrelated to their confidence and credible region counterparts in classical statistical inference (Section 2.1), which are both shrinking to a singleton as $n_m \to \infty$ (under certain regularity conditions). But while one would indeed expect convergence to some 'true' parameter value when a model's structure is correct, such an idea of 'true' parameter values becomes meaningless in the presence of structural model mismatch. By contrast, SMR does not make any assumption about the correctness of a model's structure, and a $100(1-\alpha)$% SMR region is comprised of those parameter values which are equally credible under the observation error set $\mathbf{E}$, in the sense of the regression problem at hand: a clear and unambiguous statistical interpretation.

To sum up, convergence of an SMR region Θ_R to a singleton is dependent on the choice of the measurement error set $\mathbf{E}$, but is unrelated to whether or not the model's structure is correct. A follow-up question then is identifying scenarios under which SMR regions would be asymptotically equivalent to classical confidence regions. The following result establishes one simple connection with the Wald confidence regions (8) under certain regularity conditions.

Theorem 2. *Let the error set in the SMR problem (16) be given by*

$$\mathbf{E} := \left\{ \mathbf{e} \in \mathbb{R}^{n_x n_m} \;\middle|\; \mathbf{e}^T \mathbf{V}_\mathbf{e}^{-1} \mathbf{e} \leq \chi^2_{n_\theta}(1-\alpha) \right\}, \tag{21}$$

for some confidence level $1-\alpha$, and covariance matrix $\mathbf{V}_\mathbf{e} \in \mathbb{S}_+^{n_x n_m \times n_x n_m}$. Assume that the likelihood function in (16) is twice continuously differentiable and the regression problems for $\mathbf{e} \in \mathbf{E}$ all have a unique, strict global optimum. Then, the SMR region Θ_R is asymptotically equivalent to the $100(1-\alpha)$% Wald confidence region Θ_W in (6) and (8),

$$d_H(\Theta_R, \Theta_W) \in \mathcal{O}(\mathrm{diam}(\mathbf{E})^2),$$

where d_H is the Hausdorff metric.

Proof. Let $\mathbf{e} \in \mathbf{E}$, and denote by $\bar{\boldsymbol{\theta}}(\mathbf{e}) \in \Theta_R$ the corresponding solution to the regression problem $\max_{\boldsymbol{\theta}} \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m + \mathbf{e})$, so that $\frac{\partial \log \mathcal{L}}{\partial \boldsymbol{\theta}}(\bar{\boldsymbol{\theta}}(\mathbf{e}) \mid \mathbf{x}^m + \mathbf{e}) = \mathbf{0}$. Since we also have $\frac{\partial \log \mathcal{L}}{\partial \boldsymbol{\theta}}(\hat{\boldsymbol{\theta}} \mid \mathbf{x}^m) = \mathbf{0}$ at the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}$, it follows by Taylor's theorem and the regularity assumptions that

$$\bar{\boldsymbol{\theta}}(\mathbf{e}) \in \hat{\boldsymbol{\theta}} - \hat{\mathcal{H}}^{-1} \frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}} \mid \mathbf{x}^m)}{\partial \boldsymbol{\theta} \partial \mathbf{x}} \mathbf{e} + \mathcal{O}\left(\|\mathbf{e}\|^2 \right), \tag{22}$$

with $\hat{\mathcal{H}} := \frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}} \mid \mathbf{x}^m)}{\partial \boldsymbol{\theta}^2}$. Now, let $\boldsymbol{\theta}$ be any point in Θ_R. From (22), we have

$$\boldsymbol{\theta}' = \hat{\boldsymbol{\theta}} - \hat{\mathcal{H}}^{-1} \frac{\partial^2 \log \mathcal{L}(\hat{\boldsymbol{\theta}} \mid \mathbf{x}^m)}{\partial \boldsymbol{\theta} \partial \mathbf{x}} \mathbf{e}, \tag{23}$$

for some $\boldsymbol{\theta}' \in \Theta_0$ with $\left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\| \in \mathcal{O}(\mathrm{diam}(\mathbf{E})^2)$. The image of the error set (21) under the affine transformation (23) is an ellipsoid with center $\hat{\boldsymbol{\theta}}$ and shape matrix $\mathbf{V}_{\hat{\boldsymbol{\theta}}}$ as in (6), so that $\boldsymbol{\theta}' \in \Theta_W$. Conversely, let $\boldsymbol{\theta}'$ be any point in Θ_W, and let $\mathbf{e}$ be any point in $\mathbf{E}$ satisfying (23). Clearly, the point $\bar{\boldsymbol{\theta}}(\mathbf{e}) \in \Theta_R$ is such that $\left\| \bar{\boldsymbol{\theta}}(\mathbf{e}) - \boldsymbol{\theta}' \right\| \in \mathcal{O}(\mathrm{diam}(\mathbf{E})^2)$ by (22). □

Remark 2. In the special case of a linear regression, the equivalence between the SMR and Wald confidence regions in Theorem 2 turns out to be exact, not merely asymptotic. For an ℓ_2-regression and the model $\mathbf{y} = \mathbf{F}\boldsymbol{\theta}$, we have

$$\Theta_R^{\ell_2} = \left\{ \boldsymbol{\theta} \in \Theta_0 \;\middle|\; \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right)^T \mathbf{F}^T \mathbf{V}_\mathbf{e}^{-1} \mathbf{F} \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}} \right) \leq \chi^2_{n_\theta}(1-\alpha) \right\},$$

which matches the likelihood-ratio confidence region Θ_L (2), as well as the Bayesian's HPD credible region Θ_B (11) for a uniform/non-informative prior. Both the frequentist and Bayesian inference regions are thus implied by the SMR framework in linear regression problems.

Remark 3. The key difference between the error set (21) in Theorem 2 and the $100(1-\alpha)$%-HPD region (19), is that the HPD limit in the former, namely $\chi^2_{n_\theta}(1-\alpha)$, is independent of the number of observations. This is also the reason why the error set (21) shrinks to the origin, and therefore Θ_R converges to the singleton set $\{\hat{\boldsymbol{\theta}}\}$ as $n_m \to \infty$ (under the assumptions of Theorem 2). Conversely, a $100(1-\alpha)$% confidence region may be regarded as the asymptotic equivalent to an SMR region with the confidence level $100(1-\beta)$% on the jointly Gaussian-distributed observation errors in (19) such that $\chi^2_{n_x n_m}(1-\beta) = \chi^2_{n_\theta}(1-\alpha)$. For instance, a 90%- confidence region in a two-parameter regression problem is asymptotically equivalent to an SMR region with 67%, 20% and 0.26% joint confidence for 4, 8 and 16 observations, respectively.

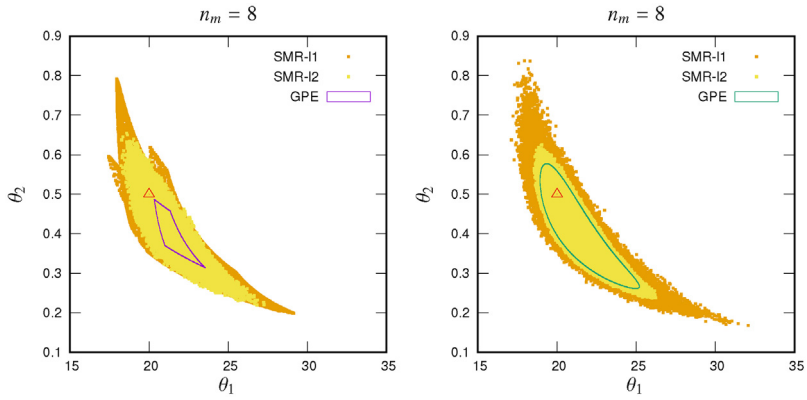


Fig. 3. Guaranteed parameter regions compared with sampled SMR-ℓ_1 and SMR-ℓ_2 regions for the BOD example with 8 measurement points. The left and right plots are for measurement error sets corresponding to 90% HPD regions on independent and joint Gaussian distributions, respectively. The triangles represent the real parameter values.

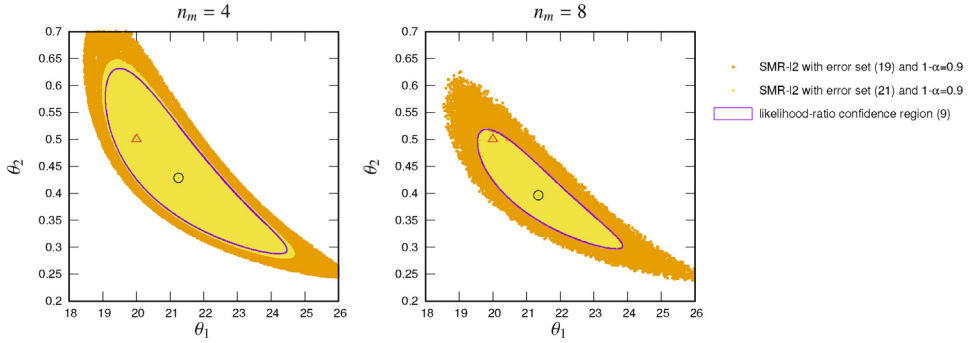


Fig. 4. 90% likelihood- ratio confidence regions compared with sampled SMR-ℓ_2 regions for the BOD example with 4 (left plot) and 8 (right plot) measurement points. The SMR regions are for two measurement error sets corresponding to the HPD regions (19) and (21) with $1-\alpha = 0.9$. The circles and triangles represent the maximum-likelihood estimates in ℓ_2- regression and the real parameter values, respectively.

Illustrative example (continued). A comparison between 90% likelihood-ratio confidence regions and two SMR-ℓ_2 regions corresponding to different measurement error sets is shown in Fig. 4, in the case of 4 and 8 measurement points. The SMR regions are inner-approximated using the same sampling strategy as previously.

The first error sets correspond to 90% HPD regions in (19) for the jointly Gaussian-distributed measurement errors—or, equivalently, the set $\mathbf{E}_2$ in (15). These SMR regions are found to be significantly larger than their 90% likelihood-ratio (or Wald) confidence counterparts. Also recall that, by Theorem 1, these SMR regions always enclose the GPE regions shown in Fig. 3 for the same error sets $\mathbf{E}_2$.

The second error sets are constructed per (21), in order to illustrate the asymptotic equivalence with classical confidence regions as established through Theorem 2; they correspond to 67% and 20% HPD regions for jointly Gaussian-distributed measurement errors with 4 and 8 measurements, respectively, as discussed in Remark 3. Such asymptotic convergence with an increasing number of measurements is clearly visible in Fig. 4, where the small discrepancy observed on the left plot for $n_m = 4$ cannot be seen anymore on the right plot for $n_m = 8$. The SMR framework is thus capable of providing equivalent confidence information as in classical statistical inference, with the attendant advantage of being able to switch between alternative error set descriptions or likelihood functions seamlessly.

4. Numerical solution and approximation

Describing the SMR region Θ_R as defined in (16) is a difficult task in general. A simple approach to enclosing Θ_R by a set of algebraic constraints, which would then allow the application of the same set-inversion techniques as for GPE (Section 2.2; Appendix A), entails a substitution of the regression problems by their optimality conditions. Since every element $\boldsymbol{\theta}$ in (the interior of) Θ_R should satisfy the first- and second-order optimality conditions

$$\frac{\partial \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m + \mathbf{e})}{\partial \boldsymbol{\theta}} = \mathbf{0} \quad \text{and} \quad \frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m + \mathbf{e})}{\partial \boldsymbol{\theta}^2} \preceq 0 \tag{24}$$

for some observation error $\mathbf{e} \in \mathbf{E}$, we have

$$\left\{ \boldsymbol{\theta} \in \Theta_0 \;\middle|\; \begin{array}{l} \exists \mathbf{e} \in \mathbf{E}: \\ \dfrac{\partial \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m + \mathbf{e})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \quad \dfrac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m + \mathbf{e})}{\partial \boldsymbol{\theta}^2} \preceq \mathbf{0} \end{array} \right\} \supseteq \Theta_R.$$

However, since the optimality conditions (24) hold for both local and global maxima of the likelihood function, as well as saddle points, this inclusion could end up being very conservative for nonlinear regression problems in general. Another important caveat with this approach is the computational penalty of applying a set-inversion algorithm in the $(n_\theta + n_x n_m)$-dimensional domain $\Theta_0 \times \mathbf{E}$, not merely in the original n_θ-dimensional domain Θ_0. The following subsections set out to develop more tractable, yet still conservative, bounding strategies to alleviate the computational burden of SMR, both in the form of confidence-like regions (Section 4.1) and polyhedral regions (Section 4.2).

4.1. Likelihood-contour enclosure

We consider the problem of enclosing the SMR region Θ_R within a confidence-like region of the form

$$\overline{\Theta}_R(\lambda) := \{\boldsymbol{\theta} \in \Theta_0 \mid \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m) \geq \lambda\}, \tag{25}$$

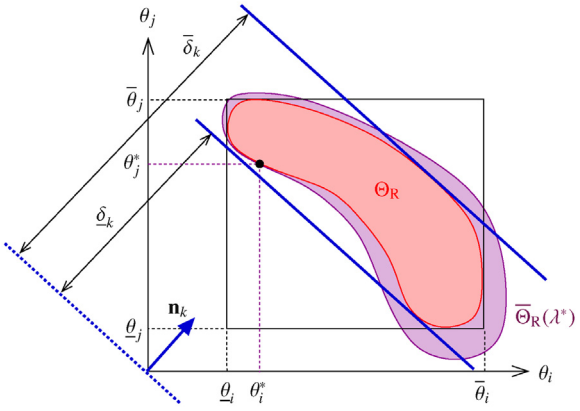


Fig. 5. Illustration of enclosure strategies for an SMR region Θ_R [red-shaded area], either in the form of a likely-contour enclosure $\overline{\Theta}_R(\lambda^*)$ [purple-shaded area], or by the box enclosure (32) [thin solid black lines] along with pairs of non- axis aligned cuts in the form (34) [thick solid blue lines]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

for some constant $\lambda \geq 0$. Notice that the computational complexity of describing, or closely approximating, the relaxed region $\overline{\Theta}_R(\lambda)$ is then comparable to describing either a likelihood-based confidence region (7) or a GPE region (12), for instance by applying a set-inversion algorithm in the original n_θ- dimensional domain Θ_0.

The following theorem provides a systematic means of computing a value λ^* such that $\overline{\Theta}_R(\lambda^*)$ is a tight enclosure of Θ_R, upon specializing $\varphi(\boldsymbol{\theta}) := \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m)$. This situation is depicted in Fig. 5.

Theorem 3. *Given any continuous function $\varphi : \mathbb{R}^{n_\theta} \to \mathbb{R}$, a valid enclosure $\{\boldsymbol{\theta} \in \Theta_0 \mid \varphi(\boldsymbol{\theta}) \geq \lambda\} \supseteq \Theta_R$ is obtained with $\lambda \geq \lambda^*$ and*

$$\begin{aligned} \lambda^* := & \min_{\boldsymbol{\theta}\in\Theta_0, \mathbf{e}\in\mathbf{E}} \ \varphi(\boldsymbol{\theta}) \\ & \text{s.t.} \quad \boldsymbol{\theta} \in \arg\max_{\boldsymbol{\varpi}\in\Theta_0} \log \mathcal{L}(\boldsymbol{\varpi} \mid \mathbf{x}^m + \mathbf{e}) \\ := & \min_{\boldsymbol{\theta}\in\Theta_0, \mathbf{e}\in\mathbf{E}} \ \varphi(\boldsymbol{\theta}) \\ & \text{s.t.} \ \forall \boldsymbol{\varpi} \in \Theta_0, \log \mathcal{L}(\boldsymbol{\varpi} \mid \mathbf{x}^m + \mathbf{e}) \leq \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m + \mathbf{e}). \end{aligned} \tag{26}$$

Moreover, the enclosure with λ^* is tight in the sense that the two sets share one or more boundary points.

Proof. Let $\overline{\boldsymbol{\theta}} \in \Theta_R$. From (16), there exists $\overline{\mathbf{e}} \in \mathbf{E}$ such that

$$\forall \boldsymbol{\varpi} \in \Theta_0, \quad \log \mathcal{L}(\boldsymbol{\varpi} \mid \mathbf{x}^m + \overline{\mathbf{e}}) \leq \log \mathcal{L}(\overline{\boldsymbol{\theta}} \mid \mathbf{x}^m + \overline{\mathbf{e}}).$$

Therefore, $(\overline{\boldsymbol{\theta}}, \overline{\mathbf{e}})$ satisfies the semi-infinite constraint in (26), and $\varphi(\overline{\boldsymbol{\theta}}) \geq \lambda^*$ follows immediately by optimality. Conversely, any optimal pair $(\boldsymbol{\theta}^*, \mathbf{e}^*)$ corresponding to the optimal value λ^* of (26) is such that $\boldsymbol{\theta}^* \in \operatorname{argmax}_{\boldsymbol{\varpi}\in\Theta_0} \log \mathcal{L}(\boldsymbol{\varpi} \mid \mathbf{x}^m + \mathbf{e}^*)$, and so $\boldsymbol{\theta}^* \in \Theta_R$. Since $\boldsymbol{\theta}^* \in \Theta_0$ and $\varphi(\boldsymbol{\theta}^*) = \lambda^*$, we have that $\boldsymbol{\theta}^*$ is also a boundary point of $\overline{\Theta}_R(\lambda^*)$, and hence a boundary point of Θ_R too. □

Specializing the function φ in Theorem 3 to the log-likelihood function in (25) gives

$$\begin{aligned} \lambda^* = & \min_{\boldsymbol{\theta}\in\Theta_0, \mathbf{e}\in\mathbf{E}} \ \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m) \\ & \text{s.t.} \quad \forall \boldsymbol{\varpi} \in \Theta_0, \quad \log \mathcal{L}(\boldsymbol{\varpi} \mid \mathbf{x}^m + \mathbf{e}) \leq \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m + \mathbf{e}). \end{aligned} \tag{27}$$

Solving this SIP problem is hard in general, since both the semi-infinite constraint and the objective function are generally nonconvex for a nonlinear regression problem. Existing solution approaches to SIP rely on either one of two key ideas [62,63]. In *local reduction methods*, a semi-infinite constraint is represented locally by a finite number of instances of the constraint, upon invoking the implicit function theorem. Alternatively, *discretization (and exchange) methods* involve replacing the uncertain parameter set with a finite discretization so as to create a relaxation of the SIP, and then iteratively refining this discretization until convergence. The focus in the remainder of this paper is on the second type of methods, for which global optimality certificates can be provided upon solving the nonlinear programming (NLP) subproblems to global optimality using complete search methods [64–66].

More specifically, we apply the cutting-plane SIP algorithm by Blankenship and Falk [67] in order to construct a sequence of decreasing upper bounds λ^k on the upper bound λ^* given by (27); that is, we construct an inclusion sequence $\overline{\Theta}_R(\lambda^k) \supseteq \overline{\Theta}_R(\lambda^*) \supseteq \Theta_R$. Within the SMR framework, this algorithm entails an iteration between:

(i) the finite-dimensional nonlinear programming (NLP) subproblems

$$\begin{aligned} (\boldsymbol{\theta}^k, \mathbf{e}^k) \in & \arg\min_{\boldsymbol{\theta}\in\Theta_0, \mathbf{e}\in\mathbf{E}} \ \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m) \\ & \text{s.t.} \quad \forall \boldsymbol{\varpi} \in \Theta_0^k, \quad \log \mathcal{L}(\boldsymbol{\varpi} \mid \mathbf{x}^m + \mathbf{e}) \leq \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m + \mathbf{e}), \end{aligned} \tag{28}$$

where $\Theta_0^k := \{\boldsymbol{\varpi}^0, \ldots, \boldsymbol{\varpi}^k\}$ is a finite subset of Θ_0; and

(ii) the feasibility subproblems

$$\boldsymbol{\varpi}^{k+1} \in \arg\max_{\boldsymbol{\varpi}\in\Theta_0} \ \log \mathcal{L}(\boldsymbol{\varpi} \mid \mathbf{x}^m + \mathbf{e}^k). \tag{29}$$

The subset Θ_0^k at iteration $k = 1$ may be initialized as the empty set, or better, as a singleton set with the maximum-likelihood estimate $\hat{\boldsymbol{\theta}}$ (see Section 2).

Under the assumptions that the likelihood function $\mathcal{L}$ is jointly continuous in $(\boldsymbol{\theta}, \mathbf{e})$ and that the parameter set Θ_0 and the error set $\mathbf{E}$ are both compact, any point of accumulation $\boldsymbol{\theta}^*$ of the sequence $\{\boldsymbol{\theta}^k\}$ will correspond to the best possible lower bound λ^* in (27) [67,Theorem 2.1]. In practice, the iterations may be interrupted when the following termination criterion is satisfied for a certain tolerance $\epsilon > 0$,

$$\log \mathcal{L}(\boldsymbol{\varpi}^{k+1} \mid \mathbf{x}^m + \mathbf{e}^k) \leq \log \mathcal{L}(\boldsymbol{\theta}^k \mid \mathbf{x}^m + \mathbf{e}^k) + \epsilon. \tag{30}$$

Naturally, such a convergence property of the cutting-plane algorithm hinges on the ability to solve all of the nonconvex subproblems (28) and (29) to global optimality. Otherwise, the resulting threshold values λ^* could be underestimated, leading to likelihood contours that exclude parts of the corresponding SMR regions. The practical applicability of this approach may thus be hindered by its computational complexity.

One way to expedite convergence of the cutting-plane algorithm is via the addition of redundant constraints, namely constraints that do not alter the optimal solution set of the SIP (27) yet tighten the relaxations in (28); see, e.g., [68,69] for more details about KKT-based tightening in SIP. Provided that the likelihood function is sufficiently smooth, one can add the first- and second-order optimality cuts (24) as redundant constraints in the subproblem (28), so that[2]

$$\begin{aligned} (\boldsymbol{\theta}^k, \mathbf{e}^k) \in & \arg\min_{\boldsymbol{\theta}\in\Theta_0, \mathbf{e}\in\mathbf{E}} \ \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m) \\ & \text{s.t.} \ \forall \boldsymbol{\varpi} \in \Theta_0^k, \quad \log \mathcal{L}(\boldsymbol{\varpi} \mid \mathbf{x}^m + \mathbf{e}) \leq \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m + \mathbf{e}) \\ & \frac{\partial \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m + \mathbf{e})}{\partial \boldsymbol{\theta}} = \mathbf{0}, \quad \frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^m + \mathbf{e})}{\partial \boldsymbol{\theta}^2} \preceq \mathbf{0}. \end{aligned} \tag{31}$$

[2] Given that most NLP solvers do not currently support constraints in the form of linear matrix inequalities (LMI), one can always substitute the LMI constraint $\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}^m+\mathbf{e})}{\partial \boldsymbol{\theta}^2} \preceq \mathbf{0}$ in (31) by standard inequality constraints on the principal minors of $\frac{\partial^2 \log \mathcal{L}(\boldsymbol{\theta}|\mathbf{x}^m+\mathbf{e})}{\partial \boldsymbol{\theta}^2}$ [70].

In the case that none of the regression problems $\max_{\boldsymbol{\varpi}\in\Theta_0} \log \mathcal{L}(\boldsymbol{\varpi} \mid \mathbf{x}^{\mathrm{m}} + \mathbf{e})$ have local (suboptimal) solutions for any $\mathbf{e} \in \mathbf{E}$, enforcing the semi-infinite constraint in (27) is of course equivalent to satisfying the optimality conditions (24), and so the cutting-plane algorithm will trivially terminate after a single iteration. Otherwise, the intermediate solution points $\boldsymbol{\theta}^k$ to the NLP subproblems (31) might correspond to local optima of the regression problems for $\mathbf{e}^k \in \mathbf{E}$, and the algorithm thus keeps iterating by adding cutting planes until all of these local optima have been excluded. At this point, satisfying both the discretized semi-infinite and optimality constraints in (31) becomes equivalent to enforcing the original semi-infinite constraint in (27), and the algorithm will then terminate exactly – optimality gap $\epsilon = 0$ in (30) – at the next iteration. This behavior will be illustrated for the case study problem in Section 5.1.

4.2. Polyhedral enclosure

Applying a set-inversion approach to describe (an enclosure of) the SMR region Θ_{R} can prove computationally expensive, if at all tractable, especially for the estimation problems encountered in real-life situations. A computationally less demanding task entails the computation of a simple (axis-aligned) box enclosure for an SMR region; for instance, by solving a pair of optimization problems for each parameter θ_i, $i = 1 \ldots n_\theta$, as

$$\underline{\theta}_i/\overline{\theta}_i := \min_{\boldsymbol{\theta}\in\Theta_0, \mathbf{e}\in\mathbf{E}} / \max_{\boldsymbol{\theta}\in\Theta_0, \mathbf{e}\in\mathbf{E}} \ \theta_i \tag{32}$$
$$\text{s.t.} \quad \forall \boldsymbol{\varpi} \in \Theta_0, \quad \log \mathcal{L}(\boldsymbol{\varpi} \mid \mathbf{x}^{\mathrm{m}} + \mathbf{e}) \le \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^{\mathrm{m}} + \mathbf{e}).$$

Clearly, these bounds may be computed by applying a similar cutting-plane algorithm as in Section 4.1 above, whereby the discretization subproblem (28) is now replaced with

$$(\boldsymbol{\theta}^k, \mathbf{e}^k) \in \arg\min_{\boldsymbol{\theta}\in\Theta_0, \mathbf{e}\in\mathbf{E}} / \operatorname{argmax}_{\boldsymbol{\theta}\in\Theta_0, \mathbf{e}\in\mathbf{E}} \ \theta_i$$
$$\text{s.t.} \quad \forall \boldsymbol{\varpi} \in \Theta_0^k, \quad \log \mathcal{L}(\boldsymbol{\varpi} \mid \mathbf{x}^{\mathrm{m}} + \mathbf{e}) \le \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^{\mathrm{m}} + \mathbf{e}),$$

and possibly supplemented with the redundant optimality cuts (24) as in (31).

As an alternative to the direct solution of the SIP problems in (32), one can also use the likelihood-contour enclosure $\overline{\Theta}_{\mathrm{R}}(\lambda^*)$ in (25) with the lower bound λ^* from (27) in order to construct an NLP relaxation of the SIP problem. A conservative box enclosure can be computed in this way by solving the auxiliary (potentially nonconvex) NLP problems

$$\underline{\theta}_i/\overline{\theta}_i := \min_{\boldsymbol{\theta}\in\Theta_0} / \max_{\boldsymbol{\theta}\in\Theta_0} \ \theta_i \tag{33}$$
$$\text{s.t.} \quad \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^{\mathrm{m}}) \ge \lambda^*, \quad i = 1 \ldots n_\theta.$$

Of course, the presence of several disconnected subsets in an SMR region cannot be detected by a simple box enclosure, and information about correlations between the parameters θ_i in the actual SMR region is also lost. Part of this information could nonetheless be recovered by constructing a polyhedral enclosure of the SMR region, e.g., expressed in the form

$$\left\{ \boldsymbol{\theta} \in [\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}] \;\middle|\; \underline{\delta}_k \le \mathbf{n}_k^T \boldsymbol{\theta} \le \overline{\delta}_k, \quad k = 1 \ldots m \right\}, \tag{34}$$

for a set of vectors $\mathbf{n}_1 \ldots \mathbf{n}_m \in \mathbb{R}^{n_\theta}$ and scalars $\underline{\delta}_1 \ldots \underline{\delta}_m, \overline{\delta}_1 \ldots \overline{\delta}_m \in \mathbb{R}$.

Specializing the function $\varphi(\boldsymbol{\theta}) := \mathbf{n}_k^T \boldsymbol{\theta}$ in Theorem 3 provides a means of constructing such non-axis-aligned polyhedral cuts. Herein, the directions $\mathbf{n}_k$ are chosen in such a way that the cuts correspond to a (face or interior) diagonal of the box enclosure $[\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}]$,

$$\mathbf{n}_k := \sum_{i=1}^{n_\theta} \frac{\nu_i}{|\boldsymbol{\nu}|} \frac{1}{\overline{\theta}_i - \underline{\theta}_i} \tag{35}$$

with $\boldsymbol{\nu} \in \{-1, 0, 1\}^{n_\theta}$ and $|\boldsymbol{\nu}| = \sum_{i=1}^{n_\theta} \nu_i \ge 2$. Further, the limits $\underline{\delta}_k$, $\overline{\delta}_k$ in (34) such that the polyhedral cuts are tight can be computed via the solution of the auxiliary SIP problems

$$\underline{\delta}_k/\overline{\delta}_k := \min_{\boldsymbol{\theta}\in[\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}], \mathbf{e}\in\mathbf{E}} / \max_{\boldsymbol{\theta}\in[\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}], \mathbf{e}\in\mathbf{E}} \ \mathbf{n}_k^T \boldsymbol{\theta} \tag{36}$$
$$\text{s.t.} \quad \forall \boldsymbol{\varpi} \in \Theta_0, \quad \log \mathcal{L}(\boldsymbol{\varpi} \mid \mathbf{x}^{\mathrm{m}} + \mathbf{e}) \le \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^{\mathrm{m}} + \mathbf{e}),$$

possibly supplemented with the redundant optimality cuts (24) once again. Similar to the box enclosure (33) earlier, conservative, yet computationally less demanding, polyhedral cuts could be derived from the likelihood-contour enclosure $\overline{\Theta}_{\mathrm{R}}(\lambda^*)$ by solving the auxiliary NLP problems

$$\underline{\delta}_k/\overline{\delta}_k := \min_{\boldsymbol{\theta}\in[\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}]} / \max_{\boldsymbol{\theta}\in[\underline{\boldsymbol{\theta}}, \overline{\boldsymbol{\theta}}]} \ \mathbf{n}_k^T \boldsymbol{\theta} \tag{37}$$
$$\text{s.t.} \quad \log \mathcal{L}(\boldsymbol{\theta} \mid \mathbf{x}^{\mathrm{m}}) \ge \lambda^*.$$

Notice that the spans $(\overline{\delta}_k - \underline{\delta}_k)$ are bounded in $[0, 1]$ by construction.

The case of a 2-dimensional face diagonal, where $\nu_i = \nu_j = 1$ are the only nonzero elements in (35), is shown in Fig. 5 for illustration. Enumerating all such pairs of parameters $(\theta_i, \theta_j)_{1\le i<j\le n_\theta}$ calls for the solution of $2n_\theta(n_\theta - 1)$ auxiliary optimization problems. More generally with $|\boldsymbol{\nu}| \ge 2$ nonzero elements in the vector $\boldsymbol{\nu}$, the number of optimization problems is equal to $2^{|\boldsymbol{\nu}|} \binom{n_\theta}{|\boldsymbol{\nu}|}$. To manage this high combinatorial complexity when the number of parameters n_θ is high, it is of course possible to include only those cuts involving combinations of $|\boldsymbol{\nu}| = 2$ or 3 parameters in the polyhedral enclosure at the price of a more conservative polyhedral enclosure.

A simple way of detecting correlations among any parameter pair $(\theta_i, \theta_j)_{1\le i<j\le n_\theta}$ is by calculating the shortest-to-longest ratio between the spans $(\overline{\delta}_k - \underline{\delta}_k)$ obtained with $\nu_i = \nu_j = 1$ on the one hand, and $\nu_i = -\nu_j = 1$ on the other hand. A ratio close to 0 indicates an elongated set projection onto (θ_i, θ_j) in one of the diagonal directions, and therefore a large correlation between θ_i and θ_j; whereas, a ratio close to 1 indicates a more spherical set projection onto (θ_i, θ_j). This approach is the counterpart to the shortest-to-longest axis ratio in an ellipsoidal (Wald) confidence region, which is also the basis for the so-called modified E-optimality criterion in experimental design [11]. More generally, shortest-to-longest-span ratios could be computed with $|\boldsymbol{\nu}| > 2$ in order to unravel correlations among more than 2 parameters likewise. Other classical criteria, such as the A-optimality and D-optimality criteria, also have counterparts in the SMR framework, given by the sum of all the parameter ranges $\overline{\theta}_i - \underline{\theta}_i$ for $i = 1 \ldots n_\theta$ and the volume of the polytope (34), respectively.

To conclude this subsection, it is worth mentioning that the construction of such polyhedral enclosures is also relevant to the approximation of classical inference regions, for instance the likelihood-ratio confidence regions (7).

Illustrative example (continued). Various enclosures of SMR-ℓ_2 regions are presented in Fig. 6 for the BOD case study, here with either 4 or 8 measurement points. The measurement error set $\mathbf{E}$ is constructed based on (21) at the confidence level $1 - \alpha = 0.9$. The threshold values λ^* in the likelihood-contour enclosures $\overline{\Theta}_{\mathrm{R}}(\lambda^*)$ (25) are computed using the cutting-plane SIP algorithm described in Section 4.1, with first-order optimality cuts as in the discretized subproblem (31) . When the subsets Θ_0^k are initialized with the corresponding maximum- likelihood estimate $\hat{\boldsymbol{\theta}}$, the cutting-plane

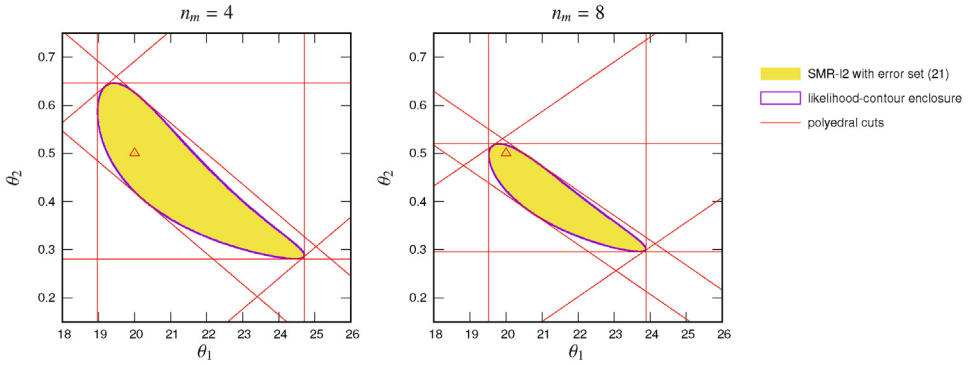


Fig. 6. Comparison of outer-approximation strategies to enclose the SMR-ℓ_2 regions for the BOD example with 4 (left) and 8 (right) measurement points: enclosures based on likelihood- contour cuts (25), and polyhedral cuts (34). The error set $\mathbf{E}$ is constructed based on (21) at the confidence level $1-\alpha = 0.9$. The triangles represent the real parameter values.

Table 1
Comparison between the thresholds λ^* and $\log \mathcal{L}(\hat{\boldsymbol{\theta}} \mid \mathbf{x}^m) - \frac{1}{2}\chi^2_2(0.9)$ corresponding to the SMR-ℓ_2 region (16) and the likelihood-based confidence region (7), respectively, for the BOD example with 4, 8 and 16 measurement points.

n_m	λ^*	$\log \mathcal{L}(\hat{\boldsymbol{\theta}} \mid \mathbf{x}^m) - \frac{1}{2}\chi^2_2(0.9)$
4	−7.66	−7.40
8	−11.90	−11.82
16	−21.20	−21.13

algorithm finds the exact solutions λ^* during the first iteration, irrespective of the number of measurement points. Even for such a simple estimation problem though, the solution of the discretized subproblem (31) to global optimality using GAMS-BARON proves computationally challenging as the number of measurement increases, here taking 404 CPU-sec for 4 measurement points, 3810 CPU-sec for 8 measurement points, and failing to close the gap within 7200 CPU-sec for 16 measurement points.[3] The GAMS code is provided as part of the Supplementary Information (see Appendix C) for the sake of reproducibility.

The likelihood-contour enclosures $\overline{\Theta}_R(\lambda^*)$ are found to provide a very close approximation of the SMR-ℓ_2 regions in Fig. 6—these enclosures are computed using the set- inversion algorithm described in Appendix A. This is expected given the fast convergence between the SMR and likelihood-based confidence regions already observed in Fig. 4, and confirmed by the comparison in Table 1 between the thresholds defining these two confidence regions.

For simplicity, the polyhedral cuts in Fig. 6 are constructed from the likelihood-contour enclosures $\overline{\Theta}_R(\lambda^*)$ rather than the actual SMR- ℓ_2 regions Θ_R here. The numerical solution of the auxiliary NLP subproblems (33) and (37) to global optimality using GAMS-BARON is fast in comparison with the SIP problems, taking <1 CPU-sec.

Finally, the shortest-to-longest-span ratios in the polyhedral enclosures of the SMR-ℓ_2 regions for 4, 8 and 16 measurement points are $\frac{0.284}{0.965} \approx 0.294$, $\frac{0.249}{0.970} \approx 0.257$ and $\frac{0.256}{0.967} \approx 0.265$, respectively. These small ratios (compared to 1) indicate that the SMR regions are 3- to 4-times flatter in one direction compared to the other direction, which unravels the presence of a strong correlation between θ_1 and θ_2 in (13), which is in agreement with the visual impression on Fig. 6.

[3] The reported CPU times are for an AMD Athlon 64 CPU at 2.2 GHz, running Red Hat 4.4.7-18, GAMS 25.0.2, and BARON 17.10.16 with default options, a relative convergence tolerance of 10^{-3} and time limit of 7200 CPU-sec.

5. Case study in temperature-dependent microbial growth

We now apply the SMR framework to a more challenging estimation problem in microbial growth, emphasizing their properties and drawing comparisons with other set-membership and statistical inference methods. Two models describing the effect of culture temperature, T on the growth rate, μ of a microbial population, each one comprising four parameters, are:

(i) The *Ratkowsky model*[71]:

$$\mu(T) = [b\,(T - T_{\min})(1 - \mathrm{e}^{c(T-T_{\max})})]^2,$$

where $T_{\min}$ and $T_{\max}$ (K) represent the minimal and maximal temperatures, respectively; while b ($\mathrm{K}^{-1}\,\mathrm{h}^{0.5}$) and c (K^{-1}) are extra parameters adding flexibility to the shape of the growth model.

The *cardinal temperature model*[72]:

$$\mu(T) = \mu_{\mathrm{opt}}\left[1 - \frac{(T - T_{\mathrm{opt}})^2}{(T - T_{\mathrm{opt}})^2 + T(T_{\max} + T_{\min} - T) - T_{\max}T_{\min}}\right],$$

where $T_{\min}$ and $T_{\max}$ (K) also represent the minimal and maximal temperatures, respectively; T_{opt} (K) corresponds to the optimal growth temperature; and μ_{opt} (h^{-1}) is the maximal growth rate attained at T_{opt}.

Experimental data used in the regression are from [71] for the bacterium *E. coli*. This data set comprises 15 measurement pairs (T_k, μ_k) within the temperature range 294–320 (K), and it is reproduced in Appendix B for completeness. The standard deviation of the growth rate measurements is taken as $\sigma_\mu = 0.1\,(\mathrm{h}^{-1})$ throughout. Results of a maximum-likelihood estimation with constant-variance and Gaussian-distributed errors – or, equivalently, a standard least-squares regression – are presented in Fig. 7. Both model predictions are found to be in good agreement with the experimental data, yet with a higher likelihood for the cardinal temperature model. Note also that errors are only taken into account for the growth rate measurements (outputs) herein, i.e. the temperature measurements (inputs) are considered to be exact.

5.1. Computational procedure and performance

For both candidate models we use the cutting-plane SIP algorithm of Section 4.1 to compute the threshold values λ^* (27), and we describe tight likelihood-contour enclosures $\overline{\Theta}_R(\lambda^*)$ (25) of the SMR regions using a set-inversion algorithm (see Appendix A) in turn. We apply a similar cutting-plane SIP algorithm to determine the box and polyhedral enclosures based on (32) and (36) with $|\boldsymbol{\nu}|$

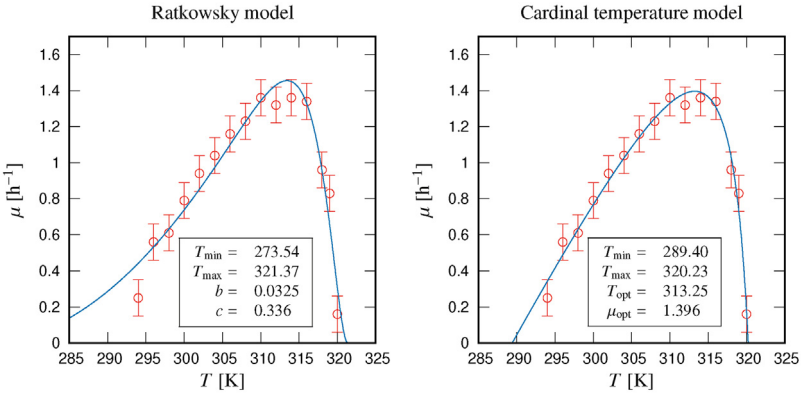


Fig. 7. Maximum- likelihood estimation results for the Ratkowsky (left) and cardinal temperature (right) models. The boxes on each plot report the maximum-likelihood parameter estimates.

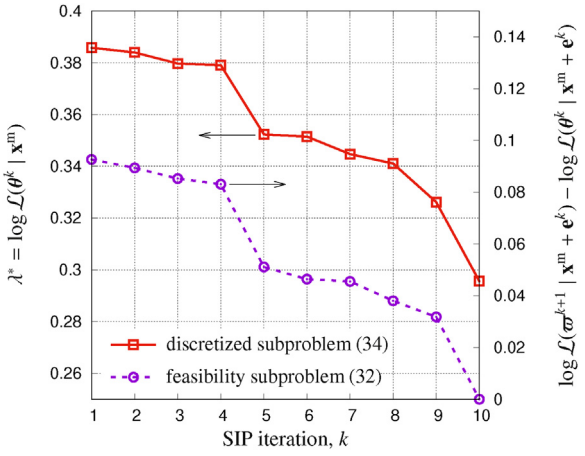


Fig. 8. Iterations of the cutting-plane SIP algorithm for computing the solution value λ^* of (27) for the Ratkowsky model. Left y- axis: solution value λ^* of subproblem (31) at iteration k. Right y-axis: optimal feasibility gap of subproblem (29) at iteration k.

= 2, as described in Section 4.2. First-order optimality cuts are added in the discretized NLP subproblems for all of the SIP problems, as in (31), in order to expedite the convergence of the cutting-plane algorithm, and the sets Θ_0^k are initialized with the maximum-likelihood estimates $\hat{\boldsymbol{\theta}}$ at iteration $k = 1$. All of the NLP subproblems in the SIP algorithm are solved with the global solver GAMS- BARON—these GAMS codes are provided as part of the Supplementary Information (see Appendix C) for reproducibility.[3] Lastly, the set-inversion computations are carried using our in-house library CRONOS [44], which is available from https://github.com/omega- icl/cronos.

In the case of the cardinal temperature model, a single iteration is needed to solve all of the SIP problems exactly – optimality gap $\epsilon = 0$ in (30). This behavior indicates that none of the regression problems for this model exhibit local, suboptimal solutions for the measurement error sets of interest in Sections 5.2 and 5.3 below. In the case of the Ratkowsky, the SIP problems are also solved exactly, but the cutting-plane algorithm terminates after several iterations due to the presence of local optima; for instance, computing the solution value λ^* of (27) for the SMR problem in Section 5.2 takes 10 iterations to terminate, as shown in Fig. 8.

Even though the cutting-plane SIP algorithms terminate exactly (after a single or several iterations), certifying global optimality for most discretized NLP subproblems is currently intractable with the state-of-the-art global solvers BARON [73] and ANTIGONE [66]. As already discussed in Section 4.1, this lack of guarantees could result in the likelihood contours or polyhedral cuts excluding parts of the actual SMR regions. The odds of missing a global optimum in a discretized NLP subproblem is nonetheless mitigated by letting BARON or ANTIGONE run up to a time limit of 7200 CPU-sec here.

5.2. SMR with jointly Gaussian-distributed errors

We consider SMR-ℓ_2 regions, where the error set $\mathbf{E}$ corresponds to the HPD region of a joint Gaussian distribution, as in (21). In order to draw on the asymptotic equivalence with a 95% confidence region in classical frequentist inference (Theorem 2), we select a 15% HPD region for the joint Gaussian distribution of the measurement errors here (see Remark 3). The likelihood-contour and polyhedral enclosures of these SMR regions are compared in Figs. 9 and 10 for the Ratkowsky and cardinal temperature models, respectively. The results from a random sampling are also shown on these plots, which lie inside the actual SMR regions.

Since the polyhedral cuts are tight by construction (Theorem 3), the seemingly large discrepancy between these cuts and the sampled SMR regions in Figs. 9 and 10 is mainly attributable to the sampling not being sufficiently exhaustive. Moreover, the comparisons between the polyhedral and likelihood-contour enclosures on these figures show that the conservatism introduced by the second remains small for both models in the present case of jointly Gaussian-distributed measurement errors.

Reported above each plot in Figs. 9 and 10 are the shortest-to-longest-span ratios in the polyhedral enclosure for the various parameter pairs (see Section 4.2). With the Ratkowsky model, all of these ratios happen to be smaller than 0.4, and even lower than 0.25 for the parameter pair ($T_{\min}$, b), thereby suggesting strong correlations in the parameter set ($T_{\min}$, $T_{\max}$, b, c). With the cardinal temperature model by contrast, most of the ratios are close to or above 0.5, suggesting much weaker correlations amongst the parameters ($T_{\min}$, $T_{\max}$, T_{opt}, μ_{opt}) thereof. Moreover, the SMR intervals for the parameters $T_{\min}$ and $T_{\max}$ – which participate and share the same interpretation in both models – are much larger for the Ratkowsky model than they are for the cardinal temperature model. On the basis of these results, a modeler would normally retain the cardinal temperature model over the Ratkowsky model.

Although the Wald confidence ellipsoids in Figs. 9 and 10 differ significantly from the SMR region enclosures, similar conclusions can nonetheless be drawn with regards to parameter precision and correlation for both the Ratkowsky and cardinal temperature models based on the main axes of the projected Wald ellipsoids. One can also compare the threshold of a likelihood-contour enclosure with its likelihood-based confidence region counterpart (7): for the Ratkowsky model, we find $\lambda^* \approx 5.9$ and $\log \mathcal{L}(\hat{\boldsymbol{\theta}} \mid \mathbf{x}^{\text{m}}) - \frac{1}{2}\chi^2_{n_\theta}(0.95) \approx 9.5$; whereas for the cardinal temperature model, we have $\lambda^* \approx 12.7$ and $\log \mathcal{L}(\hat{\boldsymbol{\theta}} \mid \mathbf{x}^{\text{m}}) - \frac{1}{2}\chi^2_{n_\theta}(0.95) \approx 14.0$. These values

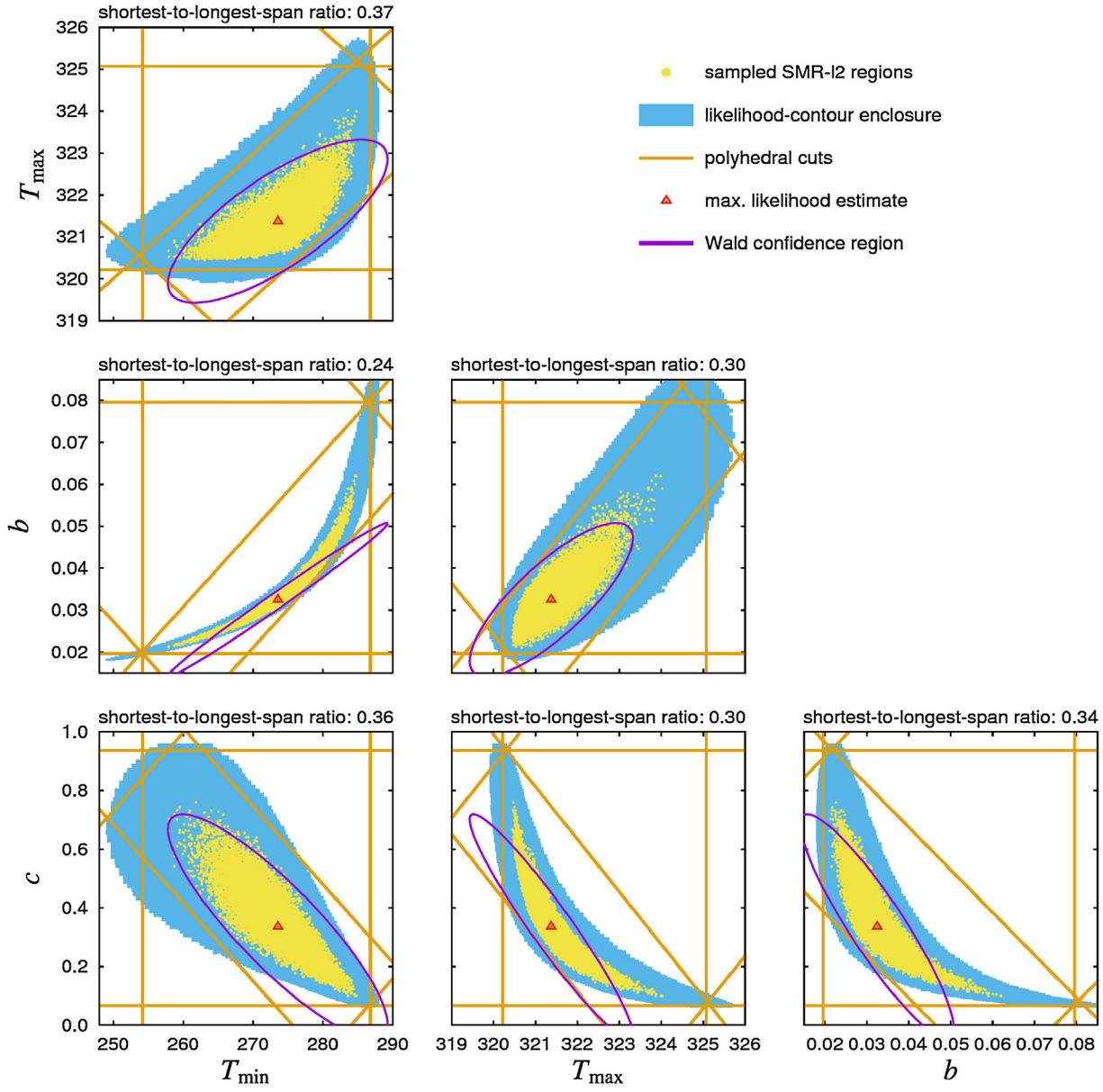


Fig. 9. Matrix plot comparing various parameter regions for the Ratkowsky model: sampled SMR regions, likelihood-contour and polyhedral enclosures of SMR regions, and Wald confidence region. The SMR region is based on ℓ_2-regression, with the error set being the 15% HPD region for the joint Gaussian distribution of the measurement errors.

being quite close to each other for both models provides yet another illustration of the asymptotic equivalence between classical statistical inference approaches and SMR for such choices of the error set (viz. Section 3.2). Also notice the higher likelihood threshold of the cardinal temperature model compared with the Ratkowsky model, which provides yet another indication of a much more confident estimation.

5.3. SMR with independently-distributed errors

We consider alternative SMR-ℓ_2 regions, where the error set $\mathbf{E}$ now comprises independent, 1-sigma error bounds on the measurements,

$$\mathbf{E} := \left\{ \mathbf{e}_\mu \in \mathbb{R}^{15} \;\middle|\; \forall k = 1 \ldots 15, \quad \left| e_{\mu,k} \right| \leq \sigma_\mu \right\}. \tag{38}$$

Similar to the jointly Gaussian-distributed case in Section 5.2 above, we compare various approximations of such an SMR region for the cardinal temperature model in Fig. 11; namely, the tight likelihood-contour and polyhedral enclosures, and an inner-approximation using a random sampling. Despite the error set $\mathbf{E}$ in (38) now being significantly different from a Gaussian HPD region, the polyhedral enclosures turn out to be comparable in shape and size to those in Fig. 10; and the shortest-to-longest-span ratios for the various parameter pairs are similar too. The likelihood-contour enclosure in Fig. 11 describes a rather close approximation of the SMR region too, albeit proving to be more conservative than for the jointly Gaussian- distributed case in Fig. 10. A similar behavior is obtained for the Ratkowsky model (results not shown).

In addition to SMR region approximations, Fig. 11 displays the guaranteed parameter region as given by (12), for the same error set (38). One can check that the inclusion property established in Theorem 1 holds. The guaranteed parameter region turns out to be much smaller than the SMR region here due to both the model mismatch and underestimating the measurement noise. For the Ratkowsky model, the guaranteed parameter region even happens to be empty for these data and error sets. Therefore, unlike SMR regions, guaranteed parameter regions do not provide a reliable means of detecting parameter correlations in the present case.

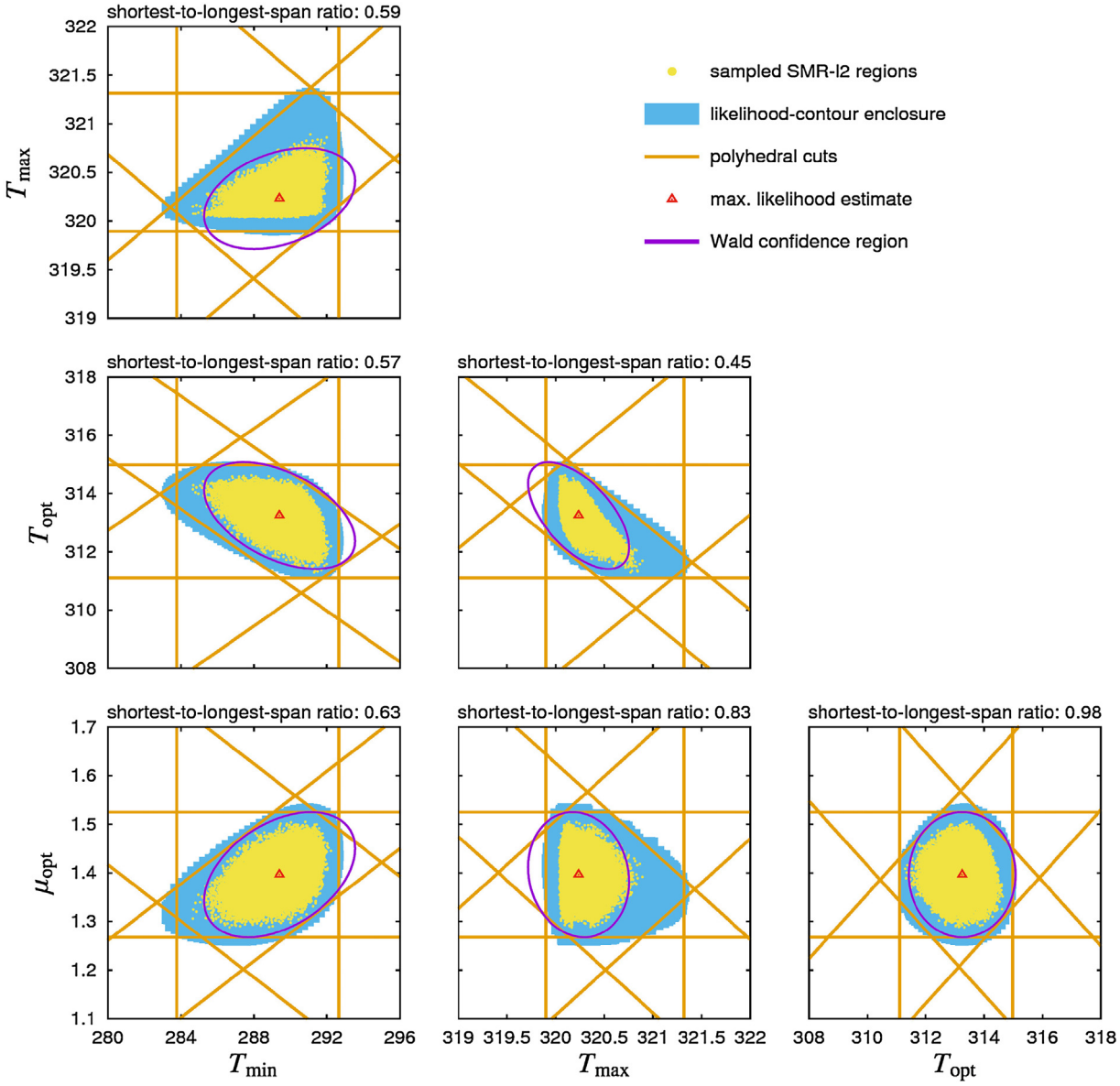


Fig. 10. Matrix plot comparing various parameter regions for the cardinal temperature model: sampled SMR regions, likelihood-contour and polyhedral enclosures of SMR regions, and Wald confidence region. The SMR region is based on ℓ_2-regression, with the error set being the 15% HPD region for the joint Gaussian distribution of the measurement errors.

6. Conclusions and future research directions

This paper has introduced set-membership regression (SMR), a new approach to parameter estimation which seeks to determine the subregion in parameter space enclosing all (global) solutions to a nonlinear regression problem subject to uncertain observations. An SMR region is thus understood as comprising those parameter values that are equally credible under the selected observation error set, in the sense of that regression problem. In particular, this interpretation is not conditional upon the model's structure being correct. Another distinctive feature of SMR is its ability to consider likelihood functions and error sets other than those corresponding to jointly Gaussian-distributed errors, including least-absolute-error (ℓ_1) regression, and independent error distributions or simple error bounds when the underlying statistics is unknown.

In a bounded-error context, SMR provides a means of robustifying existing guaranteed parameter estimation methods. By drawing on the principles of maximum likelihood estimation, an SMR region encloses the corresponding guaranteed parameter set, and unlike the latter, it may not become empty in the presence of large model mismatch or measurement errors and outliers. From a statistical inference viewpoint, SMR has been shown to be asymptotically equivalent to the Wald confidence regions for specific choices of the measurement error set. It will be important to keep developing the underlying SMR theory as part of future work, so as to better grasp the links with both frequentist and Bayesian statistical inference analysis.

Another important contribution of this paper is a computational framework for describing tight enclosures of the SMR regions, in the form of likelihood-contour and polyhedral enclosures. These enclosures can be described via the solution of auxiliary optimization problems, which are typically nonconvex and embed semi-infinite constraints. While tractable in principle using global optimization techniques based on complete search, our experience with such optimization problems is that they challenge state-of-the-art global optimization solvers such as BARON or ANTIGONE, even for small-scale estimation problems as exemplified with the

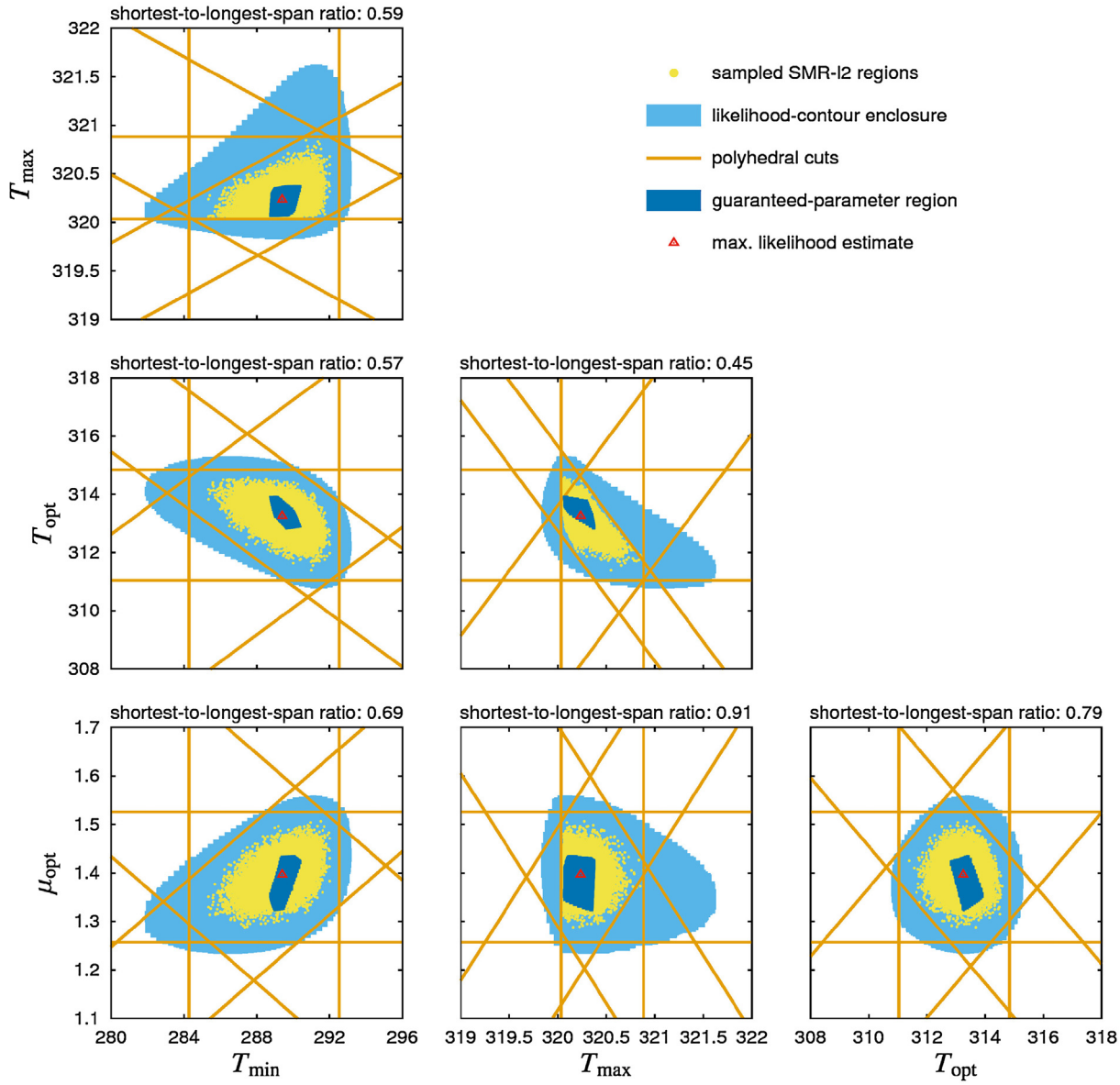


Fig. 11. Matrix plot comparing various parameter regions for the cardinal temperature model: sampled SMR regions, likelihood-contour and polyhedral enclosures of SMR regions, and Wald confidence region. The SMR region is based on ℓ_2-regression, with the error set corresponding to independent 1-sigma error bounds on the measured values.

BOD and microbial growth case studies. The tackling of larger-scale problems, including error-in-variables formulations, is a clear call for improved global search techniques; e.g., by exploiting problem structures or creating redundancy to strengthen the relaxations, or by combining with effective heuristics to increase the likelihood of finding a solution early on during the search [65].

One straightforward extension of the SMR methodology includes parameter estimation problems with other sources of uncertainty than just measurement errors. In principle, any set of nuisance parameters could be accounted for in the regression framework based on a description of the corresponding uncertainty set, similar to the measurement error set **E** in (16).

Lastly, it is worth reiterating that the SMR framework can be extended to parameter estimation in dynamic systems too. The main bottleneck in doing so is of computational rather than conceptual nature, since limited work has been published to date on SIP with differential equations embedded [74]. For instance, applying the cutting-plane SIP algorithm of Section 4 to the dynamic case should rely on efficient complete-search methods for global optimization and constraint satisfaction in dynamic optimization problems [44,75,76].

Data statement

No new data was collected in the course of this research.

Acknowledgements

NDP is grateful to EPSRC and the Department of Chemical Engineering at Imperial College London for Doctoral Training Award (DTA). RP gratefully acknowledges the contribution of the Slovak Research and Development Agency under the project APVV 15-0007. The authors would like to thank the anonymous reviewers for their thoughtful comments that led to substantial improvement of the article.

Appendix A. Set-inversion techniques

The problem in set inversion is describing, or approximating as closely as possible, a set Θ given in implicit form as

$$\Theta := \left\{ \boldsymbol{\theta} \in \Theta_0 \middle| \boldsymbol{\varphi}(\boldsymbol{\theta}) \in \Gamma \right\},$$

where $\Theta_0 \subset \mathbb{R}^{n_\theta}$ is the domain set; $\Gamma \subset \mathbb{R}^{n_\varphi}$, the target set; and $\boldsymbol{\varphi} : \mathbb{R}^{n_\theta} \to \mathbb{R}^{n_\varphi}$ is a continuous function. In other words, Θ is the pre-image of Γ under $\boldsymbol{\varphi}$ in Θ_0. The description of log-likelihood confidence regions (7) and guaranteed parameter regions (12), as well as SMR region enclosures (25), can all be cast as set- inversion problems. Existing computational approaches to set inversion range from sampling-based methods, including stochastic search [39] , support vector machines (SVM) [40] and MCMC [41], to rigorous complete-search methods based on interval analysis and other set arithmetics [42–44] or semidefinite relaxation techniques [45,46].

The focus herein is on branch-and-prune, a complete- search approach entailing the construction of partitions $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{bnd}}$ such that

$$\bigcup_{P \in \mathbb{P}_{\text{in}}} P \subseteq \Theta \subseteq \bigcup_{P \in \mathbb{P}_{\text{in}} \cup \mathbb{P}_{\text{bnd}}} P,$$

with $\mathbb{P}_{\text{bnd}}$ sufficiently small (in the sense of a certain metric). A prototypical algorithm is the following [42,61]:

Algorithm 1. Basic branch-and-prune algorithm for set inversion.

Input: Termination tolerances $\epsilon_{\text{box}} \geq 0$
Initialization: Set partitions $\mathbb{P}_{\text{bnd}} = \{\Theta_0\}$ and $\mathbb{P}_{\text{in}} = \emptyset$; Set iteration counter $k = 0$
Main Loop:
1. Select a parameter box P in the partition $\mathbb{P}_{\text{bnd}}$ and remove it from $\mathbb{P}_{\text{bnd}}$
2. Compute an enclosure $\overline{\boldsymbol{\varphi}}(P) \supseteq \{\boldsymbol{\varphi}(\boldsymbol{\theta}) \,|\, \boldsymbol{\theta} \in P\}$
3. Exclusion Tests:
   (a) **If** $\overline{\boldsymbol{\varphi}}(P) \subset \Gamma$, insert P into $\mathbb{P}_{\text{in}}$
   (b) **Else if** $\overline{\boldsymbol{\varphi}}(P) \cap \Gamma = \emptyset$, fathom P
   (c) **Else** bisect P and insert subsets back into $\mathbb{P}_{\text{bnd}}$
4. **If** $\text{width}(P) \leq \epsilon_{\text{box}}$ for all $P \in \mathbb{P}_{\text{bnd}}$, **stop**
5. Increment counter k+=1; **Return** to step 1
Output: Partitions $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\text{bnd}}$; Iteration count k

The basic requirements for finite convergence of Algorithm 1 are that: (i) the branching procedure is exhaustive; and (ii) the bounding is rigorous and convergent. Step 2 requires an enclosure of the reachable set of $\boldsymbol{\varphi}$ for the current parameter box P, which is the most critical step in Algorithm 1. Various bounding approaches are detailed in [44] (and references therein), with a focus on so-called *factorable* functions; namely, functions that can be represented by a finite number of binary sums, binary products and outer compositions with a univariate function. When the computed bounds are parameter-dependent, as is the case with McCormick relaxations or polynomial models, domain reduction techniques can be used within Step 2 in order to expedite convergence, e.g., via the solution of auxiliary optimization problems. In the case of polynomial models, these bounding subproblems may be nonconvex and it is therefore necessary to construct convex/polyhedral relaxations, for instance in the form of linear programs (LPs). This approach is the same as domain reduction in the context of branch- and-bound search for global optimization [64,65].

Appendix B. Data for the numerical case studies

The measurement data for the BOD example introduced at the end of Section 2 are reported in Table B.1 . Those for the microbial growth problem in Section 5 are reported in Table B.2 .

Table B.1
BOD concentrations at various time instants.

Times, t (day)	0.5	1.0	1.5	2.0	2.5	3.0	3.5	4.0
BOD, c (mg L^{-1})	3.696	6.197	11.289	12.626	13.589	14.702	17.105	17.425
Time, t (day)	4.5	5.0	5.5	6.0	6.5	7.0	7.5	8.0
BOD, c (mg L^{-1})	17.810	18.212	18.625	18.329	19.138	19.522	20.768	21.562

Table B.2
Specific growth rates of *E. coli* at various temperatures.

Temperature, T (K)	294	296	298	300	302	304	306	308
Spec. growth rate, μ (h^{-1})	0.25	0.56	0.61	0.79	0.94	1.04	1.16	1.23
Temperature, T (K)	310	312	314	316	318	319	320	
Spec. growth rate, μ (h^{-1})	1.36	1.32	1.36	1.34	0.96	0.83	0.16	

Appendix C. Supplementary data

Supplementary data associated with this article can be found, in the online version, at https://doi.org/10.1016/j.jprocont.2018.04.002.

References

[1] Y. Bard, Nonlinear Parameter Estimation, Academic Press, New York, 1974.
[2] R.A. Gallant, Nonlinear Statistical Models, John Wiley & Sons, New York, 1987.
[3] D. Bates, D. Watts, Nonlinear Regression Analysis and Its Applications, John Wiley & Sons, 1988.
[4] G.A.F. Seber, C.J. Wild, Nonlinear Regression, John Wiley & Sons, New York, 1989.
[5] W.E. Deming, Statistical Adjustment of Data, Wiley, New York, 1943.
[6] I.-W. Kim, M.J. Liebman, T.F. Edgar, Robust error-in-variables estimation using nonlinear programming techniques, AIChE J. 36 (1990) 985–993.
[7] W.R. Esposito, C.A. Floudas, Global optimization in parameter estimation of nonlinear algebraic models via the error-in-variables approach, Ind. Eng. Chem. Res. 37 (1998) 1841–1858.
[8] C.-Y. Gau, M.A. Stadtherr, Deterministic global optimization for error-in-variables parameter estimation, AIChE J. 48 (2002) 1192–1197.
[9] E. Walter (Ed.), Identifiability of Parametric Models, Pergamon Press, Oxford, 1987.
[10] S.P. Asprey, S. Macchietto, Statistical tools for optimal dynamic model building, Comput. Chem. Eng. 24 (2000) 1261–1267.
[11] K.A. McLean, K.B. McAuley, Mathematical modelling of chemical processes – obtaining the best model predictions and parameter estimates using identifiability and estimability procedures, Can. J. Chem. Eng. 90 (2) (2012) 351–366.
[12] D. Bonvin, C. Georgakis, C.C. Pantelides, M. Barolo, M.A. Grover, D. Rodrigues, R. Schneider, D. Dochain, Linking models and experiments, Ind. Eng. Chem. Res. 24 (2016) 6891–6903.
[13] W. Rooney, L.T. Biegler, Incorporating joint confidence regions into design under uncertainty, Comput. Chem. Eng. 23 (1999) 1563–1575.
[14] W. Rooney, L. Biegler, Design for model parameter uncertainty using nonlinear confidence regions, AIChE J. 47 (2001) 1794–1804.
[15] W. Langson, S. Raković, I. Chryssochoos, D.Q. Mayne, Robust model predictive control using tubes, Automatica 40 (1) (2004) 125–133.
[16] V. Sakizlis, N. Kakalis, V. Dua, J.D. Perkins, E.N. Pistikopoulos, Design of robust model-based controllers via parametric programming, Automatica 40 (2) (2004) 189–201.
[17] M.E. Villanueva, R. Quirynen, M. Diehl, B. Chachuat, B. Houska, Robust MPC via min–max differential inequalities, Automatica 77 (2017) 311–321.
[18] J.-L. Gouzé, A. Rapaport, Z. Hadj-Sadok, Interval observers for uncertain biological systems, Ecol. Model. 133 (2000) 45–56.
[19] B. Chachuat, O. Bernard, Probabilistic observers for a class of uncertain biological processes, Int. J. Robust Nonlinear Control 16 (3) (2006) 157–171.
[20] C.R. Rojas, J.S. Welsh, G.C. Goodwin, A. Feuer, Robust optimal experiment design for system identification, Automatica 43 (6) (2007) 993–1008.
[21] S.W. Marvel, C.M. Williams, Set membership experimental design for biological systems, BMC Syst. Biol. 6 (1) (2012) 21.
[22] A.R. Gottu-Mukkula, R. Paulen, Model-based design of optimal experiments for nonlinear systems in the context of guaranteed parameter estimation, Comput. Chem. Eng. 99 (2017) 198–213.
[23] R.D. Cook, S. Weisberg, Confidence curves in nonlinear regression, J. Am. Stat. Assoc. 85 (1990) 544–551.
[24] W.Q. Meeker, L.A. Escobar, Teaching about approximate confidence regions based on maximum likelihood estimation, Am. Stat. 49 (1) (1995) 48–53.
[25] M.J. Bayarri, J.O. Berger, The interplay of Bayesian and frequentist analysis, Stat. Sci. 19 (1) (2004) 58–80.

[26] A. Gelman, J. Carlin, H. Stern, D. Rubin, Bayesian Data Analysis, 2nd ed., Chapman & Hall/CRC, 2004.
[27] A. Smith, G. Roberts, Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods, J. R. Stat. Soc. 55 (1993) 3–23.
[28] W. Gilks, S. Richardson, D. Spiegelhalter, Markov Chain Monte Carlo in Practice, 1st ed., Chapman & Hall/CRC, 1996.
[29] E. Laloy, B. Rogiers, J. Vrugt, D. Mallants, D. Jacques, Efficient posterior exploration of a high-dimensional groundwater model from two-stage Markov chain Monte Carlo simulation and polynomial chaos expansion, Water Resour. Res. 49 (2013) 2664–2682.
[30] C.P. Robert, The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation, 2nd ed., Springer, 2001.
[31] J. Berger, The case for objective Bayesian analysis, Bayesian Anal. 1 (3) (2006) 385–402.
[32] K. Fedra, G. Van Straten, M.B. Beck, Uncertainty and arbitrariness in ecosystems modelling: a lake modelling example, Ecol. Model. 13 (1–2) (1981) 87–110.
[33] Special Issue on Parameter Identification with Error Bounds, Mathematics & Computers in Simulation 32 (1990) 447-607.
[34] M. Milanese, J.P. Norton, H. Piet- lahanier, E. Walter, Bounding Approaches to System Identication, Plenum Press, New York, 1996.
[35] J. Anderson, A. Papachristodoulou, On validation and invalidation of biological models, BMC Bioinform. 10 (2009) 132.
[36] P. Rumschinski, S. Borchers, S. Bosio, R. Weismantel, R. Findeisen, Set-based dynamical parameter estimation and model invalidation for biochemical reaction networks, BMC Syst. Biol. 4 (2010) 69.
[37] L. Jaulin, M. Kieffer, O. Didrit, E. Walter, Applied Interval Analysis, Springer-Verlag, London, 2001.
[38] S. Streif, M. Karl, R. Findeisen, Outlier analysis in set-based estimation for nonlinear systems using convex relaxations, Proceedings of the 2013 European Control Conference (2013) 2921–2926.
[39] T. Goerke, E. Engell, Application of evolutionary algorithms in guaranteed parameter estimation, 2016 IEEE Congress on Evolutionary Computation (CEC) (2016) 500–505.
[40] K.J. Keesman, R. Stappers, Nonlinear set-membership estimation: a support vector machine approach, J. Inverse Ill-Posed Probl. 12 (1) (2004) 27–41.
[41] E.W. Bai, H. Ishii, R. Tempo, A Markov chain Monte Carlo approach to nonlinear parameter system identification, IEEE Trans. Autom. Control 60 (9) (2015) 2542–2546.
[42] L. Jaulin, E. Walter, Set inversion via interval analysis for nonlinear bounded-error estimation, Automatica 29 (1993) 1053–1064.
[43] L. Jaulin, E. Walter, Guaranteed nonlinear parameter estimation from bounded-error data via interval analysis, Math. Comput. Simul. 35 (1993) 123–137.
[44] B. Chachuat, B. Houska, R. Paulen, N.D. Perić, J. Rajyaguru, M.E. Villanueva, Set-theoretic approaches in analysis, estimation and control of nonlinear systems, IFAC-PapersOnLine 48 (8) (2015) 981–995.
[45] V. Cerone, D. Piga, D. Regruto, Set-membership error-in-variables identification through convex relaxation techniques, IEEE Trans. Autom. Control 57 (2) (2012) 517–522.
[46] V. Magron, D. Henrion, J.B. Lasserre, Semidefinite approximations of projections and polynomial images of semialgebraic sets, SIAM J. Optim. 25 (4) (2015) 2143–2164.
[47] M. Milanese, Properties of least-squares estimates in set membership identification, Automatica 31 (2) (1995) 327–332.
[48] B.T. Poljak, J.Z. Tsypkin, Robust identification, Automatica 16 (1980) 53–63.
[49] A. van den Bos, Nonlinear least-absolute-values and minimax model fitting, Automatica 24 (6) (1988) 803–808.
[50] S.W. Marvel, C.M. Williams, Computational experience with confidence regions and confidence intervals for nonlinear least squares, Technometrics 29 (1) (1987) 67–82.
[51] D. Cox, D. Hinkley, Theoretical Statistics, 1st ed., Chapman and Hall, 1974.
[52] R.F. Engle, Wald, likelihood ratio, and Lagrange multiplier tests in econometrics, in: Z. Griliches, M.D. Intriligator (Eds.), Handbook of Econometrics, vol. 2, North Holland, 1984, pp. 775–826, Chapter 13.
[53] L. Jaulin, E. Walter, Guaranteed nonlinear parameter estimation via interval computations, Interval Comput. 3 (1993) 61–75.
[54] L. Jaulin, Computing minimal-volume credible sets using interval analysis; application to Bayesian estimation, IEEE Trans. Signal Process. 54 (2006) 3632–3636.
[55] B.L. Welch, H.W. Peers, On formulae for confidence points based on integrals of weighted likelihoods, J. R. Stat. Soc. B 25 (1991) 318–329.
[56] T.A. Severini, On the relationship between Bayesian and non-Bayesian interval estimates, J. R. Stat. Soc. B 53 (3) (1991) 611–618.
[57] L. Ventura, W. Racugno, A note on the relationships between Bayesian and non-Bayesian predictive inference, Atti della XLV Riunione Scientifica della SIS, Padova, 16–18 June 2010 (2010) 1–8.
[58] I. Smith, A. Ferrari, Equivalence between the posterior distribution of the likelihood ratio and a p-value in an invariant frame, Bayesian Anal. 9 (4) (2014) 939–962.
[59] L. Jaulin, Probabilistic set-membership approach for robust regression, J. Stat. Theory Pract. 4 (2010) 155–167.
[60] P.J. Rousseeuw, A.M. Leroy, Robust Regression and Outlier Detection, John Wiley & Sons, 1987.
[61] R.E. Moore, Parameter sets for bounded-error data, Math. Comput. Simul. 34 (2) (1992) 113–119.
[62] R. Hettich, K.O. Kortanek, Semi-infinite programming: theory, methods and applications, SIAM Rev. 35 (3) (1993) 380–429.
[63] M. Lopez, G. Still, Semi-infinite programming, Eur. J. Oper. Res. 180 (2007) 491–518.
[64] M. Tawarmalani, N. Sahinidis, Convexification and Global Optimization in Continuous and Mixed-Integer Nonlinear Programming: Theory, Algorithms, Software, and Applications, Kluwer Academic Publishers, 2002.
[65] A. Neumaier, Complete search in continuous global optimization and constraint satisfaction, Acta Numer. 13 (2004) 271–369.
[66] R. Misener, C.A. Floudas, ANTIGONE: algorithms for continuous/integer global optimization of nonlinear equations, J. Glob. Optim. 59 (2014) 503–526.
[67] J.W. Blankenship, J.E. Falk, Infinitely constrained optimization problems, J. Optim. Theory Appl. 19 (2) (1976) 261–281.
[68] C.A. Floudas, O. Stein, The adaptive convexification algorithm: a feasible point method for semi-infinite programming, SIAM J. Optim. 18 (4) (2007) 1187–1208.
[69] A. Mitsos, P. Lemonidis, C.K. Lee, P.I. Barton, Relaxation-based bounds for semi-infinite programs, SIAM J. Optim. 19 (1) (2008) 77–113.
[70] S. Boyd, L. El Ghaoui, E. Feron, V. Balakrishnan, Linear Matrix Inequalities in System and Control Theory Studies in Applied Mathematics, vol. 15, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, 1994.
[71] D. Ratkowsky, R. Lowry, T. McMeekin, A. Stokes, R. Chandler, A model for bacterial culture growth rate throughout the entire biokinetic temperature range, J. Bacteriol. 154 (1983) 1222–1226.
[72] J. Lobry, L. Rosso, J. Flandrois, A fortran subroutine for the determination of parameter confidence limits in non-linear models, Binary 3 (1991) 86–93.
[73] M. Tawarmalani, N.V. Sahinidis, A polyhedral branch-and-cut approach to global optimization, Math. Progr. 103 (2005) 225–249.
[74] A. Mitsos, B. Chachuat, P.I. Barton, Towards global bilevel dynamic optimization, J. Glob. Optim. 45 (1) (2009) 63–93.
[75] B. Chachuat, A. Singer, P. Barton, Global methods for dynamic optimization and mixed-integer dynamic optimization, Ind. Eng. Chem. Res. 45 (2006) 8373–8392.
[76] Y. Lin, M.A. Stadtherr, Deterministic global optimization of nonlinear dynamic systems, AIChE J. 53 (2007) 866–875.

Computers and Chemical Engineering 79 (2015) 30–45

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Improving scenario decomposition algorithms for robust nonlinear model predictive control

Rubén Martí[a,*], Sergio Lucia[b], Daniel Sarabia[c], Radoslav Paulen[b], Sebastian Engell[b], César de Prada[a]

[a] Department of Systems Engineering and Automatic Control, University of Valladolid, c/ Real de Burgos s/n, 47011 Valladolid, Spain
[b] Process Dynamics and Operations Group, Technische Universität Dortmund, Emil-Figge-Str. 70, 44221 Dortmund, Germany
[c] Department of Electromechanical Engineering, Escuela Politécnica Superior, University of Burgos, Avda. Cantabria s/n, Spain

ARTICLE INFO

Article history:
Received 13 December 2014
Received in revised form 29 March 2015
Accepted 24 April 2015
Available online 5 May 2015

Keywords:
Economic model predictive control
Uncertainty
Robust control
Distributed computing
Optimization

ABSTRACT

This paper deals with the efficient computation of solutions of robust nonlinear model predictive control problems that are formulated using multi-stage stochastic programming via the generation of a scenario tree. Such a formulation makes it possible to consider explicitly the concept of recourse, which is inherent to any receding horizon approach, but it results in large-scale optimization problems. One possibility to solve these problems in an efficient manner is to decompose the large-scale optimization problem into several subproblems that are iteratively modified and repeatedly solved until a solution to the original problem is achieved. In this paper we review the most common methods used for such decomposition and apply them to solve robust nonlinear model predictive control problems in a distributed fashion. We also propose a novel method to reduce the number of iterations of the coordination algorithm needed for the decomposition methods to converge. The performance of the different approaches is evaluated in extensive simulation studies of two nonlinear case studies.

1. Introduction

The use of optimization for the control of chemical processes is a standard technique in the process industry because it makes it possible to take into account various process objectives, such as economic ones, and calculate the best control actions subject to constraints that arise from quality, safety or environmental requirements (Engell, 2007). A family of the methods termed model predictive control (MPC) approaches evolved in this respect.

In recent years, a significant amount of research has been devoted to the use of economic cost functions within the framework of nonlinear model predictive control (NMPC), see e.g. Rawlings and Amrit (2009), Idris and Engell (2012), Prada et al. (2008), Gopalakrishnan and Biegler (2013). The optimal operation of a system according to an economic cost corresponds usually to driving the system to its constraints. For this reason, plant model mismatch or disturbances (which are always present in reality) can lead easily to constraint violations and the explicit consideration of uncertainty in the design of the MPC controller becomes very important.

The first efforts in robust MPC tried to address this problem focusing on the so-called min-max MPC (Campo and Morari, 1987). This approach obtains a sequence of control inputs that minimizes the cost of the worst-case realization of the uncertainty while satisfying the constraints for all the cases of the uncertainty. Min-max MPC however does not take into account the fact that new information will be available in the future and therefore the result may be overly conservative and may lead to infeasible optimization problems, as illustrated in Scokaert and Mayne (1998). Different methods such as the closed-loop (or feedback) min-max NMPC in Lee and Yu (1997) and Mayne (2001), or tube-based MPC (Mayne et al., 2005; Rawlings and Amrit, 2009; Rakovic et al., 2011) have been proposed to overcome the limitations of open-loop min-max MPC. However, most of the methods above cannot be applied to realistic problems, because they result in prohibitive computational cost, or they cannot be easily designed for nonlinear systems or because they are very conservative.

A different possibility is to consider the integration of the stochastic programming paradigm (Birge, 1997; Shapiro, 2009) into the framework of model predictive control. This idea has been used for the linear MPC case in Scokaert and Mayne (1998),

* Corresponding author. Tel.: +34 600880519.
E-mail addresses: ruben@autom.uva.es (R. Martí), sergio.lucia@bci.tu-dortmund.de (S. Lucia), dsarabia@ubu.es (D. Sarabia), radoslav.paulen@bci.tu-dortmund.de (R. Paulen), sebastian.engell@bci.tu-dortmund.de (S. Engell), prada@autom.uva.es (C. de Prada).

http://dx.doi.org/10.1016/j.compchemeng.2015.04.024
0098-1354/

Muñoz de la Peña et al. (2005), Bernardini and Bemporad (2009) as well as to NMPC (Lucia et al., 2013a), denoted as multi-stage NMPC which is a robust NMPC approach based on the assumption that the uncertainty can be modeled by a scenario tree. The application of multi-stage NMPC has recently provided very promising results and it is the approach to robust NMPC followed in this work.

The main drawback of the approach is that the size of the resulting optimization problem grows exponentially with the length of the prediction horizon, and with the number of uncertainties as well as with the number of different values of each uncertainty that is considered in the design of the scenario tree. For this reason, an efficient solution of the resulting Nonlinear Programming (NLP) problem is necessary to make it possible to solve such problems in real time. The main goal of this paper is to analyze the different possibilities to solve the multi-stage NMPC in an efficient way, both in terms of computation time and memory requirements.

For this purpose, we take advantage of the fact that each scenario in the scenario tree is an independent subproblem except for the non-anticipativity (or causality) constraints which make sure that the optimal inputs do not anticipate the realization of the uncertainty in the present sampling instant since it is unknown. We investigate different possibilities to relax these constraints and to solve the individual optimization problems within a coordination algorithm that enforces satisfaction of the non-anticipativity constraints upon its convergence.

We present classical coordination algorithms in the framework of economic multi-stage NMPC and propose a novel distributed algorithm that uses sensitivity information to reduce the number of iterations of the coordination algorithm needed to converge. The different distributed multi-stage NMPC algorithms are analyzed via extensive simulation studies of two industrial case studies, extending the results presented in Marti et al. (2015).

The remainder of the paper is organized as follows. Section 2 summarizes the concept of multi-stage NMPC. The different coordination algorithms used in this work are presented in Section 3 and a possible modification to achieve convergence in less iterations of the coordination algorithms is presented in Section 4. The resulting distributed NMPC algorithms are evaluated in Section 5 using an industrial hydrodesulphurization example and in Section 6 using an industrial polymerization reactor. The paper is concluded in Section 7.

2. Multi-stage NMPC

This section reviews the main concepts of the multi-stage NMPC approach presented in Lucia et al. (2013a, 2014b).

In multi-stage NMPC, the model uncertainty is taken into account by considering a tree of discrete scenarios for each possible value of the uncertainty as depicted in Fig. 1. The formulation of a scenario tree makes it possible to take explicitly into account that the future decisions can depend on the new information (measurements) that will become available in the future. Thus the future control inputs can be adapted according to the future realizations of the uncertainty and the conservativeness of the approach is reduced compared to other robust methods that search for a single sequence of control inputs to satisfy the constraints for all the possible values of the uncertainty. Formulating the uncertain decision process as a scenario tree is a well-known approach in the field of multi-stage stochastic programming, which has been extensively used in decision theory and finances (Shapiro, 2009). In the case that the uncertainty is truly discrete-valued, this is the best solution possible for a given prediction horizon. Generally this is not the case, and multi-stage NMPC is an approximation of the best solution.

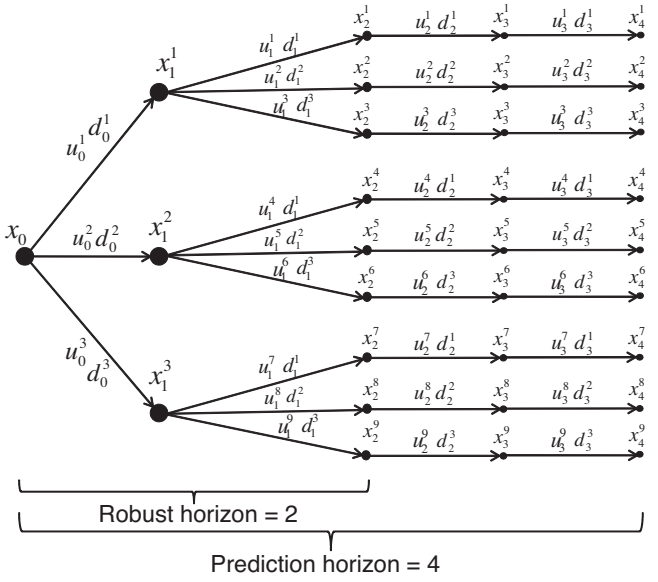


Fig. 1. Scenario tree representation of the uncertainty evolution for multi-stage NMPC.

To formulate mathematically the multi-stage NMPC approach, we consider a discrete-time nonlinear system:

$$\boldsymbol{x}_{k+1}^{j} = \boldsymbol{f}(\boldsymbol{x}_{k}^{p(j)}, \boldsymbol{u}_{k}^{j}, \boldsymbol{d}_{k}^{r(j)}), \tag{1a}$$

where each state vector $\boldsymbol{x}_{k+1}^{j} \in \mathbb{R}^{n_x}$ at stage $k+1$ and position j depends on the parent state (node) $\boldsymbol{x}_{k}^{p(j)}$ at stage k, the vector of control inputs $\boldsymbol{u}_{k}^{j} \in \mathbb{R}^{n_u}$ and the corresponding realization r of the uncertainty $\boldsymbol{d}_{k}^{r(j)} \in \mathbb{R}^{n_d}$ (e.g. in Fig. 1, $\boldsymbol{x}_{2}^{6} = \boldsymbol{f}(\boldsymbol{x}_{1}^{2}, \boldsymbol{u}_{1}^{6}, \boldsymbol{d}_{1}^{3})$). The uncertainty at the stage k is defined by $\boldsymbol{d}_{k}^{r(j)} \in \{\boldsymbol{d}_{k}^{1}, \boldsymbol{d}_{k}^{2}, \ldots, \boldsymbol{d}_{k}^{s}\}$ for s different possible combinations of values of the uncertainty. We define the set of indices (j, k) in the scenario tree as $\boldsymbol{I}$. $\boldsymbol{S}_i$ denotes the ith scenario defined as the path from the root node $\boldsymbol{x}_0$ to one of the leaf nodes and it contains all the states $\boldsymbol{x}_{k}^{j}$ and control inputs $\boldsymbol{u}_{k}^{j}$ that belong to the ith scenario.

A common way to build a scenario tree is to consider, as possible branches, a combination of values among the extreme and nominal values of all the uncertainties. For the general nonlinear case, it is not guaranteed that this results in robust constraint satisfaction for the values of the uncertainty that are not considered in the tree, but it has been shown to give very good results in practice (Lucia et al., 2012, 2013a, 2014a,b). If a rigorous guarantee for robust constraint satisfaction of all the possible values of the uncertainty (including those that are not in the tree) is required, the multi-stage approach can be combined with reachability analysis as shown in Lucia et al. (2014c).

Generating the scenario tree in a systematic way (considering the extrema of the uncertainty space) makes the size of the resulting optimization problem to grow rapidly with increasing length of prediction horizon N_p and with increasing number of uncertainties with resulting number of scenarios $N = s^{N_p n_d}$. A possible strategy to avoid the exponential growth of the scenario tree over the prediction horizon is to consider that the uncertainty remains constant after a certain stage (called robust horizon N_r) until the end of prediction horizon (Fig. 1) which gives $N = s^{N_r n_d}$.

The optimization problem that has to be solved at each sampling instant can be written as:

$$\min_{\boldsymbol{x}_{k+1}^{j}, \boldsymbol{u}_{k}^{j}, \ \forall (j,k) \in \boldsymbol{I}} \ \sum_{i=1}^{N} \omega_i J_i(\boldsymbol{X}_i, \boldsymbol{U}_i) \tag{2a}$$

subject to:

$$\boldsymbol{x}^j_{k+1} = \boldsymbol{f}(\boldsymbol{x}^{p(j)}_k, \boldsymbol{u}^j_k, \boldsymbol{d}^{r(j)}_k), \quad \forall \ (j, k+1) \in \boldsymbol{I}, \tag{2b}$$

$$0 \geq \boldsymbol{g}(\boldsymbol{x}^j_{k+1}, \boldsymbol{u}^j_k, \boldsymbol{d}^{r(j)}_k), \quad \forall \ (j, k) \in \boldsymbol{I}, \tag{2c}$$

$$\boldsymbol{u}^j_k = \boldsymbol{u}^l_k \quad \text{if} \boldsymbol{x}^{p(j)}_k = \boldsymbol{x}^{p(l)}_k, \quad \forall (j, k), (l, k) \in \boldsymbol{I}, \tag{2d}$$

where $\boldsymbol{X}_i$, $\boldsymbol{U}_i$ are the set of states and control inputs that belong to the scenario $\boldsymbol{S}_i$ with the probability of occurrence ω_i. $\boldsymbol{g}(\cdot)$ denotes the nonlinear constraints. The cost of each scenario is denoted by $J_i(\cdot)$ and can be written as:

$$J_i(\boldsymbol{X}_i, \boldsymbol{U}_i) := \sum_{k=0}^{N_p-1} L(\boldsymbol{x}^j_{k+1}, \boldsymbol{u}^j_k), \quad \forall \ \boldsymbol{x}^j_{k+1}, \boldsymbol{u}^j_k \in \boldsymbol{S}_i. \tag{3}$$

The constraints (2d) are called non-anticipativity constraints which imply that the control inputs cannot anticipate the realization of the uncertainty, i.e. the control inputs $\boldsymbol{u}^j_k$ that branch at the same parent node $\boldsymbol{x}^{p(j)}_k$ must be the same. Note that these constraints represent the coupling among the different scenarios.

The theoretical analysis of the stability properties of the NMPC controller is out of the scope of this paper and in order to simplify the presentation, we did not include any special provision for it, but terminal cost or constraints could have been added to take care of stability. Nevertheless, we check the stability of the controller by means of extensive simulation studies.

3. Distributed multi-stage NMPC based on scenario decomposition

Decomposition techniques can be used for solving large-scale optimization problems (see Birge (1997) or Ruszczynski (1997) for a review). In these approaches, the centralized problem is divided into smaller subproblems which are solved independently. For the case of multi-stage stochastic optimization problems, the different approaches are usually classified into scenario decomposition or stage decomposition. In the scenario decomposition approaches each smaller subproblem corresponds to a scenario of the tree and in the stage decomposition each subproblem corresponds to a time stage of the tree. In this work, the scenario decomposition approach is chosen because the only link between the different scenarios is the presence of the non-anticipativity constraints. Note that other techniques, such the ones based on Schur complement decomposition (Steinbach, 2000) can achieve better performance than the scenario decomposition approaches analyzed here since they can exploit the tree structure at the linear algebra level. However, they require a new implementation for its solution and general purpose solvers such as IPOPT (Wächter and Biegler, 2006) cannot be used directly. We therefore focus in this paper in decomposition techniques that can be implemented in a very simple manner using existing solvers.

Different scenario decomposition schemes have been used in several works (Lucia et al., 2013b; Marti et al., 2013a, 2014a; Patrinos et al., 2011). The main idea of any scenario decomposition approach is to decompose the centralized problem (2) into N smaller optimization problems for each scenario $i = 1, \ldots, N$ as illustrated in Fig. 2, which include copies of the variables (hence $\boldsymbol{u}^1_0, \boldsymbol{u}^2_0, \boldsymbol{u}^3_0$ in Fig. 2) that are common for the different scenarios. An optimization problem for each scenario can be formulated as:

$$\min_{\boldsymbol{x}^j_k, \boldsymbol{u}^j_k \ \forall (j,k) \in \boldsymbol{I}_i} \quad \omega_i J_i(\boldsymbol{X}_i, \boldsymbol{U}_i) \tag{4a}$$

subject to:

$$\boldsymbol{x}^j_{k+1} = \boldsymbol{f}(\boldsymbol{x}^{p(j)}_k, \boldsymbol{u}^j_k, \boldsymbol{d}^{r(j)}_k), \quad \forall \ (j, k+1) \in \boldsymbol{I}_i, \tag{4b}$$

Fig. 2. Scenario decomposition representation of the uncertainty evolution for multi-stage NMPC.

$$0 \geq \boldsymbol{g}(\boldsymbol{x}^j_{k+1}, \boldsymbol{u}^j_k, \boldsymbol{d}^{r(j)}_k), \quad \forall \ (j, k+1) \in \boldsymbol{I}_i. \tag{4c}$$

where $\boldsymbol{I}_i$ is the set of indices in the original tree of the nodes corresponding to scenario $\boldsymbol{S}_i$, that is $\boldsymbol{I}_i = \{(j,k) : \boldsymbol{x}^k_j \in \boldsymbol{S}_i\}$. This formulation results in the relaxation of the non-anticipativity constraints of the original multi-stage problem (2). Because of this, the solution obtained from the set of decomposed problems will not satisfy the non-anticipativity constraints and will not be a solution of the original multi-stage optimization problem (2). Decomposition approaches modify the decomposed optimization problem (4) so that (via an iterative algorithm) a solution is obtained which solves the original centralized optimization problem. This section describes different possibilities of implementing such an algorithm.

3.1. Augmented Lagrangean decomposition

One of the most used methods to decompose the optimization problem (2) is based on the augmented Lagrangean technique. The main idea of the approach is to enforce the non-anticipativity constraints by penalizing their violation, i.e. the difference between the control inputs that should satisfy the non-anticipativity constraints and an aggregated (assumed) common control input. In this manner, the scenarios can be solved as independent subproblems. Then the procedure is repeated until the non-anticipativity constraints are fulfilled to the specified threshold. The decomposition algorithm used in this work is the progressive hedging algorithm of Rockafellar and Wets (1991). It can be applied to nonlinear non-convex problems and it has been also proven, in Rockafellar and Wets (1991), that the solution of the algorithm always converges to a local optimum of the original cost function, if convergence of coordination is achieved. Instead of solving the original centralized problem defined in (2), the progressive hedging algorithm solves an independent problem for each scenario i that can be written as:

$$\min_{\boldsymbol{x}^j_k, \boldsymbol{u}^j_k \forall (j,k) \in \boldsymbol{I}_i} \quad \omega_i J_i(\boldsymbol{X}_i, \boldsymbol{U}_i) + \sum_{k=0}^{N_r-1} [\boldsymbol{\lambda}^{j,T}_k (\boldsymbol{u}^j_k - \hat{\boldsymbol{u}}^j_k) + \rho/2 \|\boldsymbol{u}^j_k - \hat{\boldsymbol{u}}^j_k\|^2_2] \tag{5a}$$

subject to:

$$\boldsymbol{x}^j_{k+1} = \boldsymbol{f}(\boldsymbol{x}^{p(j)}_k, \boldsymbol{u}^j_k, \boldsymbol{d}^{r(j)}_k), \quad \forall \ (j, k+1) \in \boldsymbol{I}_i, \tag{5b}$$

$$0 \leq \boldsymbol{g}(\boldsymbol{x}^j_{k+1}, \boldsymbol{u}^j_k, \boldsymbol{d}^{r(j)}_k), \quad \forall \ (j, k+1) \in \boldsymbol{I}_i, \tag{5c}$$

where $\hat{\boldsymbol{u}}^j_k \in \mathbb{R}^{n_u}$ is the aggregated value towards which the control input at stage k and position in the tree j should converge for each

of the scenarios in which the non-anticipativity constraints have to be satisfied. Note that the sum in the cost function is done until the stage $k = N_r - 1$ because the non-anticipativity constraints are applied only until that point in the scenario tree.

The choice of $\hat{\boldsymbol{u}}_k^j$ at each iteration greatly influences the performance of the algorithm. A simple approach to obtain $\hat{\boldsymbol{u}}_k^j$ is to use the averaged value of the control inputs that have to satisfy each non-anticipativity constraint when the scenarios are solved independently, that is,

$$\hat{\boldsymbol{u}}_k^j = \sum_{\forall \boldsymbol{u}_k^l | \boldsymbol{x}_k^{p(j)} = \boldsymbol{x}_k^{p(l)}} \pi^{r(l)} \boldsymbol{u}_k^l, \tag{6}$$

where $\pi^{r(l)} \in \mathbb{R}$ denotes the probability associated with the branch in the scenario tree at which the control input $\boldsymbol{u}_k^l$ is located. Note that the probabilities $\pi^{r(l)}$ are different from the scenario probabilities ω_i which are calculated as the product of the probabilities associated with the branches from the root node until the leaf node. The expression in (6) means that the aggregated values are calculated as the weighted sum of the control inputs which share the same parent, that is, for the nodes where the non-anticipativity constraints apply. The idea of the algorithm is that after several iterations of the coordination algorithm all the non-anticipativity constraints are satisfied with a desired tolerance ϵ_{tol} and therefore the augmented cost function converges towards the original one.

The parameters $\boldsymbol{\lambda}_k^j \in \mathbb{R}^{n_u}$ and $\rho \in \mathbb{R}$ are updated at each iteration of the coordination algorithm to improve the convergence. A usual update rule for $\boldsymbol{\lambda}_k^j$ is: $\boldsymbol{\lambda}_k^j \leftarrow \boldsymbol{\lambda}_k^j + \rho(\boldsymbol{u}_k^j - \hat{\boldsymbol{u}}_k^j)$. The parameter ρ is usually increased after each iteration of the coordination algorithm, with a maximum value $\rho_{\max} \in \mathbb{R}$ that avoids excessively large penalty terms $\rho \leftarrow \min(\gamma\rho, \rho_{\max})$, where $\gamma \in \mathbb{R}$ is a parameter that determines the increase of ρ. The algorithm used in this work for the distributed multi-stage NMPC with augmented Lagrangean methods is summarized in Algorithm 1. Further discussion about the updates of the parameters and about the convergence properties of the algorithm can be found in Rockafellar and Wets (1991).

Algorithm 1. Progressive Hedging Algorithm

Require: $\rho_{\max} > 0$, $\gamma > 1$; $\rho = \rho_{\text{ini}}$; $\hat{\boldsymbol{u}}_k^j = 0$; $\boldsymbol{\lambda}_k^j = 0$, $\forall (j,k) \in \boldsymbol{I}$
  Solve problem (5) $\forall i = 1, \ldots, N$
  $\hat{\boldsymbol{u}}_k^j = \sum_{\boldsymbol{u}_k^l | \boldsymbol{x}_k^{p(j)} = \boldsymbol{x}_k^{p(l)}} \pi^{r(l)} \boldsymbol{u}_k^l$
  while $\|\boldsymbol{u}_k^j - \hat{\boldsymbol{u}}_k^j\|_2^2 \geq \epsilon_{\text{tol}}$, $\forall (j,k) \in \boldsymbol{I}$ **do**
    for $i = 1 : N$ **do**
      $\boldsymbol{\lambda}_k^j \leftarrow \boldsymbol{\lambda}_k^j + \rho \left(\boldsymbol{u}_k^j - \hat{\boldsymbol{u}}_k^j \right)$, $\forall (j,k) \in \boldsymbol{I}_i$
      Solve problem (5)
    end for
    $\hat{\boldsymbol{u}}_k^j = \sum_{\boldsymbol{u}_k^l | \boldsymbol{x}_k^{p(j)} = \boldsymbol{x}_k^{p(l)}} \pi^{r(l)} \boldsymbol{u}_k^l$
    $\rho \leftarrow \min(\gamma\rho, \rho_{\max})$
  end while
  return $\boldsymbol{u}_k^j$ $\forall (j,k) \in \boldsymbol{I}$

One of the main disadvantages of this algorithm is that its performance depends on the value of the tuning parameters ρ, $\rho_{\max}$ and γ. Moreover, the approach is only globally convergent when tuned properly. To avoid the tuning process, a price-driven coordination is implemented in this work.

3.2. Price-driven coordination

The price-driven coordination approach from Jose and Ungar (1998) is suitable for solving resource distribution or auction problems. With this method a large-scale optimization problem can be decomposed into subproblems by relaxing the resource constraints which connect the subproblems together and penalizing the deviation in the constraints in the cost function increasing prices. The method presented here uses a price-adjustment algorithm based on Newton's method for updating the price or the Lagrange multiplier. For instance, this method has been implemented to coordinate a decentralized NMPC controller in an oxygen distribution network (Marti et al., 2013b).

The coordination process is similar to setting up the prices for selling common resources to different consumers. The transformation of non-anticipativity constraints, which correspond to equality constraints, to resource constraints was shown in Marti et al. (2014b).

The decomposed optimization problem solved by the price-driven coordination algorithm can be written as:

$$\min_{\boldsymbol{x}_k^j, \boldsymbol{u}_k^j \forall (j,k) \in \boldsymbol{I}_i} \omega_i J_i(\boldsymbol{X}_i, \boldsymbol{U}_i) + \sum_{k=0}^{N_r - 1} [\boldsymbol{\lambda}_k^{j,T} \|\boldsymbol{u}_k^j - \hat{\boldsymbol{u}}_k^j\| + q \|\boldsymbol{u}_k^j - \hat{\boldsymbol{u}}_k^j\|_2^2] \tag{7a}$$

subject to:

$$\boldsymbol{x}_{k+1}^j = \boldsymbol{f}(\boldsymbol{x}_k^{p(j)}, \boldsymbol{u}_k^j, \boldsymbol{d}_k^{r(j)}), \quad \forall \ (j, k+1) \in \boldsymbol{I}_i, \tag{7b}$$

$$0 \leq \boldsymbol{g}(\boldsymbol{x}_{k+1}^j, \boldsymbol{u}_k^j, \boldsymbol{d}_k^{r(j)}), \quad \forall \ (j, k+1) \in \boldsymbol{I}_i. \tag{7c}$$

The main difference with respect to the augmented Lagrangean approach is in the way the prices $\boldsymbol{\lambda}_k^j$ are updated, in this case:

$$\boldsymbol{\lambda}_k^j \leftarrow \boldsymbol{\lambda}_k^j + \left(\frac{\partial \boldsymbol{u}_k^j}{\partial \boldsymbol{\lambda}_k^j}\right)^{-1} (\boldsymbol{u}_k^j - \hat{\boldsymbol{u}}_k^j). \tag{8}$$

In this method, the coordinator sends a vector of prices ($\boldsymbol{\lambda}_k^j$) of the resources to every subsystem. After solving local optimization problems, with q as a regularization parameter (e.g., the small scalar q is chosen as 0.01 in Jose and Ungar (2000) for solving linear optimization and can be set to 0 for solving nonlinear problems, as is shown in Jose and Ungar (2000)), the subproblems inform the coordinator about the resource demands ($\boldsymbol{u}_k^j$) at current prices $\boldsymbol{\lambda}_k^j$ and their responses to the price change ($\partial \boldsymbol{u}_k^j / \partial \boldsymbol{\lambda}_k^j$). The coordinator then collects these two pieces of information to update the price. The derivative information $\partial \boldsymbol{u}_k^j / \partial \boldsymbol{\lambda}_k^j$ of local subproblems can be obtained by sensitivity analysis of the optimal local solutions.

The crucial step of this method lies in the calculation of the parametric sensitivity. In particular, in this work the procedure for parametric sensitivity analysis of nonlinear optimization problem is based on sIPOPT (Pirnay et al., 2012). However, this information can be obtained without the need to employ sIPOPT as it is shown in Marti et al. (2013a).

The price-driven algorithm for distributed multi-stage NMPC is summarized in Algorithm 2.

Algorithm 2. Price-driven Coordination Algorithm

Require: $\hat{\boldsymbol{u}}_k^j = 0$; $\boldsymbol{\lambda}_k^j = 0$, $\forall (j,k) \in \boldsymbol{I}$
  Solve problem (7) $\forall i = 1, \ldots, N$
  $\hat{\boldsymbol{u}}_k^j = \sum_{\boldsymbol{u}_k^l | \boldsymbol{x}_k^{p(j)} = \boldsymbol{x}_k^{p(l)}} \pi^{r(l)} \boldsymbol{u}_k^l$
  while $\|\boldsymbol{u}_k^j - \hat{\boldsymbol{u}}_k^j\|_2^2 \geq \epsilon_{\text{tol}}$, $\forall (j,k) \in \boldsymbol{I}$ **do**
    for $i = 1 : N$ **do**
      $\boldsymbol{\lambda}_k^j \leftarrow \boldsymbol{\lambda}_k^j + \left(\frac{\partial \boldsymbol{u}_k^j}{\partial \boldsymbol{\lambda}_k^j}\right)^{-1} \left(\boldsymbol{u}_k^j - \hat{\boldsymbol{u}}_k^j \right)$, $\forall (j,k) \in \boldsymbol{I}_i$
      Solve problem (7)
    end for
    $\hat{\boldsymbol{u}}_k^j = \sum_{\boldsymbol{u}_k^l | \boldsymbol{x}_k^{p(j)} = \boldsymbol{x}_k^{p(l)}} \pi^{r(l)} \boldsymbol{u}_k^l$
  end while
  return $\boldsymbol{u}_k^j$ $\forall (j,k) \in \boldsymbol{I}$

3.3. Sensitivity-based decomposition

Even though the algorithms presented in the previous section reduce the necessary memory for solving the problem (2) compared to a centralized approach, they might require a relatively big number of iterations of the coordination algorithm to converge and therefore the use of these methods might often result in a significant increase of computation time compared to the solution of the centralized problem. We propose here a novel modification for both presented coordination algorithms that aims at reducing the number of iterations of the coordination algorithm needed to achieve a solution that satisfies the non-anticipativity constraints. This modification is based on the use of problem-specific information to calculate the values of the aggregate variables $\hat{\boldsymbol{u}}_k^j$ instead of using a weighted sum of the local control actions as it is usually done.

When an economic objective is used in the cost function of an NMPC controller, the optimal solution is either unconstrained or it lies in the region of active path constraints. This feature stresses the importance of robustifying such a controller. When solving the decomposed multi-stage NMPC problem, the activity of the path constraints causes an inability of some of the locally calculated control inputs (independently for each scenario) to move towards the aggregated control input as this would cause violations of the constraints. This makes the number of iterations of the coordination algorithm needed to converge to a solution that satisfies the non-anticipativity constraints to grow significantly. We propose to calculate the aggregated variables $\hat{\boldsymbol{u}}_k^j$ by taking into account the possible constraint violations that a modification of the local control inputs would cause.

We take into account this in a systematic manner by solving at each iteration of the coordination algorithm, the following quadratic program once the optimal (decentralized) local solutions $\boldsymbol{x}_{k+1}^{j*}, \boldsymbol{u}_k^{j*}$ have been obtained:

$$\min_{\hat{\boldsymbol{u}}_k^j, \boldsymbol{\varepsilon} \geq \mathbf{0}} \mathbf{1}^T \boldsymbol{\varepsilon} + \sum_{(j,k)\in \boldsymbol{I}} \alpha_k^j \|\hat{\boldsymbol{u}}_k^j - \boldsymbol{u}_k^{j*}\|_2^2 \tag{9a}$$

subject to:

$$\boldsymbol{\varepsilon} \geq \boldsymbol{g}(\boldsymbol{x}_{k+1}^{j*}, \boldsymbol{u}_k^{j*}, \boldsymbol{d}_k^{r(j)}) + \frac{\partial \boldsymbol{g}}{\partial \boldsymbol{u}}|_{\boldsymbol{x}_{k+1}^{j*}, \boldsymbol{u}_k^{j*}, \boldsymbol{d}_k^{r(j)}} (\hat{\boldsymbol{u}}_k^j - \boldsymbol{u}_k^{j*}), \quad \forall (j,k) \in \boldsymbol{I}, \tag{9b}$$

where $\boldsymbol{\varepsilon} \in \mathbb{R}^{n_g}$ (n_g is the number of constraints) is a vector of slack variables that at the optimal solution has the value of the biggest constraint violation (scenario-wise) that would be (approximately) caused by moving the control inputs from the current optimal local solution ($\boldsymbol{u}_k^{j*}$) towards the current value of aggregated variable $\hat{\boldsymbol{u}}_k^j$ at each point along the prediction horizon. The expression $\partial\, \boldsymbol{g}/\partial\, \boldsymbol{u}$ represents sensitivities of the optimal value of constraint functions w.r.t. the optimal control input. The degrees of freedom of the quadratic program are the aggregated variables $\hat{\boldsymbol{u}}_k^j$ that are used in the objectives of the local problems for the next iteration of the coordination algorithm. The main idea of the optimization problem (9) is to find the aggregated variables $\hat{\boldsymbol{u}}_k^j$ such that the worst-case constraint violation for the local subproblems is minimum (given by the optimal value of $\boldsymbol{\varepsilon}$). This is equivalent to find new weights for each sceanrio to calculate the aggregated variables as a weighted sum. Since the convergence properties are developed for any value of the weights (Rockafellar and Wets, 1991), they extend to this case as well. We introduce regularization terms in the cost function weighted by α_k^j to regularize the solution for the case when no constraints violations occur, i.e. when the optimal solution is unconstrained. We propose to choose α_k^j according to the relative cost of each subproblem as:

$$\alpha_k^j = \left| \frac{\omega_i \boldsymbol{J}_i}{\sum_{i=1}^N \omega_i \boldsymbol{J}_i} \right|, \quad \forall (j,k) \in \boldsymbol{I}_i. \tag{10}$$

The employed regularization forces $\hat{\boldsymbol{u}}_k^j$ to move towards the control inputs that achieve the worst cost but are more likely to occur thus aims at improving the worst cost of the most probable scenario. Note that at the convergence of the decomposition algorithm $\hat{\boldsymbol{u}}_k^j = \boldsymbol{u}_k^{j*}$, $\forall (j,k) \in \boldsymbol{I}$ and the second term of the cost function (9a) vanishes. We will show in the simulation studies of the next section that the calculation of the aggregated variables using this quadratic program instead of a simple weighted average helps to decrease significantly the number of iterations of the coordination algorithm needed by both the augmented Lagrangean and the price-based decomposition methods. The main reason for this improvement is that more information about the centralized problem is given to the iterative algorithms via aggregated variables. In addition, this improvement is achieved at a very low cost, since the only additional requirement is to solve the quadratic program (9), which can be solved efficiently using relatively small amount of memory.

The sensitivity-based update of the aggregated variables can be applied to solve the distributed multi-stage NMPC problem using the (a) Augmented Lagrangean or the (b) Price-driven coordination methods as it is summarized in Algorithm 3.

Sensitivity-based distributed multi-stage NMPC using (a) Augmented Lagrangean or (b) price-driven algorithm

Require: $\hat{\boldsymbol{u}}_k^j = \mathbf{0}; \boldsymbol{\lambda}_k^j = \mathbf{0}, \forall (j,k) \in \boldsymbol{I}$; for (a) $\rho_{\max} > 0, \gamma > 1; \rho = \rho_{\text{ini}}$;
  Solve (a) problem (5) or (b) problem (7) $\forall i = 1, \ldots, N$
    Calculate the aggregated variables $\hat{\boldsymbol{u}}_k^j$ by solving the quadratic program (9)
  while $\|\boldsymbol{u}_k^j - \hat{\boldsymbol{u}}_k^j\|_2^2 \geq \epsilon_{\text{tol}}, \forall (j,k) \in \boldsymbol{I}$ **do**
    for $i = 1 : N$ **do**
      (a) $\boldsymbol{\lambda}_k^j \leftarrow \boldsymbol{\lambda}_k^j + \rho \left(\boldsymbol{u}_k^j - \hat{\boldsymbol{u}}_k^j\right), \ \forall (j,k) \in \boldsymbol{I}_i$
        Solve problem (5)
      (b) $\boldsymbol{\lambda}_k^j \leftarrow \boldsymbol{\lambda}_k^j + \left(\frac{\partial u_k^j}{\partial \lambda_k^j}\right)^{-1} \left(\boldsymbol{u}_k^j - \hat{\boldsymbol{u}}_k^j\right) \ \forall (j,k) \in \boldsymbol{I}_i$
        Solve problem (7)
    end for
    Calculate the aggregated variables $\hat{\boldsymbol{u}}_k^j$ by solving the quadratic program (9)
    (a) $\rho \leftarrow \min(\gamma\rho, \rho_{\max})$
  end while
  return $\boldsymbol{u}_k^j \forall (j,k) \in \boldsymbol{I}$

4. Distributed multi-stage NMPC based on bundle decomposition

One possibility to further reduce the number of the iterations of the coordination algorithm needed is to consider a hybrid approach between a centralized method and a full scenario decomposition. In this case, the subproblems that are solved in each optimization problem consist not only of one scenario but of a subset or *bundle* of scenarios as illustrated in Fig. 3.

Each bundle $\boldsymbol{B}_b$ contains some scenarios $\boldsymbol{S}_i$ and they are chosen such that $\cup_{b=1}^{n_b} \boldsymbol{B}_b = \cup_{i=1}^{N} \boldsymbol{S}_i$, where n_b is the number of bundles, and no scenarios are repeated in the bundles ($\cap_{b=1}^{n_b} \boldsymbol{B}_b = \emptyset$). Instead of solving the fully decomposed problem presented in (4), the optimization problem that has to be solved at each sampling instant for the bundle decomposition approach can be written as:

$$\min_{\boldsymbol{x}_k^j, \boldsymbol{u}_k^j \in \boldsymbol{B}_b \forall i | \boldsymbol{S}_i \in \boldsymbol{B}_b} \sum \omega_i J_i(\boldsymbol{X}_i, \boldsymbol{U}_i) \tag{11a}$$

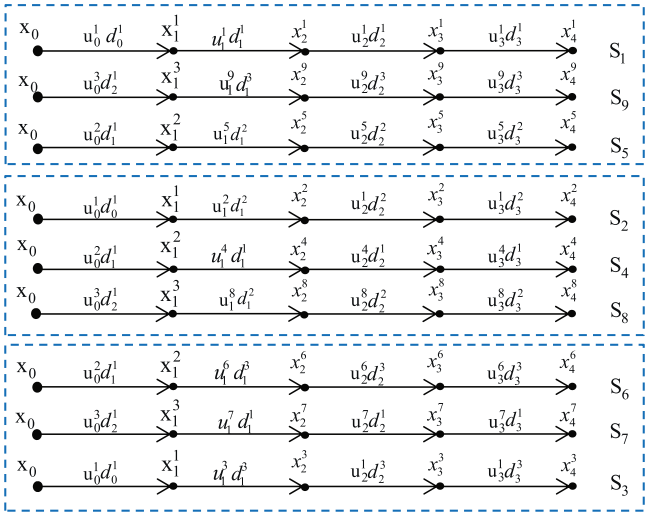


Fig. 3. Scenario tree representation of the uncertainty evolution using bundle decomposition for multi-stage NMPC. The scenarios in the same box represent the bundles.

subject to:

$$\boldsymbol{x}^j_{k+1} = \boldsymbol{f}(\boldsymbol{x}^{p(j)}_k, \boldsymbol{u}^j_k, \boldsymbol{d}^{r(j)}_k), \quad \forall \ \boldsymbol{x}^j_{k+1} \in \boldsymbol{B}_b, \tag{11b}$$

$$0 \geq \boldsymbol{g}(\boldsymbol{x}^j_{k+1}, \boldsymbol{u}^j_k, \boldsymbol{d}^{r(j)}_k), \quad \forall \ \boldsymbol{x}^j_{k+1} \in \boldsymbol{B}_b. \tag{11c}$$

The Augmented Lagrangean and Price-Driven Coordination method (explained above) can be used as well to coordinate the fulfillment of the non-anticipativity constraints between the different bundles. The use of bundles can significantly reduce the necessary number of iterations of the coordination algorithm and, despite solving larger subproblems, may result in faster computation times than the full decomposition or centralized approach.

5. Case study 1: Hydrodesulphurization process

The first case study under consideration consists of an industrial hydrodesulphurization plant (HDS), which is used to remove sulphur from hydrocarbons to fulfill environmental regulations. To do this, hydrogen is put in contact with the hydrocarbon in a fixed bed reactor with a catalyst (Bellos and Papayannakos, 2003). The optimal management of the hydrogen provided is very important in order to operate efficiently: if the quantity of hydrogen supplied is less than the minimum required, the expensive catalyst used in the desulphurization reactors can suffer important damage, while if the supply is in excess, significant economic losses will occur (Navia et al., 2014).

To understand the process, let us consider the simplified structure of the core part of an HDS plant represented in Fig. 4. It can be seen that the hydrogen comes from three sources: H_4, H_3 and *LP*. H_4 and H_3 are collectors transporting hydrogen manufactured in especially dedicated production units. Each production unit can generate hydrogen at different quantities and purity levels. On the other hand, the *LP* source is a recirculated stream with lower hydrogen concentration. The mixture goes through a compressor (C-1) and then is fed together with the hydrocarbon stream (FC), to a packed bed reactor (R_1, R_2). The products of the reaction are separated using a separation unit (T_1). One part of the excess of hydrogen feed is recirculated to the reactors while the rest (F_{10}) is purged in order to maintain a minimal hydrogen purity due to the presence of light ends.

The key operation is performed in the reactor to eliminate the undesired sulphur down to a given level at the plant output.

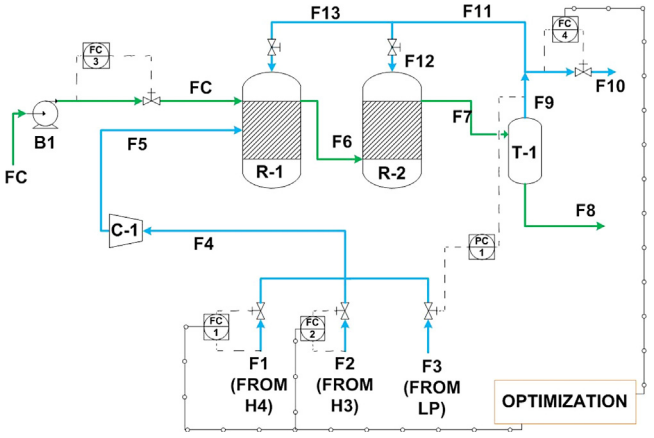


Fig. 4. Diagram of the hydrodesulphurization unit with decision variables.

Operators adjust the total hydrogen supply to the reactor, its temperature, etc. to attain this target, so that a given mode of operation implies a certain hydrogen consumption in the reactor according to the load conditions. This is a sub-system where settling time can be in the order of hours. To carry out the operation, operators modify different hydrogen streams (fresh and recycled) in order to maintain the required supply to the reactors fulfilling the constraints imposed by compressors, purities and catalysts. Since there are different ways of providing the same amount of hydrogen that the reactor is consuming, the operational target is to supply the required flow to the reactors using the best combination of these sources from an economic point of view, satisfying the set of operational constraints.

The lack of reliable information about many streams and compositions, and the uncertainty of the demands, are important problems to implement a model-based controller for the optimal management of the hydrogen. Regarding the first one, it is clear that trustworthy information from the hydrogen network is required if one wishes to perform optimal decisions. On the other hand, about the hydrogen demands, there is a significant source of uncertainty from the changes in the composition of the hydrocarbon streams being treated in the HDS plants, which are linked to the origin of the crude oil or to production policies.

The provenance of the crude oil changes quite often, which modifies the hydrogen demands in the reactor. Even if the hydrocarbon flows to the HDS plant are planned, still uncertainty exists because its compositions, that determine the specific hydrogen consumption, are not well known. A typical pattern in the operation of an HDS plant is a transient lasting some hours followed by a stable demand. This situation becomes critical when a change in the hydrocarbon load takes place. This happens approximately every two days when new products are processed and the plant takes some hours to stabilize in the new operating point, which is desired to be optimum.

5.1. Model

The operation of an HDS unit involves many aspects. The model of the HDS is over-simplified and it does not intend to represent a real industrial plant, but to capture the main sources of uncertainty and the key decisions that should be made. The dynamics of the process are described by the set of Eqs. (12)–(17). The evolution of the hydrogen consumption inside the reactor (F^{H2}_X) is approximated by first order dynamics as Eq. (12). This consumption depends on the flow of hydrocarbon to be desulphurizated (F_{HC}) and the

stochastic variable (β), the specific consumption rate characteristic of the type of hydrocarbon received.

$$\tau \frac{dF_X^{H2}}{dt} + F_X^{H2} = F_{HC}\beta, \qquad F_X^{H2}(t_0) = F_{X0}^{H2}. \tag{12}$$

For the mixing point of the three hydrogen sources and the flow through the compressor, the total and hydrogen mole balances can be defined as:

$$F_1 + F_2 + F_3 = F_5, \tag{13}$$

$$F_1X_1 + F_2X_2 + F_3X_3 = F_5X_5, \tag{14}$$

where F_i, represents a flow according to Fig. 4 and X_1, X_2 and X_3 are unknown hydrogen compositions which are considered as another stochastic unknown variables.

Inside the reactor, the pressure is maintained at constant value by adjusting the flow of *LP*, so that, the dynamics of the total holdup can be neglected, unlike the hydrogen concentration that can change over time:

$$0 = F_5 - F_{10} - F_X^{H2}, \tag{15}$$

$$\frac{VP}{ZRT}\left(P\frac{dX_{H2}}{dt}\right) = F_5X_5 - F_{10}X_{H2} - F_X^{H2}, \tag{16}$$

$$X_{H2}(t_0) = X_{H2,0}. \tag{17}$$

The right-hand side of (16) involves a bilinear term which is the source of nonlinearity in the presented model. The main operational constraints refer to the allowable ranges for the flows and the minimum purity that must be maintained in the hydrogen stream (F_5) and the reactor (H_2). More details about the model can be found in Navia et al. (2014).

5.2. Control problem

The optimization problem is formulated in the context of economic dynamic optimization. It consists on finding the cheapest combination of hydrogen sources that produces the desulphurization of a given hydrocarbon when a change in the load is produced, fulfilling the constraints of the process.

The available control inputs are the flows F_1 and F_2 from high purity collectors (H_4) and (H_3) and the purge F_{10}. The optimization problem that is solved at each sampling time for the centralized approach can be written as:

$$\min_{\boldsymbol{x}_{k+1}^j, \boldsymbol{u}_k^j \forall (j,k)\in \boldsymbol{I}} \sum_{i=1}^{N} \omega_i \sum_{k=0}^{N_{p-1}} (C_{H4}X_{H4,k}^j F_{1,k}^j + C_{H3}X_{H3,k}^j F_{2,k}^j) \tag{18a}$$

subject to:

$$\boldsymbol{x}_{k+1}^j = \boldsymbol{f}(\boldsymbol{x}_k^{p(j)}, \boldsymbol{u}_k^j, \boldsymbol{d}_k^{r(j)}), \quad \forall (j, k+1) \in \boldsymbol{I}, \tag{18b}$$

$$F_l^{LO} \le F_{l,k}^j \le F_l^{UP}, \quad \forall \ (j,k) \in \boldsymbol{I}, \tag{18c}$$

$$X_5^{LO} \le X_{5,k}^j \le 1, \quad \forall \ (j,k) \in \boldsymbol{I}, \tag{18d}$$

$$X_{H2}^{LO} \le X_{H2,k}^j \le 1, \quad \forall \ (j,k) \in \boldsymbol{I}, \tag{18e}$$

$$\boldsymbol{u}_k^j = \boldsymbol{u}_k^l \quad if \quad \boldsymbol{x}_k^{p(j)} = \boldsymbol{x}_k^{p(l)}, \quad \forall \ (j,k),(l,k) \in \boldsymbol{I}, \tag{18f}$$

with $l = 1, 2, 10$. The objective function is given by Eq. (18a), where C_{H4} and C_{H3} are the costs of pure hydrogen from each fresh hydrogen source. Besides, all the constraints are applied to all the states $\boldsymbol{x}_{k+1}^j = (X_{H2,k}^j, X_{5,k}^j, F_{X,k}^{H2j})^T$ and control inputs $\boldsymbol{u}_k^j = (F_{1,k}^j, F_{2,k}^j, F_{10,k}^j)^T$ in the scenario tree. Eq. (18b) represents the discretized dynamics of the system, Eq. (18c) represents the constraints on the control inputs and Eqs. (18d)–(18e) represents the constraints on the reactor purities states.

If standard NMPC is used and the model is not perfect, i.e. there is plant-model mismatch due to the uncertainty parameters, the standard NMPC controller is not able to satisfy the constraints on the reactor concentration. The sampling time of the controller is $t_s = 0.5$h with a prediction horizon of $N_p = 5$ steps. It is considered that full state feedback is available at each sampling time. Fig. 5 shows the results of standard NMPC for different values of the uncertain parameters β, X_1, X_2 and X_3. Note that each line in the plot (this holds for all the plots throughout this paper, unless explicitly mentioned) represents the state and control trajectories corresponding to different values of the uncertain parameter for each control simulation (varying between ±30% with respect to their nominal values) using the same controller. The parameters are kept constant along each control simulation. It is clear that the standard NMPC controller fails to satisfy the constraints for several scenarios.

On the other hand, if multi-stage approach is implemented in a centralized manner, all the constraints are fulfilled. The different trajectories are shown in Fig. 6 also with mismatches of ±30% with respect to their nominal values. We consider all the scenarios to be equiprobable and the scenario tree to be built considering the combinations of the maximum and minimum values of the uncertain parameters using a robust horizon $N_r = 1$.

The two upper graphs display the values of the time evolution of the purities X_5 and X_{H2}, with the red dotted line showing the corresponding lower constraints, while the manipulated flows F_1, F_2 and F_{10} are shown in the lower graphs.

All the optimal control problems described in this work are discretized using orthogonal collocation on finite elements using two collocation points placed in the roots of Gauss–Radau polynomial. The resulting Nonlinear Programming problems are solved via IPOPT (Wächter and Biegler, 2006) which uses exact first- and second-order derivative information that is computed with the tool CasADi (Andersson et al., 2012). The implementation in this environment, reported in Lucia et al. (2014a) is found to be very efficient giving promising results. The computations are performed on a computer with 4 cores running at 3.67 GHz and 8 GB RAM. The computation times reported in this section contain all the necessary computations necessary to obtain the solutions, including the computation of sensitivity information, if required. The memory consumption reported in these sections is the memory consumed by the implemented program when it is executed, including all the use of the libraries that are necessary for its solution.

5.3. Decomposition approaches for multi-stage optimization

This section is focused on the solution of the multi-stage NMPC using the scenario decomposition approach with coordination algorithms described earlier and with the novel sensitivity-based technique, described in Section 3.3, which determines the values of aggregate variables.

In Fig. 7 a comparison is shown between multi-stage NMPC using the centralized and decomposition algorithms based on price-driven and augmented Lagrangean coordination ($\rho = 0.01$, $\rho_{max} = 1000$ and $\gamma = 2$) with and without the use of the sensitivity-based calculation of the values of the aggregated variables. The results are presented for a single realization of the uncertainty ($\beta = 14.553$, $X_3 = 0.803$, $X_1 = 0.971$ and $X_2 = 0.9114$). It can be seen that the solutions are similar, without violations of the constraints on the reactor concentration (18d) and (18e).

Tables 1 and 2 show the main results for each studied approach: number of iterations of the coordination algorithm to reach an optimal solution (Iter), computational memory of the optimization problem in megabytes (Mem), the number of decision variables of the multi-stage NMPC (Var) and the average time spent in each sampling instant to reach the optimal solution in seconds. The

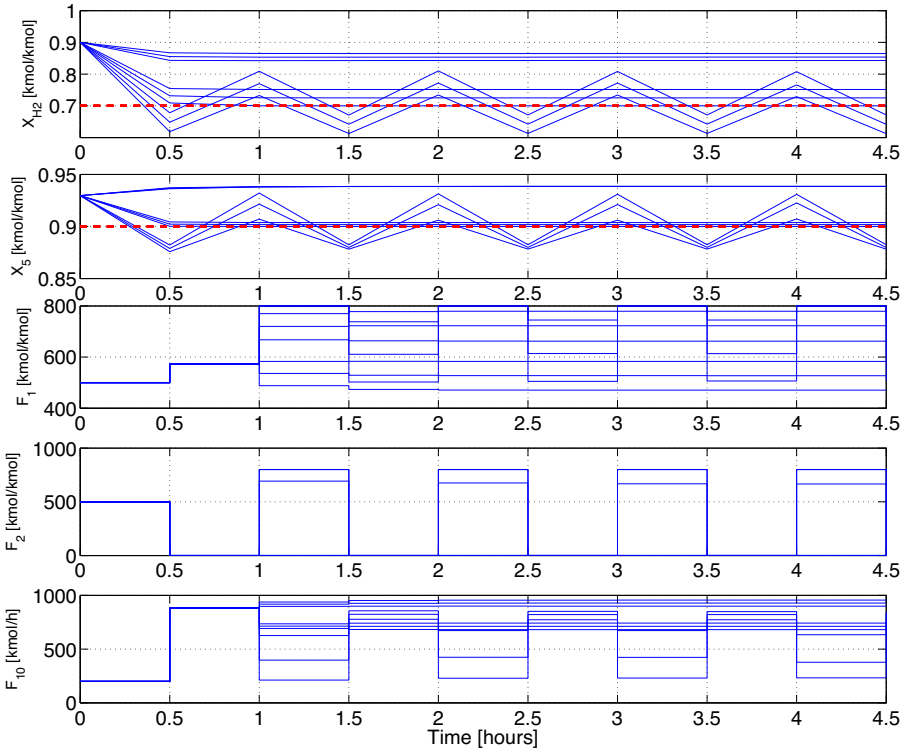


Fig. 5. Performance of standard NMPC for each considered scenario.

increasing number of scenarios corresponds to the consideration of more intermediate values of the allowed range of the uncertain parameters when building the scenario tree. Table 2 also shows results with $N_r \geq 1$. The three decomposition approaches fulfill all the constraints. In addition, if decomposition based on augmented Lagrangean is well tuned, it is able to reach the same results as price-driven coordination. The centralized approach spends less time to reach the solutions but the size of the optimization

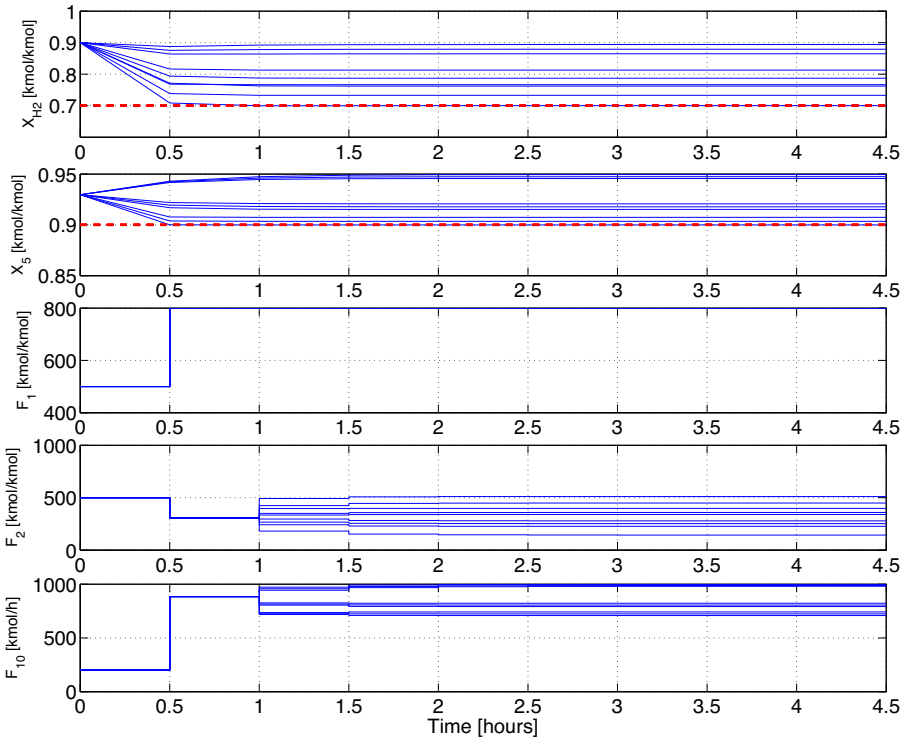


Fig. 6. Concentration X_{H2} and X_5 in hydrodesulphurization unit using monolithic multi-stage NMPC algorithm. The different trajectories show the different scenarios of the uncertainty with mismatches of ±30% with respect to their nominal values. (For interpretation of reference to color in this figure, the reader is referred to the web version of this article.)

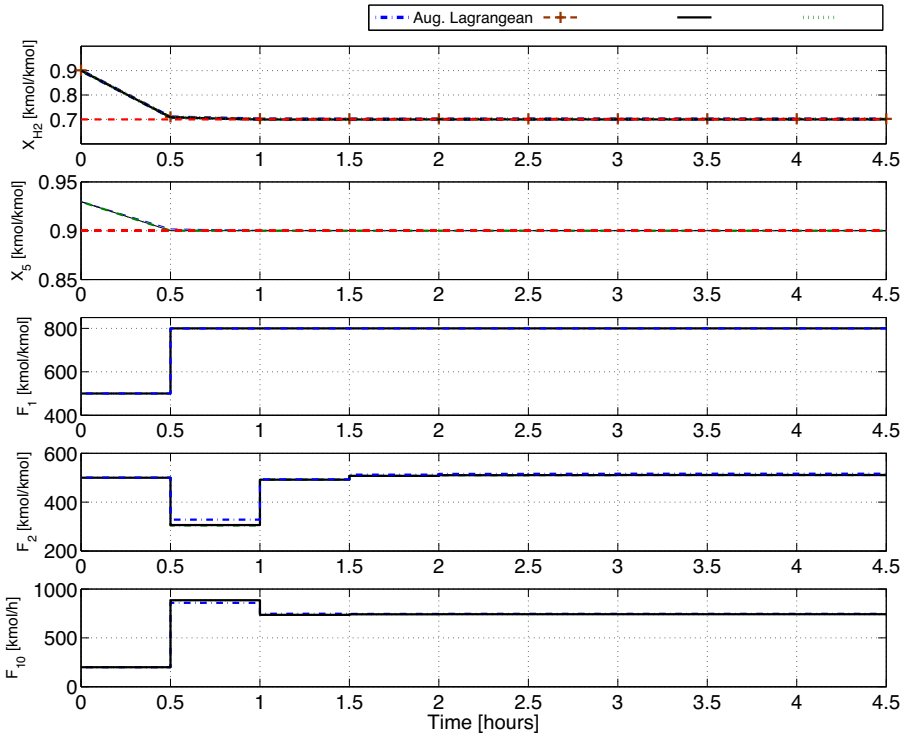


Fig. 7. Hydrodesulphurization concentration X_{H2} and X_5 trajectories for multi-stage NMPC problem using centralized approach, and decomposition with coordination based on augmented Lagrangean, price-driven coordination and sensitivity-based algorithm with augmented Lagrangean (S-AL) for a prediction horizon of $N_p = 5$ and a robust horizon $N_r = 1$ step.

problem is larger compared to decomposition approaches. It can be concluded that the decomposition approaches without any kind of modification are not competitive in comparison to centralized approach if enough RAM memory is available, but the memory requirements of large problems can limit the use of centralized approach and leave the decomposition as the only alternative. Note that the total number of scenarios solved is determined using $N = s^{N_r n_d}$.

As expected, the proposed sensitivity decomposition algorithm of both standard coordination methods results in a substantial decrease of the number of iterations of the coordination algorithm and, thus, CPU time. For cases with $N_r \geq 1$, i.e., when there are several non-anticipativity constraints, the centralized problem becomes more difficult and the proposed sensitivity decomposition algorithm achieves (slightly) faster computation times with an important reduction of the memory usage. In general we can observe that, for higher number of scenarios, the distributed approach becomes competitive. The price to pay for the improved performance of the sensitivity-based decomposition with respect to the augmented Lagrangean and price-based

Table 1
Number of iterations of the coordination algorithm, memory usage [MB], number of variables and average CPU time [s] to solve the problem for different number of scenarios and for each algorithm for case study 1 when $N_r = 1$ is applied: centralized multi-stage NMPC (CA), Augmented Lagrangean Decomposition (AL), Price-Driven Coordination (PDC), sensitivity-based decomposition algorithm with Price-Driven (S-PDC).

	CA			AL			PDC			S-PDC		
Sce.	81	624	2401	81	624	2401	81	624	2401	81	624	2401
N_r	1	1	1	1	1	1	1	1	1	1	1	1
Ite.	1	1	1	39	48	68	35	46	67	1	1	1
Mem.	60.3	302	1306	10.1	10.1	10.1	10.1	10.1	10.1	12.2	38.6	192.2
Var.	7455	57,503	220,895	92	92	92	92	92	92	92	92	92
CPU (s)	0.34	17.3	179.6	86.7	568.4	3944.6	83.3	562.1	3941.4	3.2	74.6	224.7

Table 2
Number of iterations of the coordination algorithm, memory usage [MB], number of variables and average CPU time [s] to solve the problem for different number of scenarios and for each algorithm for case study 1 when $N_r = 2$ and $N_r = 3$ are applied: Centralized multi-stage NMPC (CA), Augmented Lagrangean Decomposition (AL), Price-Driven Coordination (PDC), sensitivity-based decomposition algorithm with Price-Driven (S-PDC).

	CA			AL			PDC			S-PDC		
Sce.	64	64	256	64	64	256	64	64	256	64	64	256
N_r	2	3	2	2	3	2	2	3	2	2	3	2
Ite.	1	1	1	30	48	64	29	47	64	2.6	3.4	4.1
Mem.	34.1	51.1	173.9	10.1	10.1	10.2	10.1	10.1	10.2	10.5	11.1	11.74
Var.	4827	3839	18,995	92	92	92	92	92	92	92	92	92
CPU (s)	3.53	4.17	24.6	44.16	61.44	196.6	43.96	61.35	196.3	2.95	4.03	21.51

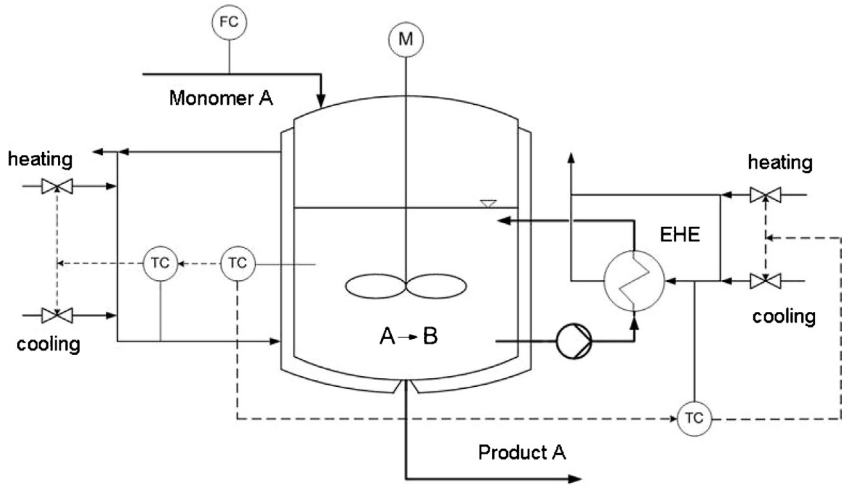


Fig. 8. Industrial batch polymerization reactor with an external heat exchanger.

decomposition is a increased memory consumption, compared to schemes without sensitivity-based decomposition, which is present due to the memory required for solving the problem (9). Note that the presented CPU times are reported for the application of the decomposition algorithms in a sequential way. The times can thus be further improved if some computations are parallelized.

Since the proposed sensitivity decomposition method already achieves convergence in few iterations of the coordination algorithm no further reduction in the number of iterations of the coordination algorithm is needed. Therefore, the performance of distributed multi-stage based on bundle decomposition will be shown in the following case study.

6. Case study 2: polymerization reactor

6.1. Model

We consider, as a second case study, the industrial polymerization reactor presented in Lucia et al. (2014b). For the sake of completeness we present here a short description of the process and the reader is referred to Lucia et al. (2014b) for a more detailed explanation of the model. A scheme of the system under consideration can be seen in Fig. 8. The system consists of a reactor into which monomer is fed. The monomer turns into a polymer via a very exothermic chemical reaction. The reactor is equipped with a jacket and with an external heat exchanger (EHE) that can both be used to control the temperature inside the reactor.

The process is modeled by a set of 8 ordinary differential equations:

$$\frac{dm_W}{dt} = \dot{m}_F\,\omega_{W,F}, \tag{19a}$$

$$\frac{dm_A}{dt} = \dot{m}_F\omega_{A,F} - k_{R1}\,m_{A,R} - k_{R2}\,m_{AWT}\,m_A/m_{ges}, \tag{19b}$$

$$\frac{dm_P}{dt} = \quad k_{R1}\,m_{A,R} + p_1\,k_{R2}\,m_{AWT}\,m_A/m_{ges}, \tag{19c}$$

$$\frac{dT_R}{dt} = 1/(c_{p,R}m_{ges})\ \ [\dot{m}_F\ \ c_{p,F}\,(T_F - T_R) + \Delta H_R k_{R1} m_{A,R} - k_K A\,(T_R - T_S) - \dot{m}_{AWT}\,c_{p,R}\,(T_R - T_{EK})], \tag{19d}$$

$$\frac{dT_S}{dt} = 1/(c_{p,S}m_S)\ \ [k_K A\,(T_R - T_S) - k_K A\,(T_S - T_M)], \tag{19e}$$

$$\frac{dT_M}{dt} = 1/(c_{p,W}m_{M,KW})\ \ [\dot{m}_{M,KW}\,c_{p,W}\left(T_M^{IN} - T_M\right) + k_K A\,(T_S - T_M)], \tag{19f}$$

$$\frac{dT_{EK}}{dt} = 1/(c_{p,R}m_{AWT})\ \ [\dot{m}_{AWT}c_{p,W}\,(T_R - T_{EK}) - \alpha\,(T_{EK} - T_{AWT}) + k_{R2}\,m_A\,m_{AWT}\Delta H_R/m_{ges}], \tag{19g}$$

$$\frac{dT_{AWT}}{dt} = [\dot{m}_{AWT,KW}\,c_{p,W}\,(T_{AWT}^{IN} - T_{AWT}) - \alpha\,(T_{AWT} - T_{EK})]/(c_{p,W}m_{AWT,KW}), \tag{19h}$$

where:

$$U = m_P/(m_A + m_P), \tag{19i}$$

$$m_{ges} = m_W + m_A + m_P, \tag{19j}$$

$$k_{R1} = \ k_0 e^{\frac{-E_a}{R(T_R+273.15)}}\,(k_{U1}\,(1-U) + k_{U2}U), \tag{19k}$$

$$k_{R2} = \ k_0 e^{\frac{-E_a}{R(T_{EK}+273.15)}}\,(k_{U1}\,(1-U) + k_{U2}U), \tag{19l}$$

$$k_K = (m_W\,k_{WS} + m_A\,k_{AS} + m_P\,k_{PS})/m_{ges}, \tag{19m}$$

$$m_{A,R} = m_A - m_A m_{AWT}/m_{ges}. \tag{19n}$$

The model includes mass balances for the water, monomer and product hold-ups (m_W, m_A, m_P) and energy balances for the reactor (T_R), the vessel (T_S), the jacket (T_M), the mixture in the external heat exchanger (T_{EK}) and the coolant leaving the external heat exchanger (T_{AWT}). The variable U denotes the polymer–monomer ratio in the reactor, m_{ges} represents the total mass, k_{R1} is the reaction rate inside the reactor and k_{R2} is the reaction rate in the external heat exchanger. The total heat transfer coefficient of the mixture inside the reactor is denoted as k_K and $m_{A,R}$ represents the current amount of monomer inside the reactor. The available control inputs are the feed flow $\dot{m}_F$, the coolant temperature at the inlet of the jacket T_M^{IN} and the coolant temperature at the inlet of the external heat exchanger T_{AWT}^{IN}. The complete set of parameters of the model can be seen in Lucia et al. (2014b).

6.2. Control problem

The control task for this case study is to produce one batch of polymer in the minimum time possible. This task is approximated by a maximization of the produced polymer in a finite horizon,

which produces very similar results as shown in simulations. A batch is considered finished when the desired amount of polymer has been produced ($m_P^{end} = 20,680$ kg).

The control task has to be achieved while satisfying quality and safety constraints. The temperature at which the polymerization reaction takes place strongly influences the quality of the resulting polymer. For this reason, the temperature of the reactor should be maintained in a range of ±2.0 °C around the desired reaction temperature $T_{set} = 90$ °C in order to ensure that the produced polymer has the required properties.

The safety constraint in this example is the maximum temperature that the reactor would reach in the case of a cooling failure, which is constrained to be below 109 °C and can be calculated as:

$$T_{adiab} = \frac{\Delta H_R}{c_{p,R}} \frac{m_A}{m_{ges}} + T_R. \tag{20}$$

The model is augmented with an additional differential equation that describes the accumulated monomer that has been fed to the reactor $\dot{m}_A^{acc} = \dot{m}_F$. The maximum amount of monomer that can be fed in the reactor in one batch is given by $\int \dot{m}_F \, dt = 30,000$ kg. Then, a constraint is included such that $0 < m_F^{acc} < m_F^{max} = 30,000$ kg.

The different quality and safety constraints must be fulfilled also in the presence of uncertainty. We consider that the reaction rate constant k_0 and the reaction enthalpy ΔH_R are uncertain and vary ± 30% around their nominal value. These parameters, which have a strong influence on the behavior of the system, are assumed to be uncertain but constant throughout the batch.

The optimization problem that has to be solved at each sampling time for the centralized multi-stage NMPC is:

$$\min_{\boldsymbol{x}_k^j, \boldsymbol{u}_k^j \forall (j,k) \in \boldsymbol{I}} \quad \sum_{i=1}^{N} \omega_i \sum_{k=0}^{N_p-1} - m_{P,k+1}^j + r_1 (\Delta \dot{m}_{F,k}^j)^2 + r_2 (\Delta T_{M,k}^{IN,j})^2 + r_3 (\Delta T_{AWT,k}^{IN,j})^2, \quad \forall \quad m_{P,k+1}^j, \dot{m}_{F,k}^j, T_{M,k}^{IN,j}, T_{AWT,k}^{IN,j} \in \boldsymbol{S}_i. \tag{21a}$$

subject to:

$$\boldsymbol{x}_{k+1}^j = f(\boldsymbol{x}_k^{p(j)}, \boldsymbol{u}_k^j, \boldsymbol{d}_k^{r(j)}), \quad \forall \quad (j, k+1) \in \boldsymbol{I}, \tag{21b}$$

$$T_{set} - 2.0 \le T_{R,k}^j \le T_{set} + 2.0, \quad \forall \quad (j,k) \in \boldsymbol{I}, \tag{21c}$$

$$0 \le T_{S,k}^j, T_{M,k}^j, T_{EK,k}^j, T_{AWT,k}^j \le 100, \quad \forall \quad (j,k) \in \boldsymbol{I}, \tag{21d}$$

$$0 \le T_{adiab,k}^j \le 109, \quad \forall \quad (j,k) \in \boldsymbol{I}, \tag{21e}$$

$$0 \le m_{A,k}^{acc,j} \le m_A^{max}, \quad \forall \quad (j,k) \in \boldsymbol{I}, \tag{21f}$$

$$0 \le \dot{m}_{F,k}^j \le 30,000, \quad \forall (j,k) \in \boldsymbol{I}, \tag{21g}$$

$$60 \le T_{M,k}^{IN,j} \le 100, \quad \forall \quad (j,k) \in \boldsymbol{I}, \tag{21h}$$

$$60 \le T_{AWT,k}^{IN,j} \le 100, \quad \forall \quad (j,k) \in \boldsymbol{I}, \tag{21i}$$

$$\boldsymbol{u}_k^j = \boldsymbol{u}_k^l \quad if \quad \boldsymbol{x}_k^{p(j)} = \boldsymbol{x}_k^{p(l)} \quad \forall \quad (j,k), (l,k) \in \boldsymbol{I}, \tag{21j}$$

where the constraints are applied to all the states and all control inputs along each scenario with

$$\boldsymbol{x}_k^j = [m_{W,k}^j, m_{A,k}^j, m_{P,k}^j, T_{R,k}^j, T_{S,k}^j, T_{M,k}^j, T_{EK,k}^j, T_{AWT,k}^j, T_{adiab,k}^j, m_{A,k}^{acc,j}]^T,$$

$$\boldsymbol{u}_k^j = [\dot{m}_{F,k}^j, T_{M,k}^{IN,j}, T_{AWT,k}^{IN,j}]^T.$$

As shown in (21a), the cost function includes for each scenario the maximization of the polymer mass and a penalty term (weighted by the tuning parameters $r_1 = 0.01$, $r_2 = 0.1$ and $r_3 = 0.1$) for the control movements to avoid unwanted oscillations on the control inputs.

To motivate the use of multi-stage NMPC, we show in Fig. 9 the results obtained by a standard NMPC controller for different realizations of the uncertain parameters. To solve the arising optimal control problems we use method of orthogonal collocation with three collocation points placed in the roots of Gauss-Radau polynomial. It is clear that standard NMPC cannot satisfy the constraints on the reactor temperature T_R or on the adiabatic temperature

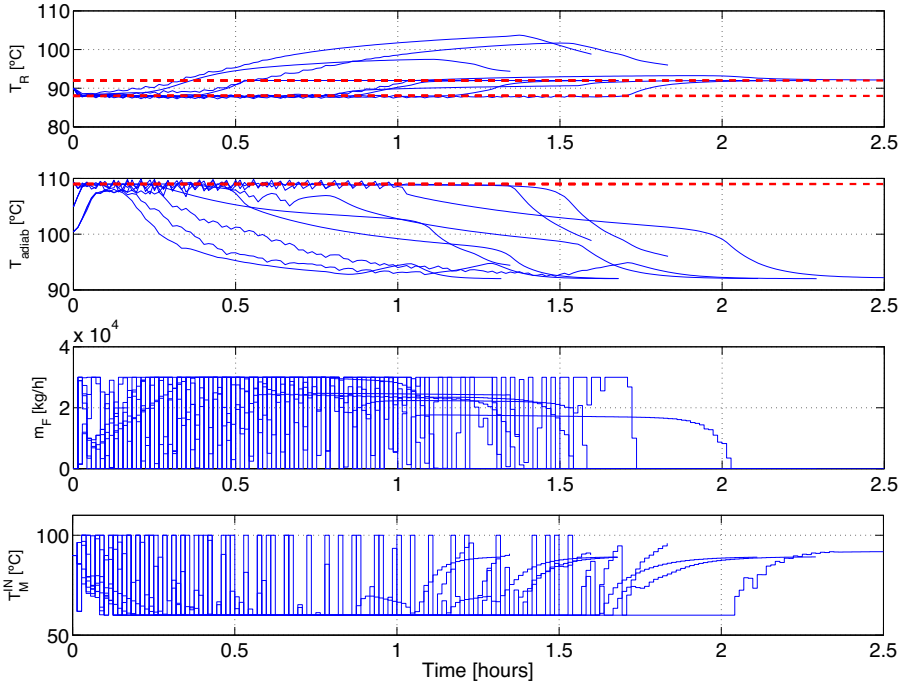


Fig. 9. Results of one deterministic NMPC for each scenario considered.

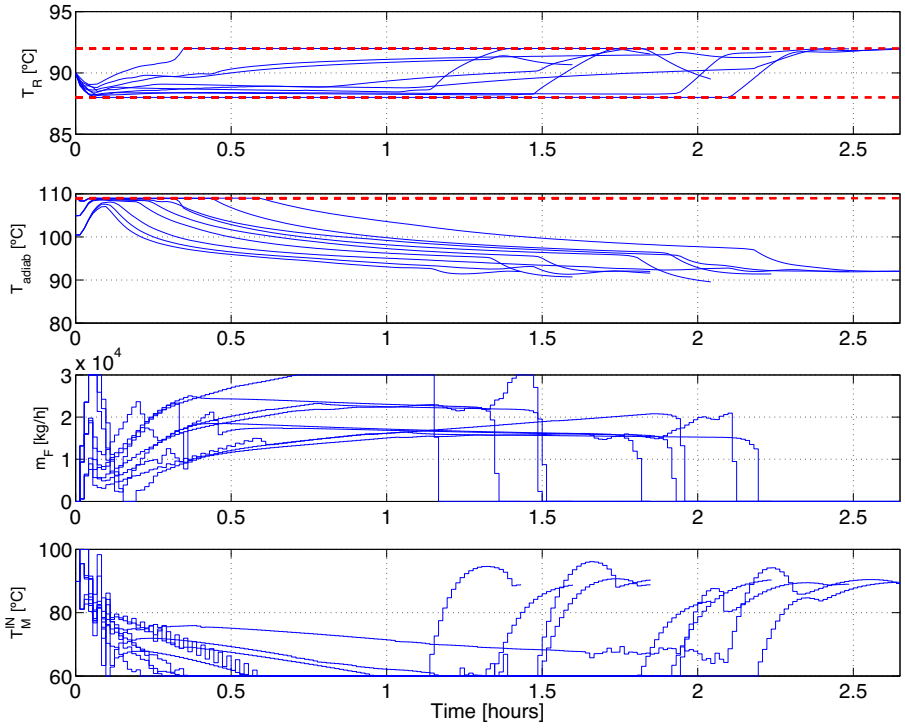


Fig. 10. Industrial batch polymerization reactor temperature, safety temperature (with constant constraints indicated), monomer feed jacket temperature trajectories using monolithic multi-stage NMPC algorithm. The different trajectories show the different scenarios of the uncertainty with mismatches of ±30% with respect to their nominal values.

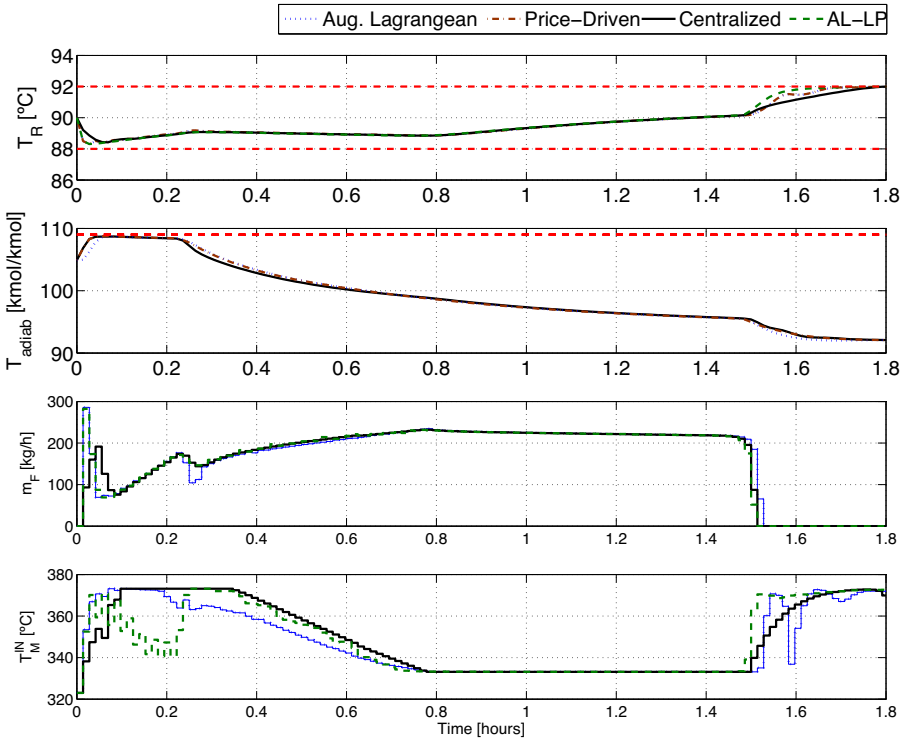


Fig. 11. Industrial batch polymerization reactor temperature, safety temperature (with constant constraints indicated), monomer feed jacket temperature trajectories using Centralized, Augmented Lagrangean, Price-Driven Coordination and proposed algorithm with Augmented Lagrangean approaches for a prediction horizon of $N_p = 20$ steps and a robust horizon $N_r = 1$ step.

T_{adiab}. This results in unfeasible optimization problems that cause the oscillations on the control inputs. Even if soft constraints are implemented to avoid the infeasible optimization problems, the constraints cannot be satisfied due to the wrong predictions of the standard NMPC controller.

6.3. Decomposition approaches for multi-stage optimization

This section compares the centralized solution to different decomposition approaches, which have been presented above. If multi-stage NMPC is used, in which a scenario tree is generated

Table 3
Polymerization reactor case study, number of iterations, memory (Mb), number of variables and average CPU time (s) to solve the problem for different number of scenarios and for each algorithm when robust horizon (N_r) equal to 1 is applied: Centralized multi-stage NMPC (CA), Augmented Lagrangean decomposition (AL), price-driven Coordination (PDC), sensitivity-based algorithm with Augmented Lagrangean (S-AL).

	CA			AL			PDC			S-AL		
Sce.	9	25	81	9	25	81	9	25	81	9	25	81
N_r	1	1	1	1	1	1	1	1	1	1	1	1
Ite.	1	1	1	10.6	16.3	35.3	10.6	16.6	35.2	3.7	4.8	7.9
Mem.	159.3	331.4	930.4	35.8	35.8	35.8	35.8	36.1	35.8	35.9	36.2	36.4
Var.	7754	24,428	69,420	680	680	680	680	680	680	680	680	680
CPU(s)	1.13	3.3	12.2	8.7	37.5	263.1	8.7	37.1	263.8	3.1	11.1	46.1
$CPU_{Par}(s)$	-	-	-	4.4	16.5	106.3	4.4	16.5	106.2	1.54	5.24	21.25

Table 4
Polymerization reactor case study, number of iterations, memory (Mb), number of variables and average CPU time (s) to solve the problem for different number of scenarios and for each algorithm when robust horizon (N_r) greater to 1 is applied: Centralized multi-stage NMPC (CA), Augmented Lagrangean decomposition (AL), price-driven coordination (PDC), sensitivity-based algorithm with Augmented Lagrangean (S-AL).

	CA			AL			PDC			S-AL		
Sce.	16	64	81	16	64	81	16	64	81	16	64	81
N_r	2	3	2	2	3	2	2	3	2	2	3	2
Ite.	1	1	1	12.4	24.5	46.3	12.4	24.3	46.3	4.6	5.4	8.6
Mem.	225.9	731.1	809.9	35.8	35.8	35.8	35.8	36.1	35.8	35.9	36.2	36.4
Var.	13,199	50,207	66,324	680	680	680	680	680	680	680	680	680
CPU(s)	3.46	8.83	16.35	14.09	142.7	303.8	14.08	142.5	303.7	5.22	24.5	50.15
$CPU_{Par}(s)$	-	-	-	7.1	59.7	134.2	7.1	59.6	134.2	3.14	11.32	23.16

using the combinations of the maximum, minimum and nominal values of the uncertain parameters, the results obtained for different scenarios of the uncertain parameters can be seen in Fig. 10. These results show the evolution of manipulated variables and states of the plant if the realization of true parameters are as ones assumed by building a scenario tree. We use for all the results in this case study a sampling time of $t_s = 20$ s, a prediction horizon $N_p = 20$ and a robust horizon $N_r = 1$. It can be seen that the constraints are satisfied for all the values of the uncertainty. The different lengths of the batches correspond to different realizations of uncertainty represented by the reaction rate constant k_0, that determines speed of the reaction, and the reaction enthalpy ΔH_R, that implicitly determines the rate of monomer feeding. A comparison of multi-stage NMPC with other robust NMPC techniques can be seen in Lucia et al. (2014b).

The decomposition approaches presented in this paper achieve a very similar control performance in comparison to the centralized computation. This can be illustrated in Fig. 11 for a single realization of the uncertainty ($k_0 = 7.0$ and $\Delta H_R = 950.0$). We can observe that all decomposition approaches reach practically the same performance with some small differences that occur due to local optima of the problem. The similarity of the approaches is best evidenced by the achieved batch time which is almost the same in all cases.

The results of the different algorithms for distributed multi-stage NMPC are summarized in Table 3 and in Table 4 for cases with $N_r \geq 1$. The tuning parameters for the augmented Lagrangean approached were selected as $\rho = 100$, $\rho_{max} = 10{,}000$ and $\gamma = 2$. In this case study, since several constraints are active and the problem is highly nonlinear, the number of iterations of the coordination algorithm needed by the decomposition algorithms is large. This results in much slower solutions when compared to the centralized solution. The proposed sensitivity-based method to compute the aggregate variables significantly reduces the number of iterations of the coordination algorithm needed, and hence the necessary computation time. However, an efficient computation of the centralized solution, using as in this case orthogonal collocation on finite elements and exact first- and second-order derivatives, is faster than any of the proposed decomposition methods. Tables 3 and 4 show another advantage of the decomposition approaches: the possibility of parallelization. If the computations are parallelized (using 4 cores), the computation times obtained by the proposed sensitivity-based approach are very similar to the ones obtained with the centralized approach (especially when $N_r \geq 1$), with a significant reduction of the memory usage. Further improvements can be achieved if a higher degree of parallelization is performed.

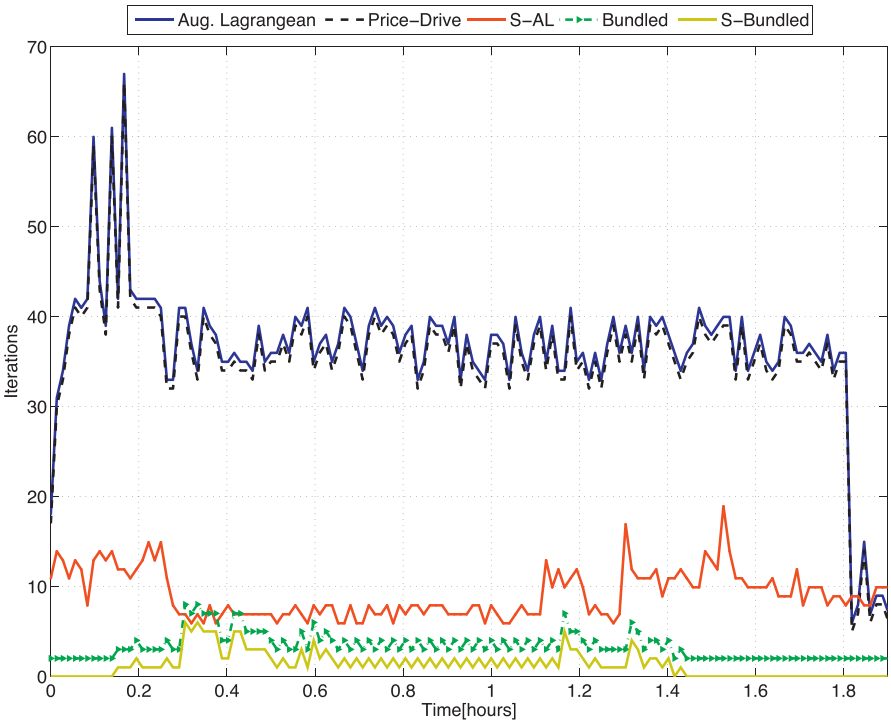
6.4. Further reduction of the number of iterations of the coordination algorithms

In order to further reduce the number of iterations needed for the decomposition algorithms to converge, we implement in this case study the bundle decomposition method.

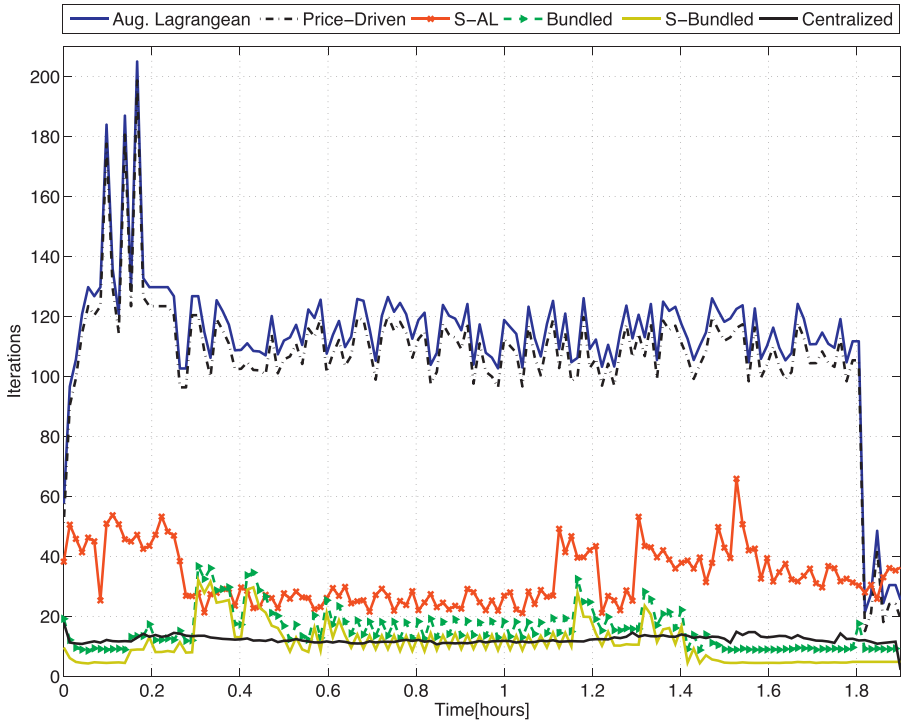
We generate bundles in a random fashion. For the case of 9 scenarios, there are 3 bundles chosen with 3 scenarios each one, which are chosen randomly. On the other hand, the monolithic optimization problem with 25 scenarios is solved using 2 bundles with 8 scenarios and one more with 9 scenarios. Finally, the last problem which corresponds to 81 scenarios is solved using 3 bundles with 20 scenarios and one more with 21 scenarios.

As it can be seen in Table 5, the number of iterations of the coordination algorithm needed is significantly lower than in the fully decomposition-based approach for this case study, and it is comparable to the sensitivity-based method proposed in this work. It is then possible to combine both methods, in which each subproblem is a bundle of scenarios and the aggregated control inputs are calculated using the proposed sensitivity-based method. As shown in Table 5, this further reduces the number of iterations of the coordination algorithm needed, achieving (slightly) shorter computation times and lower memory requirements compared to the centralized approach, for the problems with a high number of scenarios.

Fig. 12(a) shows a comparison of the number of iterations of the coordination algorithm needed for each algorithm to converge to a solution that satisfies the non-anticipativity constraints at each sampling time. The combination of heuristics (use of bundles) and a systematic use of the sensitivities proposed in this paper achieves a significant and consistent reduction in the number of iterations



(a) Number of iterations of the coordination algorithm.



(b) Computational time at each NMPC iteration.

Fig. 12. (a) Number of iterations of the coordination algorithm and (b) computation time for Augmented Lagrangean, Price-Driven Coordination (Price-Driven), sensitivity-based decomposition with Augmented Lagrangean (S-AL), bundled approach (Bundled) and sensitivity-based decomposition with bundled approach (S-Bundled). The results show the simulations with $N = 81$ and $N_r = 1$.

Table 5
Polymerization reactor case study when the robust horizon (N_r) is equal to 1, number of iterations, memory (Mb), number of variables and average CPU time (s) to solve the problem for different number of scenarios and for each algorithm: centralized multi-stage NMPC (CA), Bundled algorithm (Bundled) and sensitivity-based algorithm with Bundled approach (S-Bundled).

	CA			Bundled			S-Bundled		
Sce.	9	25	81	9	25	81	9	25	81
N_r	1	1	1	1	1	1	1	1	1
Ite.	1	1	1	3.2	3.8	3.4	2.56	2.84	2.41
Mem.	159.3	331.4	930.4	55.4	124.1	237.3	58.6	125.5	238.5
Var.	7754	24,428	69,420	2580	7740	17,351	2580	7740	17,351
CPU (s)	1.13	3.3	12.2	2.49	7.66	28.9	1.97	6.24	20.54
CPU_{Par} (s)	-	-	-	1.43	4.28	15.8	1.36	2.86	11.208

of the coordination algorithm. The decrease in the number of iterations of the coordination algorithm is done at the price of a higher memory consumption, compared to fully distributed approach, due to the size of the bundles. It can be seen in Fig. 12(b) as illustrated in Table 5 that the bundle sensitivity-based decomposition has a very similar performance compared to the centralized approach (slightly faster in average), with lower memory requirements.

7. Conclusions and future work

This paper presents different possibilities to solve the large-scale nonlinear programming problems that result from the multi-stage NMPC formulation in an efficient manner, achieving the same (or slightly faster) computation times, with an important reduction in terms of memory consumption. In particular we focus on scenario decomposition approaches to relax the centralized problem and solve smaller subproblems which are iterated using different coordination algorithms until a solution that satisfies the global constraints is achieved. We compare existing methods such as the augmented Lagrangean decomposition and a price-based coordination scheme with a novel algorithm that uses sensitivity information to reduce the number of iterations of the coordination algorithm needed to converge. We also show results that use a hybrid method between a fully centralized and a fully distributed approach, which can be used to further reduce the number of iterations of the coordination algorithm. The results obtained for industrial case studies from the chemical industry domain show that a price-based coordination algorithm is equivalent to a well-tuned Augmented Lagrangean decomposition. It has been also shown that introducing problem-specific information for the computation of the aggregate variables as we proposed significantly reduces the number of iterations of the coordination algorithm needed by the decomposition algorithms to converge. In addition, the simulation results suggest that for slightly nonlinear examples where the decomposition methods need few iterations to converge the proposed sensitivity-based decomposition method can achieve faster computation times with an important decrease in the memory usage. For highly nonlinear examples with different constraints (as the second case study), an efficient implementation of the centralized problem results in faster solution than the decomposition methods at the cost of a higher memory consumption. In those cases, the use of parallelization or some heuristics (such as the use of bundles) can help to obtain similar computation times with reduced memory requirements.

Acknowledgements

RM acknowledges the financial support of the project DPI2012-37859, MICINN Spain. SL, RP and SE acknowledge the support of European Commission under grant agreement number 291458 (MOBOCON) and 611281 (DYMASOS) and the support from the Deutsche Forschungsgemeinschaft under grant agreement number EN 152/39-1.

References

Andersson J, Äkesson J, Diehl M. Casadi: a symbolic package for automatic differentiation and optimal control. In: Forth S, Hovland P, Phipps E, Utke J, Walther A, editors. Recent advances in algorithmic differentiation, Vol. 87 of lecture notes in computational science and engineering. Springer; 2012]. p. 297–307.

Bellos G, Papayannakos N. The use of a three phase microreactor to investigate hds kinetics. Catal Today 2003];79-80:349–55.

Bernardini D, Bemporad A. Scenario-based model predictive control of stochastic constrained linear systems. In: Proc. of the 48th IEEE conference on decision and control, 2009; 2009]. p. 6333–8.

Birge J. Stochastic programming computation and applications. INFORMS J Comp 1997];9(2):111–33.

Campo PJ, Morari M. Robust model predictive control. In: Proc. of the American control conference; 1987]. p. 1021–6.

Engell S. Feedback control for optimal process operation. J Process Control 2007];17(3):203–19.

Gopalakrishnan A, Biegler L. Economic nonlinear model predictive control for periodic optimal operation of gas pipeline networks. Comp Chem Eng 2013];52:90–9.

Idris E, Engell S. Economics-based NMPC strategies for the operation and control of a continuous catalytic distillation process. J Process Control 2012];22:1832–43.

Jose R, Ungar L. Auction-driven coordination for plantwide optimization. In: Foundations of computer-aided process operation (FOCAPO); 1998].

Jose RA, Ungar LH. Pricing interprocess streams using slack auctions. AIChE J 2000];46(3):575–87.

Lee JH, Yu ZH. Worst-case formulations of model predictive control for systems with bounded parameters. Automatica 1997];33(5):763–81.

Lucia S, Finkler T, Basak D, Engell S. A new robust NMPC scheme and its application to a semi-batch reactor example. In: Proc. of the international symposium on advanced control of chemical processes; 2012]. p. 69–74.

Lucia S, Finkler T, Engell S. Multi-stage nonlinear model predictive control applied to a semi-batch polymerization reactor under uncertainty. J Process Control 2013a];23:1306–19.

Lucia S, Subramanian S, Engell S. Non-conservative robust nonlinear model predictive control via scenario decomposition. In: Proc. of the 2013 IEEE multi-conference on systems and control; 2013b]. p. 586–91.

Lucia S, Andersson J, Brandt H, Bouaswaig A, Diehl M, Engell S. Efficient robust economic nonlinear model predictive control of an industrial batch reactor. In: Prof. of the 19th IFAC world congress; 2014a]. p. 11093–8.

Lucia S, Andersson J, Brandt H, Diehl M, Engell S. Handling uncertainty in economic nonlinear model predictive control: a comparative case-study. J Process Control 2014b];24:1247–59.

Lucia S, Paulen R, Engell S. Multi-stage nonlinear model predictive control with verified robust constraint satisfaction. In: Proc. of the 54th IEEE conference on decision and control; 2014c]. p. 2816–21.

Marti R, Sarabia D, Navia D, De Prada C. Coordination of distributed model predictive controllers using price driven coordination and sensitivity analysis. In: Proc. of the 10th IFAC international symposium on dynamics and control of process systems; 2013a]. p. 215–20.

Marti R, Sarabia D, Navia D, De Prada C. A method to coordinate decentralized NMPC controllers in oxygen distribution networks. Comp Chem Eng 2013b]:122–37.

Marti R, Navia D, Sarabia D, de Prada C. Distributed stochastic optimization of a process plant start-up. In: Proc. of 19th world congress the international federation of automatic control; 2014a]. p. 2734–9.

Marti R, Sarabia D, de Prada C. Price-driven coordination for distributed NMPC using feedback control law. In: Maestre JM, Negenborn RR, editors. Distributed MPC made easy. Springer; 2014b] [Chapter 3].

Marti R, Lucia S, Sarabia D, Paulen R, Engell S, de Prada C. An efficient distributed algorithm for multi-stage robust nonlinear predictive control. In: Proc. of the European control conference; 2015] [in Press].

Mayne D, Seron M, Rakovic S. Robust model predictive control of constrained linear systems with bounded disturbances. Automatica 2005];41:219–24.

Mayne D. Control of constrained dynamic systems. Eur J Control 2001];7:87–99.
Muñoz de la Peña D, Bemporad A, Alamo T. Stochastic programming applied to model predictive control. In: Proc. of the 44th IEEE conference on decision and control; 2005]. p. 1361–6.
Navia D, Sarabia D, Gutiérrez G, Cubillos F, Prada Cd. A comparison between two methods of stochastic optimization for a dynamic hydrogen consuming plant. Comp Chem Eng 2014]:219–33.
Patrinos ST, Panagiotis A, Bemporad. Stochastic mpc for real-time market-based optimal power dispatch. In: Decision and control and European control conference (CDC-ECC); 2011]. p. 7111–6.
Pirnay H, Lopez-Negrete R, Biegler L. Optimal sensitivity based on IPOPT. Math Program Computat 2012];4:307–31.
Prada C, Sarabia D, Cristea S, Mazaeada R. Plant-wide control of a hybrid process. Int Jf Adapt Control Signal Process 2008];22(2):124–41.
Rakovic S, Kouvaritakis B, Cannon M, Panos C, Findeisen R. Fully parameterized tube MPC. In: Proc. of the 18th IFAC World Congress Milano; 2011]. p. 197–202.
Rawlings J, Amrit R. Optimizing process economic performance using model predictive control. In: Nonlinear model predictive control. Berlin, Heidelberg: Springer; 2009]. p. 119–38.
Rockafellar RT, Wets RJ-B. Scenarios and policy aggregation in optimization under uncertainty. Math Oper Res 1991];16(1):119–47.
Ruszczynski A. Decomposition methods in stochastic programming. Math Program 1997];79:333–53.
Scokaert P, Mayne D. Min–max feedback model predictive control for constrained linear systems. EEE Trans Autom Control 1998]:1136–42.
Shapiro A. Lectures on stochastic programming: modeling and theory. In: SIAM; 2009].
Steinbach M. Hierarchical sparsity in multistage convex stochastic programs. In: Stochastic optimization: algorithms and applications. Kluwer Academic Publishers; 2000]. p. 363–88.
Wächter A, Biegler L. On the implementation of a primal-dual interior point filter line search algorithm for large-scale nonlinear programming. Math Program 2006];106:25–57.

Journal of Process Control 72 (2018) 39–51

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

Dual robust nonlinear model predictive control: A multi-stage approach

S. Thangavel [a,*], S. Lucia [a,b], R. Paulen [a,c], S. Engell [a]

[a] Process Dynamics and Operations Group, Department of Chemical and Biochemical Engineering, Technische Universität Dortmund, Emil-Figge-Strasse 70, 44227 Dortmund, Germany
[b] Fakultät Elektrotechnik und Informatik, Technische Universität Berlin, Ernst-Reuter-Platz 7, 10587 Berlin, Germany
[c] Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Radlinského 9, 812 37 Bratislava, Slovakia

ARTICLE INFO

Article history:
Received 2 February 2018
Received in revised form 8 October 2018
Accepted 11 October 2018

Keywords:
Dual control
Multi-stage decision making
Nonlinear model predictive control
Adaptive control
Economically optimal operation
Robust control

ABSTRACT

Dual control seeks to explicitly deal with the trade-off between the excitation of the controlled system by probing actions, which lead to a more accurate estimation of the unknown parameters of the plant model, and performance (set-point tracking, economic optimality, etc.) of the controlled system under the imperfect knowledge of the plant behavior. This paper presents a dual-control approach that extends a nonlinear model predictive controller, the control actions of which are robust against the effect of model uncertainties. The robustness is achieved via the multi-stage approach that uses a scenario-tree representation of the propagation of the uncertainties over the prediction horizon of the controller and includes the adaptation of the control actions on the basis of the information that is gained in the future in the optimization problem. The dual-control aspect of the proposed scheme is realized via the direct consideration of the reduction of the range of the parameter uncertainty that is predicted as a result of the parameter estimation using the future measurements. This implicit dual-control mechanism does not require a-priori tuning with respect to the relative importance of the probing actions against the optimal operation of the system, as proposed in other recent approaches. The results from a reactor control example show the advantage of using Dual Multi-stage NMPC over its robust adaptive counterpart, where the reduction of the uncertainty is not predicted and optimized, but only obtained a posteriori when the measurements have arrived.

1. Introduction

Model-based optimal control of dynamic systems is a well-established paradigm where open-loop predictions, on the basis of some form of a mathematical model, and feedback, enabled by the measurement of the response of the plant, are combined to minimize or maximize a performance criterion of the system under control with time-varying degrees of freedom. Among the different model-based controllers, model predictive control originally proposed in [1,2] has gained an important position in applications due to its ability to handle multi-variable systems and hard constraints [3–5]. The performance of model-predictive controllers critically depends on the accuracy of the model used.

In general, it is difficult to obtain a precise model of a plant and there often exists an uncertainty that can be represented by a time-varying disturbance or by uncertain parameters. In the presence of a significant level of uncertainty and when tight constraints are imposed on the system states and outputs, the control actions calculated by MPC based on a nominal model are suboptimal and can even be infeasible. A common way to handle this situation is to design a cautious (robust) controller, which satisfies the constraints for all possible realizations of the uncertainty [4,6]. This leads to the well-known min–max formulation [7], tube-based methods [8,9], or Multi-stage NMPC [10].

Min–max MPC [7] extends the idea behind min–max optimal control of sampled linear systems as presented in [11] to the MPC framework. Min–max MPC computes the optimal inputs for the worst-case realization of the uncertainty. However, it does not take into account the presence of feedback information, which is available in the future and is therefore conservative and may even result in an infeasible optimization problem [12]. To reduce the conservativeness of Open-loop Min-max MPC by taking the future feedback information into account Closed-loop Min-max MPC [13,14] has

* Corresponding author.
E-mail address: sakthi.thangavel@tu-dortmund.de (S. Thangavel).

https://doi.org/10.1016/j.jprocont.2018.10.003
0959-1524/

been introduced, but it requires the optimization over feedback policies or the assumption of a fixed controller structure.

Tube-based MPC [8,15] is a robust MPC strategy, which is based on set-theoretic methods. It uses a nominal controller with tightened constraints and an ancillary controller, which keeps the system state in the neighborhood of the nominal trajectory in the presence of the uncertainty. Several extensions of Tube-based MPC were proposed in the literature with different levels of conservativeness and computational complexity [9,16–19]. The main differences among these approaches lie in the computation of the uncertainty region around the nominal trajectory and in the ancillary controller chosen.

Multi-stage NMPC [10] is another recently developed promising technique for robust NMPC-based control of dynamic systems under uncertainty. It models the realization of the uncertainty by a tree of discrete scenarios and includes recourse actions that take the realization of the uncertainty into account in the open-loop optimization. This results in the optimal closed-loop policy when the uncertainty is correctly represented by the scenario tree.

The application of any robust control mechanism inevitably results in conservatism and a loss of optimality compared to the case when perfect information (a perfect model) of the system is available. In order to reduce this conservatism and to improve the closed-loop performance of the controller, measurements from the controlled system can be used to enhance the knowledge about the system via parameter estimation, e.g., in the least-squares sense, which reduces the range of uncertainty. This is usually referred to as adaptive robust control [20–23].

In order to achieve reliable and precise parameter estimation, the information content of the data that is gathered from the controlled system should be rich. This necessitates the application of a sufficiently exciting control input, also called a probing action, which disturbs the system states and outputs in order to reveal important system characteristics. However, such control inputs generally are not in line with the goal of optimal operation, as they can significantly decrease the closed-loop performance. This poses a challenging decision problem with two conflicting goals; to apply probing inputs that maximize the information about the uncertain system and to apply inputs that optimize the desired performance. This control problem was introduced under the name *dual* control in 1960 by Feldbaum [24].

The solution to the dual-control problem is in most cases impossible to compute analytically and it might even be computationally intractable [25,26]. These challenges can be overcome if approximations are introduced. One can distinguish between explicit and implicit dual-control schemes [27]. Explicit dual control tries to resolve the trade-off by minimizing the weighted sum of a performance-related objective and the objective of reducing the range of uncertainty, where the latter is formulated as minimization in the sense of a design of experiments [28,29]. The main drawback of this approach is the necessity to decide a-priori upon the relative importance of the two objectives.

Implicit dual-control approaches consider the adaptation of the parameters along the prediction horizon that results when the optimized control input is applied and compute the expected gain in performance due to the improved parameter estimates in the optimization of the inputs. Ideally, the accuracy of the prediction due to the future measurements should be incorporated. Some implicit dual-control approaches reduce the complexity of the problem by truncating the adaptation to a horizon that is shorter than the prediction horizon [30–33] or by simplifying the adaptation laws, e.g., by linearization around a nominal trajectory of the system [34,35]. For a comprehensive review of dual-control methods the reader is referred to [27,21].

In this work, we propose an approach to implicit dual control using Multi-stage NMPC. This, on one hand, enables the explicit consideration of (parametric) model uncertainty to achieve the optimal robust performance of the closed-loop system. On the other hand, it is straightforward to incorporate the adaptation within the prediction horizon as the control actions along the prediction horizon depends on the range of the uncertainty that is influenced by the previous inputs. The latter feature makes the computation of dual-control actions possible.

This study is an extension of our previous work [31] where it is assumed that the uncertain parameters that are obtained from the parameter estimation problem converge to the true plant parameters. The novel approach that is presented in this paper eliminates this assumption by incorporating a region of uncertainty of the future parameter estimates, which depends upon the amount of information gained using the past and future measurements. We provide a comparison of the proposed method with the approach of [31] along with a comparison to nondual robust controllers, which are also based on the multi-stage formulation. This is done by means of extensive simulation studies for an example of economics-oriented control of a batch chemical reactor.

The remainder of the paper is structured as follows. Section 2 states the problem at hand with the necessary preliminaries. Section 3 presents the dual-control approach that builds on Multi-stage NMPC and Adaptive Multi-stage NMPC. Section 4 introduces the case study. The simulation results obtained with the different robust NMPC approaches are presented and discussed in Section 5. The paper is concluded by Section 6.

2. Problem statement

We consider the problem of finding an optimizing control input of a plant (dynamic system) that can be stated in the framework of receding-horizon model predictive control (MPC) as follows:

$$\min_{\boldsymbol{x}_k, \boldsymbol{u}_k, \forall k \in \{t,\ldots,N_p-1\}} \sum_{k=t}^{t+N_p-1} \ell_k(\boldsymbol{x}_{k+1}, \boldsymbol{u}_k), \tag{1a}$$

$$\text{s.t. } \boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k, \boldsymbol{u}_k, \boldsymbol{d}), \quad \forall k \in \{t, \ldots, t+N_p-1\}, \tag{1b}$$

$$\boldsymbol{g}(\boldsymbol{x}_{k+1}, \boldsymbol{u}_k) \leq \boldsymbol{0}, \quad \forall k \in \{t, \ldots, t+N_p-1\}, \tag{1c}$$

where $\ell_k(\cdot)$ expresses the control performance at time k over the prediction horizon of the length N_p. The set of equality constraints (1b) represents the model equations of the system, where $\boldsymbol{f}(\cdot)$ is a sufficiently many times continuously differentiable nonlinear mapping, and $\boldsymbol{g}(\cdot)$ is a continuously differentiable nonlinear mapping that describes the constraint on the state and input variables, $\boldsymbol{x} \in \mathbb{R}^{n_x}$ and $\boldsymbol{u} \in \mathbb{R}^{n_u}$.

The variable $\boldsymbol{d}$ represents an n_d-dimensional vector of uncertainties of the model. Here we assume that the uncertainties are parametric, time-invariant and bounded such that $\boldsymbol{d} \in \boldsymbol{D}_0 := \boldsymbol{d}_0^{nom} + [\underline{\Delta}\boldsymbol{d}_0, \bar{\Delta}\boldsymbol{d}_0]$ where $\boldsymbol{d}_0^{nom}$ represents the nominal values of the uncertain parameters. $\Delta\boldsymbol{d}_0 := [\underline{\Delta}\boldsymbol{d}_0, \bar{\Delta}\boldsymbol{d}_0]$ is the associated range of uncertainty, where $\underline{\Delta}\boldsymbol{d}_0$ and $\bar{\Delta}\boldsymbol{d}_0$ denote the component wise lower and upper bounds on the range of the uncertainty. The subscript 0 indicates the a-priori knowledge about the uncertain parameters at time 0. If parameter adaptation is considered based on the available measurements, the uncertainty interval can evolve through time such that at time t the uncertainty range is given by $\boldsymbol{D}_t := \boldsymbol{d}_t^{nom} + [\underline{\Delta}\boldsymbol{d}_t, \bar{\Delta}\boldsymbol{d}_t] \subseteq \boldsymbol{D}_0$. We assume full-state measurement, so the initial conditions of the state variables are given as $\boldsymbol{x}_t := \boldsymbol{x}(t)$, where $\boldsymbol{x}(t)$ is the measurement at time instant t.

In the absence of full state information, observers such as the nominal model Extended Kalman filter [36] or the moving-horizon estimator [37] can be used to estimate the unmeasured states. Here, we assume for simplicity that full state information from the plant is available.

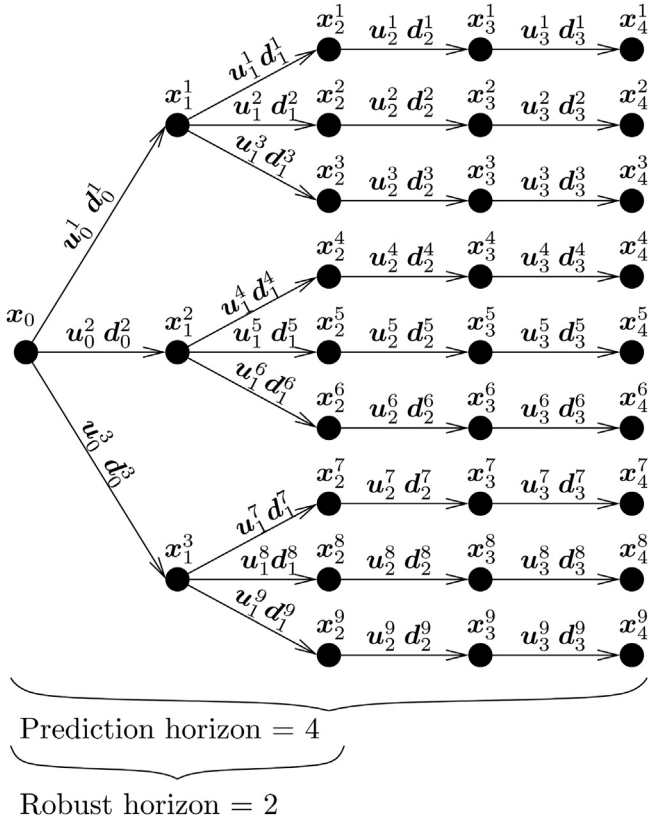


Fig. 1. Scenario-tree representation of the uncertainty evolution for Multi-stage NMPC.

Problem (1) can be solved at each sampling time using the nominal value $\boldsymbol{d}:=\boldsymbol{d}_0^{nom}$, to determine the optimal input $\boldsymbol{u}_t^*$. Given the uncertainty present in $\boldsymbol{d}$, if nominal MPC is used, the determined control input is neither guaranteed to be optimizing nor necessarily feasible, which motivates the further developments presented in this study.

3. Formulation of the Dual Multi-stage controller

The formulation of the proposed implicit dual multi-stage controller involves the combination of a robust controller, an adaption mechanism for the model parameters and the incorporation of the adaptation into the future stages of the uncertainty model. The following sections provide details on each of these elements.

3.1. Multi-stage NMPC

Multi-stage NMPC [10] is a robust NMPC strategy that is based on the description of the evolution of the uncertainty by a tree of discrete scenarios as represented in Fig. 1. Each branch of the tree represents a trajectory of the system states under a certain realization of the uncertainty dependent on the control sequence. The main advantage of such a formulation is that the availability of information that is provided by the future measurements can be taken into account so that the future control inputs can be optimized taking into account the future reaction to the realization of the uncertainty (the control inputs beyond the next time step become scenario-dependent recourse variables). In this way, a feedback problem can be solved as an open-loop optimization problem. This significantly improves the performance of the controller in comparison to open-loop min–max MPC as reported in [12,38], and makes the incorporation of the predicted reduction of the uncertainty of the parameter estimates possible, as required for implicit dual control.[1]

The system dynamics along the branches of the scenario tree that are based on different realizations of the uncertainty, can be formulated as:

$$\boldsymbol{x}_{k+1}^{j} = \boldsymbol{f}(\boldsymbol{x}_k^{p(j)}, \boldsymbol{u}_k^{j}, \boldsymbol{d}_k^{r(j)}), \tag{2}$$

where $\boldsymbol{x}_{k+1}^{j}$, the state vector at stage $k+1$ and position in the tree (node) j, is obtained as a function of the parent state node $\boldsymbol{x}_k^{p(j)}$, the control input $\boldsymbol{u}_k^{j}$ and the realization of the uncertainty $\boldsymbol{d}_k^{r(j)} \in \boldsymbol{D}_t$.

In Fig. 1, we illustrate the situation at $t=0$. Here the node $\boldsymbol{x}_1^1$ in the scenario tree is given by the parent state $\boldsymbol{x}_0$, control input $\boldsymbol{u}_0^1$ under the uncertainty realization $\boldsymbol{d}_0^1$. The nodes $\boldsymbol{x}_0$, $\boldsymbol{x}_1^1$, $\boldsymbol{x}_2^1$, $\boldsymbol{x}_3^1$, and $\boldsymbol{x}_4^1$ form a scenario S_1 of the scenario tree under the realization of the uncertainty $\boldsymbol{d}_0^1$, $\boldsymbol{d}_1^1$, $\boldsymbol{d}_2^1$, and $\boldsymbol{d}_3^1$. We assume that the tree has the same number of branches at each node, given by $\boldsymbol{d}_k^{r(j)} \in \{\boldsymbol{d}_k^1, \boldsymbol{d}_k^2, \ldots, \boldsymbol{d}_k^s\}$, where s is the number of branches at each node.

The consideration of the combinations of the minimal, maximal and nominal values of the uncertain parameters usually results in a robust behavior at a manageable computational burden as shown in [10]. This results in 3^{n_d} branches that are considered at each node in the scenario tree. In this notation, the nominal scenario ($\boldsymbol{d}_k^{nom}$) considers $\boldsymbol{d}_k^{nom}:=\boldsymbol{d}_t^{nom}$, the scenario obtained using minimal values of the uncertain parameters ($\underline{\boldsymbol{d}}_k^{nom}$) considers $\underline{\boldsymbol{d}}_k^{nom}:=\boldsymbol{d}_t^{nom}+\underline{\Delta}\boldsymbol{d}_t$ and the scenario obtained using maximal values of the uncertain parameters ($\bar{\boldsymbol{d}}_k^{nom}$) considers $\bar{\boldsymbol{d}}_k^{nom}:=\boldsymbol{d}_t^{nom}+\underline{\Delta}\boldsymbol{d}_t$ for $t \le k \le t+N_p-1$. For simplicity, we denote the operation of building the scenario tree, i.e., assignment of the different extreme discrete realizations of parametric uncertainty from the set $\boldsymbol{D}_t$, as $\boldsymbol{d}_k^{r(j)} \in \mathbb{D}(\boldsymbol{d}_t^{nom}, \underline{\Delta}\boldsymbol{d}_t, \bar{\Delta}\boldsymbol{d}_t)$, $\forall (j,k) \in \mathcal{I}$. The standard (non-adaptive) Multi-stage NMPC considers $\boldsymbol{D}_t := \boldsymbol{D}_0, \forall\, t \ge 0$, hence $\boldsymbol{d}_k^{r(j)} \in \mathbb{D}(\boldsymbol{d}_0^{nom}, \underline{\Delta}\boldsymbol{d}_0, \bar{\Delta}\boldsymbol{d}_0)$, $\forall (j,k) \in \mathcal{I}$.

In order to avoid the exponential growth of the tree, we assume that the uncertainty remains constant after a certain point in time, which is called the robust horizon N_r. To simplify the notation, the set of indices (j,k) that occur in a given scenario tree is denoted by $\mathcal{I}$.

The optimization problem that is solved at each sampling time using the multi-stage formulation of robust NMPC reads as:

$$\min_{\boldsymbol{x}_k^j, \boldsymbol{u}_k^j, \forall (j,k) \in \mathcal{I}} \sum_{i=1}^{N} \omega_i \sum_{k=t}^{t+N_p-1} \ell_k(\boldsymbol{x}_{k+1}^j, \boldsymbol{u}_k^j), \qquad \forall\ \boldsymbol{x}_{k+1}^j, \boldsymbol{u}_k^j \in S_i \tag{3a}$$

subject to:

$$\boldsymbol{x}_{k+1}^{j} = \boldsymbol{f}(\boldsymbol{x}_k^{p(j)}, \boldsymbol{u}_k^{j}, \boldsymbol{d}_k^{r(j)}), \qquad \forall (j, k+1) \in \mathcal{I}, \tag{3b}$$

$$\boldsymbol{g}(\boldsymbol{x}_{k+1}^{j}, \boldsymbol{u}_k^{j}) \le \boldsymbol{0}\ , \qquad \forall\ (j, k+1) \in \mathcal{I}, \tag{3c}$$

$$\boldsymbol{u}_k^j = \boldsymbol{u}_k^l \ \text{ if } \boldsymbol{x}_k^{p(j)} = \boldsymbol{x}_k^{p(l)}, \qquad \forall\ (j,k),(l,k) \in \mathcal{I}, \tag{3d}$$

$$\boldsymbol{d}_k^{r(j)} \in \mathbb{D}(\boldsymbol{d}_0^{nom}, \underline{\Delta}\boldsymbol{d}_0, \bar{\Delta}\boldsymbol{d}_0), \qquad \forall\ (j,k) \in \mathcal{I} \tag{3e}$$

where N is the number of scenarios. The constraints $\boldsymbol{g}(\boldsymbol{x}_{k+1}^{j}, \boldsymbol{u}_k^{j}) \le 0$ are applied to each node in the tree. The coefficient ω_i is the

[1] The presented formulation is rigorously valid only for the case of a discrete-valued uncertainty. In the case of a general nonlinear system and continuous parameter uncertainty, this formulation does not guarantee robust constraint satisfaction for those values of the uncertainty that are not explicitly included in the scenario tree. However, very often a scenario tree that is generated using the combinations of the assumed maximum, minimum and nominal values of the uncertain model parameters provides very good results [10,39]. If a rigorous guarantee is required, the multi-stage approach can be combined with reachability analysis as shown in [40].

non-negative weighting coefficient of the ith scenario (S_i) and is chosen based on the relative importance of this scenario over the other ones with $\sum_{i=1}^{N}\omega_i = 1$. In the absence of information about the relative importance of each scenario, ω_i is chosen as $\frac{1}{N}$, i.e., all the scenarios are given equal importance. Constraints (3d) are the non-anticipativity (causality) constraints [10]. They enforce the causality of the controller by making sure that the control decisions taken at a subsequent node are the same if the preceding parent node is mutual to them as they are based on the same information (e.g., in Fig. 1 $u_0^1 = u_0^2 = u_0^3$; $u_1^1 = u_1^2 = u_1^3$; $\cdots$).

Remark 1. The stability properties of Multi-stage MPC have been previously analyzed in the literature, using standard tracking objective functions. The first analysis was presented in [12] for linear systems and using a min–max cost function. Results have been presented for linear systems based upon stochastic considerations in which exponential stability in the mean square sense has been established for linear systems [41,42]. The setting considered in this paper is more complex, as it considers nonlinear systems and a deterministic representation of the uncertainty. First results for the stability analysis of this formulation can be found in [43]. The consideration of different weights for the branches in the tree that are added in the cost function (as opposed to a worst-case or probabilistic formulation) is a challenge for stability analysis. This is achieved in [43] via continuity assumptions on the model equations. As with other robust approaches, only convergence to a neighborhood of the equilibrium point can be achieved, because the different branches of the uncertainty are considered at each sampling time, even if the disturbance vanishes. Convergence can be recovered by using a dual-mode approach as used in other MPC schemes [44].

3.2. Adaptive Multi-stage NMPC

The principle of adaptive control is to use the available information about the system for the improvement of the performance of the control system. The measurement information can be used to reduce the range of the uncertainty of the uncertain parameters $\boldsymbol{D}_t$ at time instant t. This can be resolved in the framework of least-squares estimation (LSE) where the (nominal) estimate of the uncertain parameters using the measurements that are available until a certain time instant t, i.e., $\boldsymbol{x}(t)$, $t \geq 0$, is found via

$$\boldsymbol{d}_t^{nom} := \underset{\boldsymbol{d}}{\operatorname{arg\,min}} \sum_{k=0}^{tn_s} (\boldsymbol{x}(k) - \boldsymbol{x}_k)^T \mathbf{Q} (\boldsymbol{x}(k) - \boldsymbol{x}_k) \tag{4a}$$

$$\text{s.t. } \boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k, \boldsymbol{u}_k, \boldsymbol{d}), \quad \forall k \in \{0, \ldots, tn_s - 1\}, \tag{4b}$$

for a given sequence $\boldsymbol{u}_k$, where $\boldsymbol{Q}$ denotes the inverse of the variance–covariance matrix of the measurement noise. We consider here that the sampling time of the measurements can differ from sampling time of the controller (4b), where n_s denotes the number of measurements between consecutive sampling times of the controller.

The bounds $\underline{\Delta}\boldsymbol{d}_t$ and $\bar{\Delta}\boldsymbol{d}_t$ represent a joint-confidence interval of parameter estimates, which depends on the information content of the data and can be obtained as follows. If we assume that the model is structurally identifiable, and that a white Gaussian noise affects the data, according to the Cramer–Rao inequality [45], the parametric variance–covariance matrix can be over-approximated as

$$\left(\mathbf{P}_{0,t}^{\boldsymbol{d}}\right)^{-1} \approx \sum_{k=0}^{tn_s} \boldsymbol{s}_k^T(\boldsymbol{d}) \mathbf{Q} \boldsymbol{s}_k(\boldsymbol{d}). \tag{5}$$

$\left(\mathbf{P}_{0,t}^{\boldsymbol{d}}\right)^{-1}$ is called the Fisher information matrix. $\mathbf{P}_{0,t}^{\boldsymbol{d}}$ gives an upper bound on the parameter covariance matrix, and $\boldsymbol{s}_k \in \mathbb{R}^{n_y \times n_d}$ represents the matrix of the parametric output sensitivities. n_y is the number of observed outputs ($n_y = n_x$ in case of full state measurement).

It is then possible to construct a joint-confidence ellipsoid centered at the value of the least-squares parameter estimate that bounds all the possible values of the least-squares parameter estimates (under different realizations of the measurement noise) with a prescribed confidence level α. This in turn implies a certain level of uncertainty in the parameters. The joint-confidence ellipsoid can be determined using [45]:

$$(\boldsymbol{d} - \boldsymbol{d}_t^{nom})^T \left(\mathbf{P}_{0,t}^{\boldsymbol{d}_t^{nom}}\right)^{-1} (\boldsymbol{d} - \boldsymbol{d}_t^{nom}) \leq n_d \mathcal{F}_{n_d, tn_s n_y - n_d, \alpha}, \tag{6}$$

where $\mathcal{F}$ represents the upper α quantile of the Fisher distribution with n_d and $tn_s n_y - n_d$ degrees of freedom in the numerator and denominator. The ellipsoid computed with tn_s measurement points provides the bounds $\boldsymbol{D}_t$ on the uncertain parameters as the projections of the ellipsoid onto the parametric axes which can be obtained as:

$$\underline{\Delta}\boldsymbol{d}_t := \max\left\{\underline{\Delta}\boldsymbol{d}_{t-1}, -\text{diag}^{\frac{1}{2}}\left(n_d \mathcal{F}_{n_d, tn_s n_y - n_d, \alpha} \mathbf{P}_{0,t}^{\boldsymbol{d}_t^{nom}}\right)\right\}, \tag{7a}$$

$$\bar{\Delta}\boldsymbol{d}_t := \min\left\{\bar{\Delta}\boldsymbol{d}_{t-1}, \text{diag}^{\frac{1}{2}}\left(n_d \mathcal{F}_{n_d, tn_s n_y - n_d, \alpha} \mathbf{P}_{0,t}^{\boldsymbol{d}_t^{nom}}\right)\right\}, \tag{7b}$$

where the operator diag$(\cdot)$ gives the vector of diagonal elements of a matrix. Because of the assumption of constant uncertain parameters, the min and max operators are used to clip the bounds should the a-priori knowledge of the parametric bounds (at time $t-1$) be violated by the estimation at time t.

The problem (3) can then be solved at the time instance t with the updated scenario tree with the uncertainty interval $\boldsymbol{D}_t$. The nominal scenario ($\boldsymbol{d}_k^{nom}$) of the evolution of the process considers $\boldsymbol{d}_k^{nom} := \boldsymbol{d}_t^{nom}$, the scenario obtained using minimal values of the uncertain parameters ($\underline{\boldsymbol{d}}_k^{nom}$) considers $\underline{\boldsymbol{d}}_k^{nom} := \boldsymbol{d}_t^{nom} + \underline{\Delta}\boldsymbol{d}_t$ and the scenario obtained using the maximal value of uncertain parameters ($\bar{\boldsymbol{d}}_k^{nom}$) considers $\bar{\boldsymbol{d}}_k^{nom} := \boldsymbol{d}_t^{nom} + \bar{\Delta}\boldsymbol{d}_t$ for $t \leq k \leq t + N_p - 1$. Equivalently, in the notation introduced above, the scenarios of the Adaptive Multi-stage NMPC consider $\boldsymbol{d}_k^{r(j)} \in \mathbb{D}(\boldsymbol{d}_t^{nom}, \underline{\Delta}\boldsymbol{d}_t, \bar{\Delta}\boldsymbol{d}_t)$, $\forall (j,k) \in \mathcal{I}$, which replaces (3e).

The presented Adaptive Multi-stage NMPC can significantly improve the performance of Multi-stage NMPC, as shown in [46], as it narrows the width of the scenario tree based on the available information. It is, however, not a dual-control approach as it does not take into account that the probing actions can improve the accuracy of the parameter estimates, which can then result in a further improvement of the performance of Multi-stage NMPC.

Remark 2. The computational complexity of the optimization problem (4) increases as $t \to \infty$. This can be overcome by solving the parameter estimation problem (4) in a moving horizon fashion, where instead of using all the available measurements from the plant only the measurements within a window of a chosen size are considered.

Remark 3. If there exists a strong correlation between the estimated uncertain parameters, the resulting joint-confidence ellipsoid will be considerably skewed. The consideration of the minimal and the maximal values of the uncertain parameters for the creation of the scenario tree might then introduce a large degree of overestimation and thus the benefits of adaptive and dual approaches might be insignificant. This can be avoided by building the scenario tree using the so-called sigma points. A study in this direction was presented recently [47].

Remark 4. In the presence of biased parameter estimates, the confidence region of the uncertain parameters obtained from (6) may not enclose the true values of the uncertain parameters. As a result, the scenario tree of Adaptive Multi-stage NMPC built using the minimal and maximal values of the uncertain parameters may not enclose the true values of the uncertain parameters, and this may lead to constraint violations. In such cases, one can replace (6) by a set of guaranteed parameter estimates using set-membership estimation [48,49]. The minimum and maximum values of the uncertain parameters can be obtained by solving an optimization problem as shown in [50] and the scenario tree of the Multi-stage NMPC can then be updated along the prediction horizon [51]. However, this is computationally expensive.

3.3. Dual Multi-stage NMPC

As the aim of the dual control is to strike a balance between the (optimizing) control actions and the probing actions, a dual predictive controller has to possess the ability of predicting the impact of the probing actions and the subsequent reduction of the range of uncertainty (width of the scenario tree) on the robust optimality of the controlled system. We propose Dual Multi-stage NMPC that uses the predicted parametric bounds that result from the future parameter estimation, subject to the constraints that arise from the different realizations of the uncertainty. The scenario-tree formulation is suitable for this purpose since the uncertainty can be treated as being time-varying in a straightforward manner and the effect of the choice of the inputs can be included directly in the optimization.

We present two formulations of a dual controller. One optimistically assumes that the least-squares estimates converge rapidly to the true values of the uncertain parameters and remain constant over the prediction horizon. The second approach uses the amount of information gained from past and future data in order to predict the bounds on the future parameter estimates.

The challenge for the computation of the future bounds on the uncertain parameters is the lack of measurements, and the implied uncertainty about the future parameter estimates. We addressed this issue in [31] by extending the certainty-equivalence principle, assuming that the least-squares estimate which is used as the nominal value of the uncertainty when building the scenario tree, remains constant over the prediction horizon. The range of the uncertainty along the scenario tree can then be found by (7). It depends on the control inputs that were applied previously as these influence the parametric sensitivities in (5). We refer to this strategy as *Dual Multi-stage NMPC 1*.

The overall formulation of the Dual Multi-stage NMPC 1 scheme is given by

$$\min_{\boldsymbol{x}^j_k,\boldsymbol{u}^j_k,\boldsymbol{d}^{r(j)}_k,\forall(j,k)\in\mathcal{I}} \sum_{i=1}^{N}\omega_i \sum_{k=t}^{t+N_p-1} \ell_k(\boldsymbol{x}^j_{k+1},\boldsymbol{u}^j_k), \qquad \forall\ \boldsymbol{x}^j_{k+1},\boldsymbol{u}^j_k \in S_i, \tag{8a}$$

subject to:

$$\boldsymbol{x}^j_{k+1} = \boldsymbol{f}(\boldsymbol{x}^{p(j)}_k,\boldsymbol{u}^j_k,\boldsymbol{d}^{r(j)}_k), \quad \forall (j,k+1)\in\mathcal{I}, \tag{8b}$$

$$\boldsymbol{s}^j_{k+1} = \frac{\mathrm{d}\boldsymbol{f}(\boldsymbol{x}^{p(j)}_k,\boldsymbol{u}^j_k,\boldsymbol{d}^{r(j)}_k)}{\mathrm{d}\boldsymbol{d}^{r(j)}_k}, \quad \forall (j,k+1)\in\mathcal{I}, \tag{8c}$$

$$\boldsymbol{g}(\boldsymbol{x}^j_{k+1},\boldsymbol{u}^j_k)\le\boldsymbol{0}\ , \quad \forall\ (j,k+1)\in\mathcal{I}, \tag{8d}$$

$$\boldsymbol{u}^j_k = \boldsymbol{u}^l_k \quad \text{if}\ \boldsymbol{x}^{p(j)}_k = \boldsymbol{x}^{p(l)}_k, \quad \forall\ (j,k),(l,k)\in\mathcal{I}, \tag{8e}$$

$$\left(\mathbf{P}^{\boldsymbol{d}^{nom}_t}_{0,t}\right)^{-1} = \sum_{\nu=0}^{n_s t}\boldsymbol{s}^T_\nu(\boldsymbol{d}^{nom}_t)\mathbf{Q}\boldsymbol{s}_\nu(\boldsymbol{d}^{nom}_t), \tag{8f}$$

$$\left(\mathbf{P}^{\boldsymbol{d}^{r(j)}_{k-1}}_{t,t+k}\right)^{-1} = \sum_{\nu=n_s t+1}^{n_s t+n_s k}\boldsymbol{s}^T_\nu(\boldsymbol{d}^{r(j)}_{k-1})\mathbf{Q}\boldsymbol{s}_\nu(\boldsymbol{d}^{r(j)}_{k-1}), \quad \forall (j,k)\in\mathcal{I}, \tag{8g}$$

$$\left(\mathbf{P}^{\boldsymbol{d}^j_k}_{0,t+k}\right)^{-1} = \left(\mathbf{P}^{\boldsymbol{d}^{nom}_{N_m}}_{0,t}\right)^{-1} + \left(\mathbf{P}^{\boldsymbol{d}^{r(j)}_{k-1}}_{t,t+k}\right)^{-1}, \quad \forall (j,k)\in\mathcal{I}, \tag{8h}$$

$$\underline{\Delta}\boldsymbol{d}^j_{t+k} = \max\left\{\underline{\Delta}\boldsymbol{d}^j_{t+k-1}, -\mathrm{diag}^{\frac{1}{2}}\left(n_d\mathcal{F}_{n_d,(t+k)n_s n_y-n_d,\alpha}\mathbf{P}^{\boldsymbol{d}^j_t}_{0,t+k}\right)\right\}, \quad \forall (j,k)\in\mathcal{I}, \tag{8i}$$

$$\bar{\Delta}\boldsymbol{d}^j_{t+k} = \min\left\{\bar{\Delta}\boldsymbol{d}^j_{t+k-1}, \mathrm{diag}^{\frac{1}{2}}\left(n_d\mathcal{F}_{n_d,(t+k)n_s n_y-n_d,\alpha}\mathbf{P}^{\boldsymbol{d}^j_t}_{0,t+k}\right)\right\}, \quad \forall (j,k)\in\mathcal{I}, \tag{8j}$$

$$\boldsymbol{d}^{r(j)}_k \in \mathbb{D}(\boldsymbol{d}^{nom}_t, \underline{\Delta}\boldsymbol{d}^j_{t+k}, \bar{\Delta}\boldsymbol{d}^j_{t+k}), \quad \forall (j,k)\in\mathcal{I}. \tag{8k}$$

The parametric sensitivities are obtained using (8c). They can be computed using numerical techniques with high accuracy [52,53]. The Fisher information matrices w.r.t., the past measurements and the future predictions are obtained from (8f) and from (8g), respectively. The upper bound on the parameter covariance matrix of the uncertain parameters can be obtained using the Fisher information matrix of the past measurements and the future predictions using (8h) [45,54]. The minimum and maximum values of the uncertain parameters and the realizations of the uncertainty in the scenario tree are updated along the prediction horizon (8k) in contrast to Multi-stage and Adaptive Multi-stage NMPC.

The Dual Multi-stage NMPC 1 strategy is optimistic as it assumes that the least-squares estimator (4a) has converged to the true values of the uncertain parameters. If this assumption is not satisfied, the robustness of the control scheme cannot be guaranteed. In order to overcome this potential problem, we propose a second approach where preference can be given to either robustness or optimality of the control actions by adjusting a tuning constant. This strategy is called *Dual Multi-stage NMPC 2*.

The central idea here is the use of an additional uncertainty in the future parameter estimates that are obtained from possible future measurements and estimates until some point in time for the prediction of the future deviation of the parameter estimates. In Dual Multi-stage NMPC 2, an over-approximation factor $\delta\boldsymbol{d}^j_k \in \mathbb{R}^{n_d}$ is added to $\Delta\boldsymbol{d}_{N_m+kn_s}$. The over-approximation factor $(\delta\boldsymbol{d}^j_k)$ is computed via

$$\delta^a\boldsymbol{d}^j_k := \sqrt{n_d\mathcal{F}_{n_d,(t+k)n_s n_y-n_d,\beta}}\left(\mathrm{diag}^{\frac{1}{2}}\left(\mathbf{P}^{\boldsymbol{d}^{nom}_t}_{0,t}\right) - \mathrm{diag}^{\frac{1}{2}}\left(\mathbf{P}^{\boldsymbol{d}^{r(j)}_k}_{t,t+k}\right)\right), \quad \forall (j,k)\in\mathcal{I}, \tag{9a}$$

$$\delta d^j_{k,s} := \begin{cases} \delta^a d^j_{k,s}, & \text{if}\ \delta^a d^j_{k,s} > 0, \\ 0, & \text{otherwise}, \end{cases} \quad \forall s\in\{1,\ldots,n_d\}, \forall (j,k)\in\mathcal{I}, \tag{9b}$$

where $\delta^a d^j_{k,s}$ represents an a-priori estimate of the over-approximation factor for the sth parameter. The over-approximation factor depends on the information carried by the past and the future measurements. If the future measurements carry more information about the uncertain parameters than the past measurements, it is highly likely that the future parameter estimates change depending on the future measurements, hence an over-approximation factor is added to the parameter estimates. On the contrary, if the past measurements carry more information about the uncertain parameters than the future measurements, then the over-approximation factor $(\delta d^j_{k,s})$ is set to 0. One might want to avoid too pessimistic predictions here, therefore we suggest to use a confidence level (β) that can be different from the confidence level used to obtain the bounds on the uncertain parameters (α), which is used as a tuning parameter. The over-approximation factor $(\delta d^j_{k,s})$ increases with the increase in the confidence level (β) chosen.

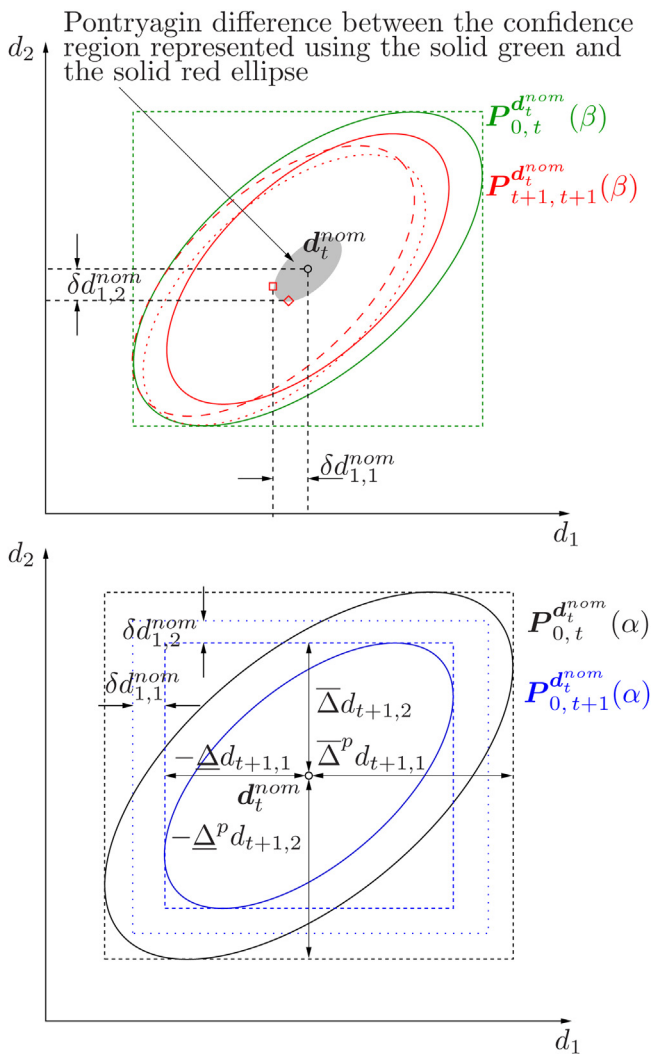


Fig. 2. Illustration of the projection of information gathered from the past measurements on the future evolution of the parametric bounds for Dual Multi-stage NMPC 2. Top figure – procedure to obtain the over-approximation factor $\delta \boldsymbol{d}$. Bottom figure – past and predicted confidence regions of the uncertain parameters.

We illustrate this idea in the top plot of Fig. 2 for the root node and child node on the nominal branch of the scenario tree. The same procedure is applied to the rest of the nodes of the scenario tree. We first determine the information gain of the measurements on the horizon $[0, t]$ represented by the green ellipsoid and the information gain of the measurements on the horizon $(t+1, t+k]$ represented by the red ellipsoid. By performing a Pontryagin difference between the confidence regions defined by $\boldsymbol{P}_{0,t}^{\boldsymbol{d}_t^{nom}}(\beta) = n_d \mathcal{F}_{n_d,(t+k)n_s n_y - n_d,\beta} \mathbf{P}_{0,t}^{\boldsymbol{d}_t^{nom}}$ and $\boldsymbol{P}_{t,t+k}^{\boldsymbol{d}_t^{nom}}(\beta) = n_d \mathcal{F}_{n_d,(t+k)n_s n_y - n_d,\beta} \mathbf{P}_{t,t+k}^{\boldsymbol{d}_t^{nom}}$ (this is indicated by the shift of the red ellipsoid to the dashed and dotted ellipsoids in the bottom plot of Fig. 2), we obtain a region (in gray) that indicates where least-squares estimate might be shifted when the measurements from the interval $(t+1, t+k]$ are gathered. Projections of this region on the parametric axes give $\delta \boldsymbol{d}_1^{nom} := [\delta d_{1,1}^{nom}, \delta d_{1,2}^{nom}]$. If the future measurements carry less information than the past measurements, then the red ellipsoid will be bigger than the green ellipsoid, and the resulting Pontryagin difference between the confidence regions would result in 0. The over-approximation factor can be obtained directly from (9). The same procedure is applied along the whole prediction horizon to build up the scenario tree. The range of uncertainty along the scenario tree of Dual Multi-stage NMPC 2 is given by

$$\underline{\Delta}^p \boldsymbol{d}_{t+k}^j := \underline{\Delta} \boldsymbol{d}_{t+k} - \delta \boldsymbol{d}_k^j, \quad k = \{0, \ldots, N_{p-1}\}, \tag{10a}$$

$$\bar{\Delta}^p \boldsymbol{d}_{t+k}^j := \bar{\Delta} \boldsymbol{d}_{t+k} + \delta \boldsymbol{d}_k^j, \quad k = \{0, \ldots, N_{p-1}\}. \tag{10b}$$

In the bottom plot of Fig. 2, we show a comparison of the confidence ellipsoids obtained using Dual Multi-stage NMPC 1 and Dual Multi-stage NMPC 2 after taking measurements until time t and $t+1$. The information content of the measurements obtained on the horizon $[t, t+1]$ is reflected in the smaller area of the blue ellipsoid. The black and blue dashed rectangles show the projections of the ellipses to the parametric bounds. This illustrates the evolution of the range of uncertainty for the control scheme Dual Multi-stage NMPC 1. Dual Multi-stage NMPC 2 considers an additional over-approximation factor $\delta \boldsymbol{d}_1^{nom}$ that is added to the range of uncertainty described by the Dual Multi-stage NMPC 1. The blue dotted line illustrates the range of uncertainty for the control scheme Dual Multi-stage NMPC 2.

The overall formulation of the Dual Multi-stage NMPC 2 scheme is given by

$$\min_{\boldsymbol{x}_k^j, \boldsymbol{u}_k^j, \boldsymbol{d}_k^{r(j)} \ \forall (j,k) \in \mathcal{I}} \sum_{i=1}^{N} \omega_i \sum_{k=t}^{t+N_p-1} \ell_k(\boldsymbol{x}_{k+1}^j, \boldsymbol{u}_k^j), \quad \forall \ \boldsymbol{x}_{k+1}^j, \boldsymbol{u}_k^j \in S_i, \tag{11a}$$

subject to:

$$\boldsymbol{x}_{k+1}^j = \boldsymbol{f}(\boldsymbol{x}_k^{p(j)}, \boldsymbol{u}_k^j, \boldsymbol{d}_k^{r(j)}), \quad \forall (j, k+1) \in \mathcal{I}, \tag{11b}$$

$$\boldsymbol{s}_{k+1}^j = \frac{\mathrm{d}\boldsymbol{f}(\boldsymbol{x}_k^{p(j)}, \boldsymbol{u}_k^j, \boldsymbol{d}_k^{r(j)})}{\mathrm{d}\boldsymbol{d}_k^{r(j)}}, \quad \forall (j, k+1) \in \mathcal{I}, \tag{11c}$$

$$\boldsymbol{g}(\boldsymbol{x}_{k+1}^j, \boldsymbol{u}_k^j) \leq \boldsymbol{0} \quad , \quad \forall \ (j, k+1) \in \mathcal{I}, \tag{11d}$$

$$\boldsymbol{u}_k^j = \boldsymbol{u}_k^l \ \text{ if } \boldsymbol{x}_k^{p(j)} = \boldsymbol{x}_k^{p(l)}, \quad \forall \ (j,k), (l,k) \in \mathcal{I}, \tag{11e}$$

$$\left(\mathbf{P}_{0,t}^{\boldsymbol{d}_t^{nom}}\right)^{-1} = \sum_{v=0}^{n_s t} \boldsymbol{s}_v^T(\boldsymbol{d}_t^{nom}) \mathbf{Q} \boldsymbol{s}_v(\boldsymbol{d}_t^{nom}), \tag{11f}$$

$$\left(\mathbf{P}_{t,t+k}^{\boldsymbol{d}_{k-1}^{r(j)}}\right)^{-1} = \sum_{v=n_s t+1}^{(t+k)n_s} \boldsymbol{s}_v^T(\boldsymbol{d}_{k-1}^{r(j)}) \mathbf{Q} \boldsymbol{s}_v(\boldsymbol{d}_{k-1}^{r(j)}), \quad \forall (j,k) \in \mathcal{I}, \tag{11g}$$

$$\left(\boldsymbol{P}_{0,t+k}^{\boldsymbol{d}_k^j}\right)^{-1} = (\boldsymbol{P}_{0,t}^{\boldsymbol{d}_t^{nom}})^{-1} + (\boldsymbol{P}_{t,t+k}^{\boldsymbol{d}_{k-1}^{r(j)}})^{-1}, \quad \forall (j,k) \in \mathcal{I}, \tag{11h}$$

$$\underline{\Delta} \boldsymbol{d}_{t+k}^j = \max \left\{ \underline{\Delta} \boldsymbol{d}^j{}_{t+k-1}, -\mathrm{diag}^{\frac{1}{2}} \left(n_d \mathcal{F}_{n_d,(t+k)n_s n_y - n_d,\alpha} \mathbf{P}_{0,t+k}^{\boldsymbol{d}_t^j} \right) \right\}, \quad \forall (j,k) \in \mathcal{I}, \tag{11i}$$

$$\bar{\Delta} \boldsymbol{d}_{t+k}^j = \min \left\{ \bar{\Delta} \boldsymbol{d}_{t+k-1}^j, \mathrm{diag}^{\frac{1}{2}} \left(n_d \mathcal{F}_{n_d,(t+k)n_s n_y - n_d,\alpha} \mathbf{P}_{0,t+k}^{\boldsymbol{d}_t^j} \right) \right\}, \quad \forall (j,k) \in \mathcal{I}, \tag{11j}$$

$$\delta^a \boldsymbol{d}_k^j = \sqrt{n_d \mathcal{F}_{n_d,(t+k)n_s n_y - n_d,\beta}} \left(\mathrm{diag}^{\frac{1}{2}} \left(\mathbf{P}_{0,t}^{\boldsymbol{d}_t^{nom}} \right) - \mathrm{diag}^{\frac{1}{2}} \left(\mathbf{P}_{t,t+k}^{\boldsymbol{d}_k^{r(j)}} \right) \right), \quad \forall (j,k) \in \mathcal{I}, \tag{11k}$$

$$\delta d_{k,s}^j = \begin{cases} \delta^a d_{k,s}^j, & \text{if } \delta^a d_{k,s}^j > 0, \\ 0, & \text{otherwise,} \end{cases} \quad \forall s \in \{1, \ldots, n_d\}, \quad \forall (j,k) \in \mathcal{I}, \tag{11l}$$

$$\underline{\Delta}^p \boldsymbol{d}_{t+k}^j = \max \left\{ \underline{\Delta} \boldsymbol{d}_{t+k} - \delta \boldsymbol{d}_k^j, \underline{\Delta}^p \boldsymbol{d}_{t+k-1}^j \right\}, \quad k = \{0, \ldots, N_{p-1}\}, \tag{11m}$$

$$\bar{\Delta}^p \boldsymbol{d}^j_{t+k} = \min\left\{\bar{\Delta}\boldsymbol{d}_{t+k} + \delta\boldsymbol{d}^j_k, \bar{\Delta}^p \boldsymbol{d}^j_{t+k-1}\right\}, \quad k = \{0, \ldots, N_{p-1}\}, \tag{11n}$$

$$\boldsymbol{d}^{r(j)}_k \in \mathbb{D}(\boldsymbol{d}^{nom}_t, \underline{\Delta}^p \boldsymbol{d}^j_{t+k}, \bar{\Delta}^p \boldsymbol{d}^j_{t+k}), \quad \forall (j,k) \in \mathcal{I}. \tag{11o}$$

The constraints (11b)–(11j) of Dual Multi-stage NMPC 2 are the same as the constraints of Dual Multi-stage NMPC 1. The over-approximation factor is obtained using (11k) and (11l). The range of uncertainty of the uncertain parameters considered by Dual Multi-stage NMPC 2 is given by (11m) and (11n).

Remark 5. The computation of over-approximation factor increases the computational complexity of Dual Multi-stage NMPC 2 due to the presence of the discontinuous constraint (11l). This can be avoided by computing an estimate of the over-approximation factor using the optimal control problem solved at the previous time step $(t-1)$. The predicted control inputs $(u^j_t, \ldots u^j_{t+N_p-1})$ are extracted from the optimization problem solved at time $t-1$ and are used to compute the over-approximation factor $\delta^j_{k,s}$ using (9). The computed over-approximation factor $\delta^j_{k,s}$ is supplied to the optimization problem (11) solved at time step t, thereby eliminating the need of the constraints (11k) and (11l).

3.4. Implementation of dual multi-stage NMPC

The implementation of Dual Multi-stage NMPC can be done by dividing it into a preparation and an optimization phase.

1. *Preparation phase.* The bounds on the uncertain parameters are obtained from the past measurements as described in Section 3.2. If Dual Multi-stage NMPC 2 is used and some plant measurements are available, $\delta\boldsymbol{d}^j_k$ is computed based on the predicted control inputs obtained from the previous optimization phase. Note that this shifts the computational burden for the calculation of the over-approximation factors from the optimization phase (see Remark 5).
2. *Optimization phase.* After setting up the scenario tree in the preparation phase, problem (8) or (11) is solved depending on the Dual Multi-stage NMPC variant chosen.

The main features of the proposed dual-control scheme can be summarized as follows:

1. The dual aspect of the NMPC controller is realized via including the future reduction of the range of uncertainty along the prediction horizon.
2. Robustness (w.r.t. both the objective and the constraints) is achieved via the scenario-tree representation of the possible uncertainty realizations in a nonconservative manner because recourse is taken into account.
3. The estimation of bounds on the uncertain parameters is achieved via projection of the confidence ellipsoids that are computed using the Fisher information matrix and the parametric sensitivities along the prediction horizon.

Table 1 outlines the main differences between the multi-stage robust control schemes according to the distribution of the computational burden among the preparation and the optimization of the future inputs. The Multi-stage NMPC controller considers the given a-priori confidence in the uncertain parameters and the parametric bounds remain constant throughout the NMPC run. The scenario tree of the Adaptive Multi-stage NMPC scheme is updated at each sampling time. The computational burden is shifted to the preparation phase where the least-squares estimation and the a-posteriori confidence analysis are performed. The scenario tree of Dual Multi-stage NMPC 1 is not only updated at each sampling time of the NMPC as in Adaptive Multi-stage NMPC but also an estimate of the future confidence regions, as a result of applying the future control inputs must be obtained along the prediction horizon, under the assumption that the least-squares estimate has converged to the true values of the uncertain parameters.

Table 1
Comparison of the computational burden for determining the parametric bounds of the presented multi-stage schemes, where $\left(\boldsymbol{P}^{\boldsymbol{d}^j_k}_{0,t+k}\right)^{-1} := (\boldsymbol{P}^{\boldsymbol{d}^{nom}_{N_m}}_{0,N_m+N_r n_s})^{-1} + (\boldsymbol{P}^{\boldsymbol{d}^{r(j)}_{k-1}}_{N_m+N_r n_s, N_m+kn_s})^{-1}$, $\Delta\boldsymbol{d}^j_{t+k} := [\underline{\Delta}\boldsymbol{d}^j_{t+k}, \bar{\Delta}\boldsymbol{d}^j_{t+k}]$ and $\Delta^p\boldsymbol{d}^j_{t+k} := [\underline{\Delta}^p\boldsymbol{d}^j_{t+k}, \bar{\Delta}^p\boldsymbol{d}^j_{t+k}]$.

	Optimization phase ($\boldsymbol{P}_{t,t+k}$)	Preparation phase
Multi-stage	–	$\boldsymbol{P}_0$
Adaptive Multi-stage	–	$\boldsymbol{d}^{nom}_t$, $\boldsymbol{P}^{\boldsymbol{d}^{nom}_t}_{0,t}$
Dual Multi-stage 1	$\Delta\boldsymbol{d}^j_{t+k}$, $\boldsymbol{P}^{\boldsymbol{d}^j_k}_{0,t+k}$	$\boldsymbol{d}^{nom}_t$, $\boldsymbol{P}^{\boldsymbol{d}^{nom}_t}_{0,t}$
Dual Multi-stage 2	$\Delta\boldsymbol{d}^j_{t+k}$, $\Delta^p\boldsymbol{d}^j_{t+k}$, $\boldsymbol{P}^{\boldsymbol{d}^j_k}_{0,t+k}$	$\boldsymbol{d}^{nom}_t$, $\delta\boldsymbol{d}^j_k$, $\boldsymbol{P}^{\boldsymbol{d}^{nom}_t}_{0,t}$

Remark 6. Dual Multi-stage NMPC requires the sensitivity information about the uncertain parameters to compute the range of the uncertain parameters along the prediction horizon. The computation of the sensitivity information of the uncertain parameters increases the computational complexity of the resulting problem considerably ($n_x \times n_d$ explicit sensitivity equations are required to compute the output sensitivities w.r.t. uncertain parameters). This limits the scalability of the proposed approach. For large-scale case studies, the computational complexity can be reduced by replacing the explicit sensitivity equations by finite difference approximations to compute the sensitivity information of the uncertain parameters. The predictions obtained for different realizations of the uncertain parameters along the prediction horizon can be used to compute the sensitivity information of the uncertain parameters using finite differences. This reduces the computational complexity of the proposed approach but also the accuracy of the sensitivity information.

4. Case study

We study an exothermic reaction A + B → C that takes place in a semi-batch reactor equipped with a cooling jacket, where reactants A (of concentration c_A) and B (of concentration c_B) react to produce a component C (of concentration c_C). The operational goal is to maximize the amount of the product C at the end of the fixed batch time t_f under constraints on the reactor temperature T and volume V. The mathematical model of the reactor reads as:

$$\frac{dV}{dt} = \dot{V}_{in}, \tag{12}$$

$$\frac{dc_A}{dt} = -\frac{\dot{V}_{in}}{V}c_A - kc_A c_B, \tag{13}$$

$$\frac{dc_B}{dt} = \frac{\dot{V}_{in}}{V}(c_{B,in} - c_B) - kc_A c_B, \tag{14}$$

$$\frac{dc_C}{dt} = -\frac{\dot{V}_{in}}{V}c_C + kc_A c_B, \tag{15}$$

$$\frac{dT}{dt} = \frac{\dot{V}_{in}}{V}(T_{in} - T) - \frac{UA_w(T - T_J)}{\rho V c_p} - \frac{kc_A c_B H}{\rho c_p}, \tag{16}$$

$$\frac{dT_J}{dt} = \frac{\dot{V}_{J,in}}{V_J}(T_{J,in} - T_J) + \frac{UA_w(T - T_J)}{\rho V_J c_p}, \tag{17}$$

Table 2
Model parameters, operating bounds and initial conditions.

Parameter	Value	Units
H	$[-461.5, -248.5]$	kJ mol^{-1}
k	$[2.343, 4.351] \times 10^{-7}$	m^3 mol^{-1} s^{-1}
ρ	1000.0	kg m^{-3}
c_p	4.2	kJ kg^{-1} K^{-1}
r	0.092	m
V_J	2.22×10^{-3}	m^3
$\dot{V}_{J,\text{in}}$	9.167×10^{-5}	m^3 s^{-1}
U	0.14844	kJ K^{-1} s^{-1} m^{-2}
τ_c	900	s
$c_{B,\text{in}}$	3000	mol m^{-3}
T_{in}	300	K
$\dot{V}_{\text{in}}$	$[0.0, 9.0] \times 10^{-6}$	m^3 s^{-1}
$T_{Jin,set}$	[280, 350]	K
$V_{\max}$	7×10^{-3}	m^3
T	[321, 325]	K
V_0	3.5×10^{-3}	m^3
$c_{A,0}$	2000	mol m^{-3}
$c_{B,0}$	0	mol m^{-3}
$c_{C,0}$	0	mol m^{-3}
$T_0, T_{J,0}, T_{J,\text{in},0}$	325	K

$$\frac{dT_{J,\text{in}}}{dt} = \frac{1}{\tau_c}(T_{J,\text{in,set}} - T_{J,\text{in}}), \tag{18}$$

with:

$$A_w = \frac{2V}{r} + \pi r^2, \tag{19}$$

where V and V_J are the volumes of the reactor content and of the jacket. $\dot{V}_{\text{in}}$ is the feed rate of component B of concentration $c_{B,\text{in}}$. T, T_{in}, T_J stand for the temperatures of the reactor, the feed and the jacket. The variables $T_{J,\text{in}}$ and $T_{J,\text{in,set}}$ are the jacket inlet temperature and its set point, τ_c is the time constant of the cooling system. The reaction rate constant and the reaction enthalpy are considered as uncertain (±30% w.r.t. their nominal values) and are denoted by k and H. The heat transfer between the reactor of radius r and the jacket is governed by the surface of the reactor wall covered by the reaction mixture A_w, and its heat transfer coefficient U. The density and the heat capacity of the reaction mixture, ρ and c_p, are both assumed to be constant. The control inputs are summarized in $\boldsymbol{u} = (\dot{V}_{\text{in}}, T_{J,\text{in,set}})^T$. The initial conditions are $\boldsymbol{x}_0 = (V_0, c_{A,0}, c_{B,0}, c_{C,0}, T_0, T_{J,0}, T_{J,\text{in},0})^T$ and all the states are measured with measurement noise with standard deviations $(\sigma_V, \sigma_A, \sigma_B, \sigma_C, \sigma_T, \sigma_{T_J}, \sigma_{T_{J,\text{in}}})^T$. Table 2 summarizes the values of the model parameters, operating bounds, and initial conditions. The nominal MPC optimization problem solved at time t is given by

$$\min_{\dot{V}_{\text{in}}(k),\, T_{J,\text{in,set}}(k)} \sum_{k=t}^{t+N_p-1} -V(k)C_C(k) + 10^{-1}\Delta\dot{V}_{\text{in}}^2(k) + 10^{-6}\Delta T_{J,\text{in,set}}^2(k), \tag{20}$$

subject to (12) to (19)

$$0\,\text{m}^3 \le V(k) \le 0.007\,\text{m}^3, \tag{21}$$

$$321\,\text{K} \le T(k) \le 325\,\text{K}, \tag{22}$$

$$0\,\text{m}^3\,\text{s}^{-1} \le \dot{V}_{\text{in}}(k) \le 9 \times 10^{-6}\,\text{m}^3\,\text{s}^{-1}, \tag{23}$$

$$280\,\text{K} \le T_{J,\text{in,set}}(k) \le 350\,\text{K}, \tag{24}$$

where $\Delta u(k) = u(k) - u(k-1)$ gives the deviation between two consecutive control moves. The optimization criterion is the maximization of the mass of product C (n_C) along the prediction horizon while penalizing the deviation between two consecutive control moves. The constraints on the volume of the reactor content and the reactor temperature are given by (21) and (22). The constraints on the manipulated variables are given by (23) and (24).

5. Results

The multi-stage NMPC, adaptive multi-stage NMPC, dual multi-stage NMPC 1 and Dual Multi-stage NMPC 2 schemes were implemented for the aforementioned case study. The arising dynamic optimization problems were solved using orthogonal collocation on finite elements, CasADi [52] and IPOPT [55].

In all presented simulation studies the scenario tree is generated at the initial time by considering all combinations of the maximum, minimum and nominal values of the uncertain parameters. The sampling time of the controller is 120 s and the samples from the process are available each 60 s, i.e., $n_s = 2$. The end time (t_f) is chosen as 1200 s. We are interested in the behavior of the plant only until the end of the batch. The degree of the Lagrange polynomials for the state variables in orthogonal collocation is chosen as 2. This results in an optimization problem with 2733 decision variables and 2700 constraints for solving a Multi-stage NMPC and Adaptive Multi-stage NMPC formulation, and an optimization problem with 8712 decision variables and 8625 constraints for solving a Dual Multi-stage NMPC problem. The measurement errors are considered as uncorrelated with the variances given as $\sigma_V = 10^{-7}$ m^3, $\sigma_A = \sigma_B = \sigma_C = 5$ mol m^{-3} and $\sigma_T = \sigma_{T_J} = \sigma_{T_{J,\text{in}}} = 0.2$ K. The confidence level α which is used for building the confidence ellipsoids, is 99.7%. The optimization criterion used for all controllers is given by (20) subject to the constraints on the manipulated variables, the reactor temperature and the volume of the reactor. The evaluation criterion used to compare the performance of different NMPC schemes is given by

$$\text{Performance Improvement} \quad [\%] = \frac{n_c^A - n_c^B}{n_c^B} \times 100, \tag{25}$$

where "Performance Improvement" represents the improvement in the performance obtained using the NMPC controller A over the NMPC controller B, where n_c^i denotes the number of moles of the product C produced at the end of the batch when the reactor is controlled by the controller NMPC i.

5.1. Assessment of Dual Multi-stage NMPC

An assessment of the performance of the dual multi-stage controllers is shown in the following for different controller parameters. The base case used for the comparison of different Multi-stage NMPC strategies considers a prediction horizon $N_p = 10$, and a robust horizon $N_r = 1$. The tuning parameter β that is used to obtain the over-approximation factors $\delta\boldsymbol{d}_k^j$ for Dual Multi-stage NMPC 2 is chosen as 68%. The realization of the parametric uncertainty is considered to be such that the true values of the uncertain parameters are 15% smaller than their nominal values, i.e., the reaction is slower and less exothermic in reality as compared to the model. The influence of the tuning parameters for different true values of the uncertain parameters is studied below.

Fig. 3 shows the base case results for the control of one batch using Multi-stage NMPC, Adaptive Multi-stage NMPC and the Dual Multi-stage NMPC schemes. Nominal NMPC is not shown because it violates the constraints and renders the optimization problem infeasible. The optimal operation of the process consists in feeding as much as possible while respecting the constraints. In the first stage of the batch, a balance has to be found between pumping the cold feed stream and heating up the reactor such that the temperature does not go below its lower limit. The later stages are characterized by the increase of the reactor temperature because the cold feed stream entering into the reactor is stopped once the reactor volume reaches its upper bound. As a consequence, the reactor temperature reaches a maximum value and then starts to decrease because the reaction rate decreases once the reac-

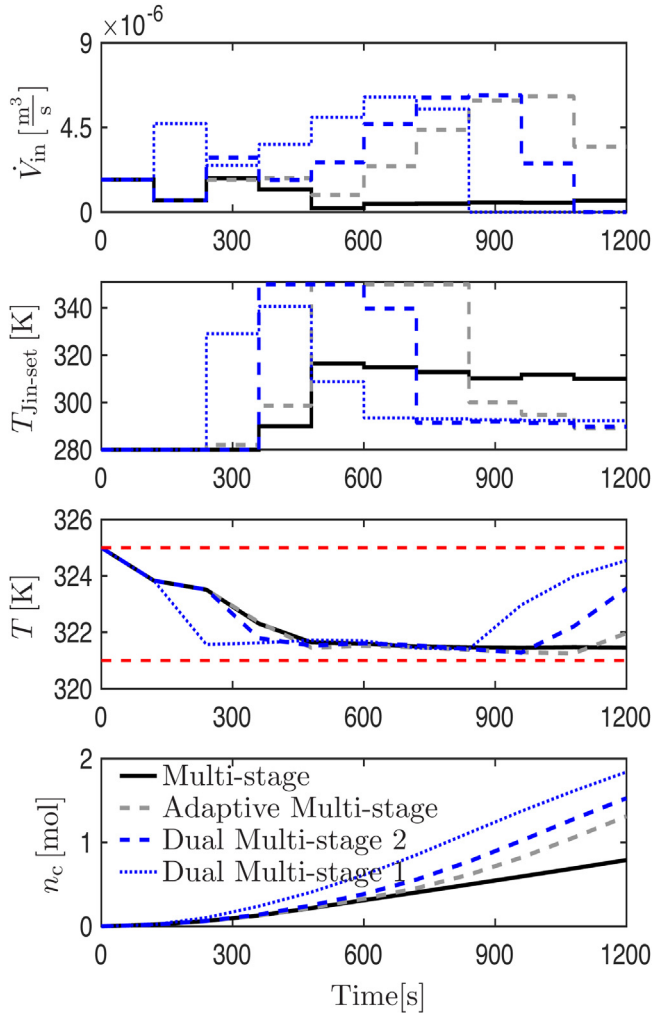


Fig. 3. Input feed, set-point for the temperature of the jacket, reactor temperature and number of moles of product C for one batch using different controllers.

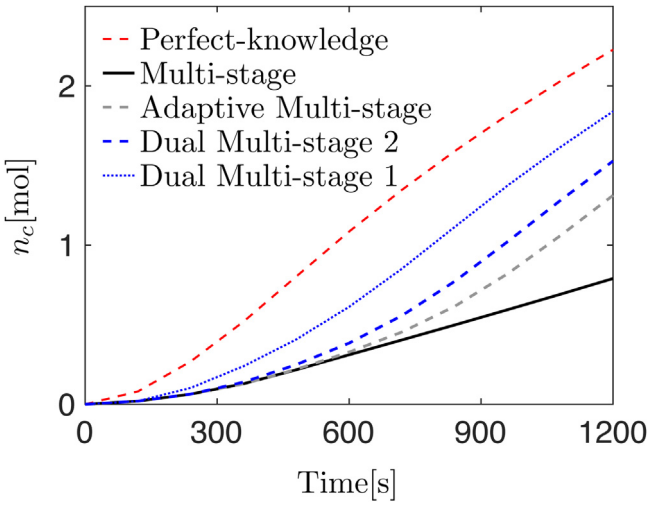


Fig. 4. Performance of different control schemes over time in terms of mass of product.

tant B does not enter the reactor. Only the set-point of the inlet temperature of the cooling jacket is adjusted. While the input-affine structure of the control problem might lead to singular-type solutions, this phenomenon is not present for the multi-stage formulations. This occurs because the solutions along the scenario tree are dominated by the active input and path constraints due to tight operation constraints and also due to the nature of the multi-stage method (i.e., branches of the scenario tree may exhibit different active constraints).

Unlike nominal NMPC, which assumes the nominal values of the uncertain parameters, Multi-stage NMPC is able to satisfy the constraints even though the true realization of the uncertainty is not part of the scenario tree. At the end of the batch, 0.8 mol of product C are obtained due to a relatively cautious feeding strategy. Due to the large uncertainty, the results of the application of Multi-stage NMPC differ significantly from the nominally optimal control inputs, especially at the later stages of the batch.

When Adaptive Multi-stage NMPC is used, the controller increases the set-point of the jacket inlet temperature $T_{J,in,set}$ and the feed $\dot{V}_{\text{in}}$ after the third control interval, once more measurements are available, and better estimates (and bounds) of the uncertain parameters are obtained. The control actions satisfy the operational constraints and as the feed to the reactor is increased, a larger amount of product C (by around 66%) is obtained at the end of the batch as compared to Multi-stage NMPC without adaptation. The estimation of the uncertainty results in a reduction of the conservativeness (back-off from the temperature constraints) of the controller as compared to Multi-stage NMPC.

When using Dual Multi-stage NMPC 2, the performance can be further increased. In this case the set-point of the jacket inlet temperature is increased one control step earlier (in the third step) hence more feed is introduced into the reactor when compared to Adaptive Multi-stage NMPC. This results in a larger amount of product C at the end of the batch. The performance can be further increased using Dual Multi-stage NMPC 1. The set-point of the jacket inlet temperature is increased at $t = 1$ and more feed is introduced into the reactor when compared to Dual Multi-stage NMPC 2. The set-point of the jacket inlet temperature obtained using all the NMPC schemes does not reach the lower limit at the later stages. This infers that the reactor temperature obtained along the prediction horizon decreases after a certain point in time.

One can compare the performance of the presented robust multi-stage controllers to a perfect-knowledge NMPC controller, i.e., the controller that possesses the knowledge about the true behavior of the plant. In Fig. 4, the evolution of the number of moles of product obtained under different NMPC strategies is shown. The dashed red line represents the performance of the perfect-knowledge controller over time. It can be seen that Dual Multi-stage NMPC reaches a performance that is closer to the perfect-knowledge controller compared to the other robust multi-stage control strategies. There is a 17% increase in the amount of the product produced using the Dual Multi-stage NMPC 2 scheme when compared to Adaptive Multi-stage NMPC and a 40% increase when using Dual Multi-stage NMPC 1 over Adaptive Multi-stage NMPC. The improved performance of Dual Multi-stage NMPC 1 with respect to Dual Multi-stage NMPC 2 is due to the optimistic assumption of having a constant least-squares estimate in the predictions. We discuss this below in more detail.

The are two reasons behind the improved performance of Dual Multi-stage NMPC. Firstly, Dual Multi-stage NMPC is aware of the fact that the parametric bounds will improve in the future, as this effect is present in the prediction horizon of the controller, unlike in the case of nondual control schemes. Secondly, the controller is aware of the influence of the control inputs on the bounds of the uncertain parameters. The inputs are thus calculated to increase the output sensitivities w.r.t. the uncertain parameters to shape the underlying confidence intervals in the desired manner.

Fig. 5 shows the confidence ellipsoids obtained for two different scenarios of the tree. The parameter values are scaled from 0

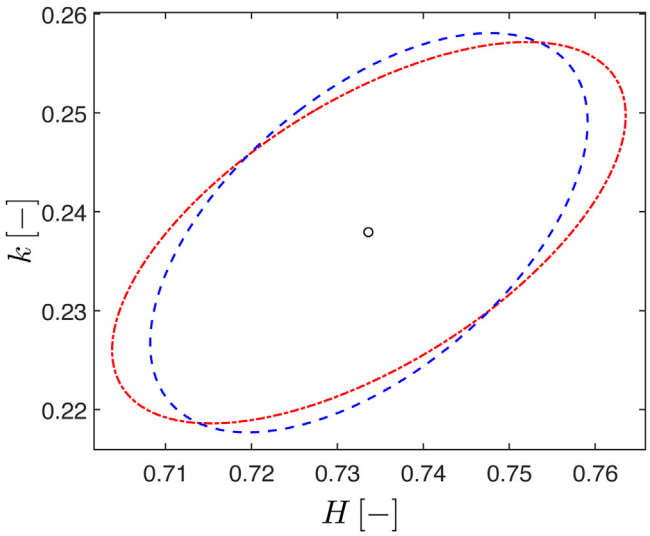


Fig. 5. Confidence ellipsoids obtained in scenarios with $(H^{nom}, \underline{k})$ (red dash-dotted) and $(\bar{H}, k^{nom})$ (blue dashed). Parameters are normalized to the interval [0, 1]. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

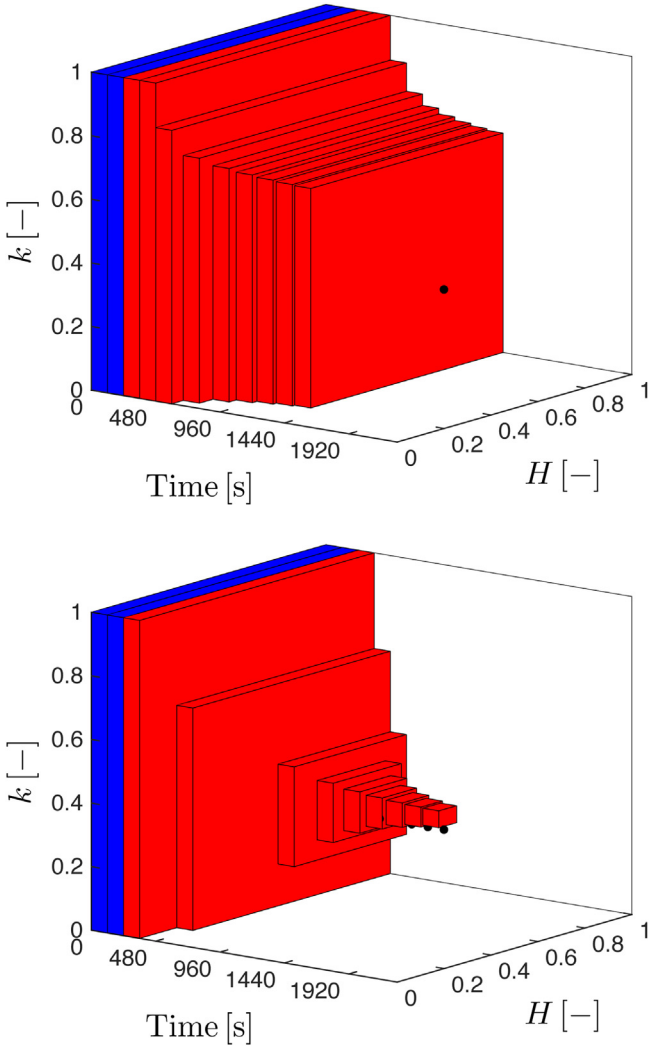


Fig. 6. A priori (blue) and predicted (red) parametric bounds along the prediction horizon used by the Dual Multi-stage NMPC 2 (top plot) and by the Dual Multi-stage NMPC 1 (bottom plot) at time $t = 240$ s. The black dot represents the true value of the uncertain parameters. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

to 1, which corresponds to the minimum and to the maximum value of each parameter. In the scenario with the nominal value of the enthalpy and the lower bound on the reaction rate k, i.e., $(H^{nom}, \underline{k})$, the value of H^{nom} is fixed to the least-squares estimate, $\underline{k}$ is determined from the parameter covariance matrix obtained along the prediction horizon. The number of moles of product C produced increases with the increase in the value of the reaction constant k. The value of $\underline{k}$ can be increased by the decreasing the uncertainty associated with the reaction constant k. Hence, the controller adjusts the control actions along this scenario such that the uncertainty decreases in the reaction rate constant. A similar tendency can be seen when plotting the confidence ellipsoid (red dash-dotted) obtained for scenario $(\bar{H}, k^{nom})$, where the amount of heat produced in the reactor increases with a decrease in the value of $\bar{H}$. Hence, more reactant can be fed into the reactor during the initial stages where the lower constraint on the reactor temperature is active.

Fig. 6 shows the evolution of the parametric bounds along the prediction horizon after the second step of the NMPC controller. Depicted in blue are the a-priori bounds obtained from the parameter estimates using the measured data. The bounds predicted for the nominal scenario are shown in red. We can observe that the bounds enclose the true parameter in all prediction steps for Dual Multi-stage NMPC 2 (top plot of Fig. 6). In contrast, the true parameters are not enclosed within the predicted bounds for Dual Multi-stage NMPC 1 (bottom plot of Fig. 6), which assumes that the least-squares estimate remains constant along the prediction horizon. Nonetheless, the controller does not fail to satisfy the constraints along the batch run. As the predicted parametric bounds shrink rapidly in this case, the controller is less cautious and improves the product yield by 20%.

The initial range of the uncertain parameters affects the performance of the multi-stage NMPC much more when compared to Adaptive and Dual Multi-stage NMPC. Multi-stage NMPC decreases the amount of feed that enters into the reactor if the initial range of uncertainty of the uncertain parameters is increased due to the tight specification on the reactor temperature, whereas this affects only the initial stages of Adaptive and Dual Multi-stage NMPC because of the updates of the range of the uncertain parameters whenever new information is available from the plant. For example, if the initial range of the uncertain parameters is increased from ±30% of their nominal values to ±50% of their nominal values, the number of moles of product C produced at the end of the batch decreases from 0.79 to 0.33 for Multi-stage NMPC, whereas it decreases from 1.31 to 1.02 for Adaptive Multi-stage NMPC and from 1.53 to 1.17 when using Dual Multi-stage NMPC 2. The decrease in the number of moles of the product C produced using Adaptive Multi-stage NMPC and Dual Multi-stage NMPC 2 is due to the reduction in the information gained about the uncertain parameters implied by the cautious control decisions taken at the beginning of the batch.

Further we evaluate the performance of the robust multi-stage controllers dependent on the tuning parameters. In order to avoid the influence of particular realization of the true values of the parameters, we chose 25 points on a uniform grid with values of the parameter H ranging from $-408.3\,\mathrm{kJ\,mol^{-1}}$ to $-301.8\,\mathrm{kJ\,mol^{-1}}$ and k ranging from $2.85 \times 10^{-7}\,\mathrm{m^3\,mol^{-1}\,s^{-1}}$ to $3.85 \times 10^{-7}\,\mathrm{m^3\,mol^{-1}\,s^{-1}}$. The performance is reported as statistics obtained over the 25 batches with different values of reaction rate and reaction enthalpy.

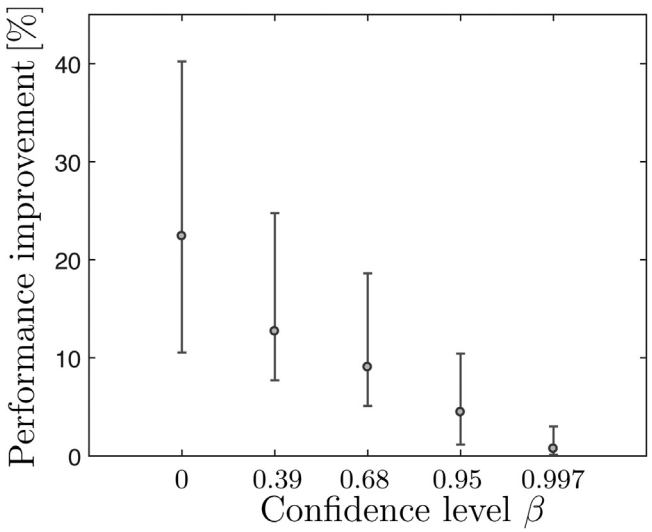


Fig. 7. Comparison of the maximum, average and minimum performance improvement of Dual Multi-stage NMPC 2 over Adaptive Multi-stage NMPC using 25 different realizations of the uncertain parameters for different choices of the parameter β.

Table 3
The value of the over-approximation factor $\delta \boldsymbol{d}_k^j$ for different choices of β obtained along the prediction horizon ($1 \le k \le 9$) at the nominal scenario (H^{nom}, k^{nom}) of the Dual Multi-stage NMPC 2 problem solved at 240 s. The parameters are normalized to the interval [0, 1].

k	β				
	0	0.39	0.68	0.95	0.997
1	$(0, 0)^T$	$(0.27, 0.20)^T$	$(0.42, 0.31)^T$	$(0.62, 0.44)^T$	$(1.68, 1.20)^T$
2	$(0, 0)^T$	$(0.28, 0.21)^T$	$(0.47, 0.35)^T$	$(1.10, 0.81)^T$	$(2.99, 2.22)^T$
3	$(0, 0)^T$	$(0.29, 0.21)^T$	$(0.48, 0.36)^T$	$(1.15, 0.85)^T$	$(3.16, 2.33)^T$
4	$(0, 0)^T$	$(0.29, 0.22)^T$	$(0.49, 0.36)^T$	$(1.17, 0.86)^T$	$(3.21, 2.37)^T$
5	$(0, 0)^T$	$(0.29, 0.22)^T$	$(0.49, 0.37)^T$	$(1.18, 0.87)^T$	$(3.24, 2.39)^T$
6	$(0, 0)^T$	$(0.30, 0.22)^T$	$(0.50, 0.37)^T$	$(1.18, 0.88)^T$	$(3.25, 2.41)^T$
7	$(0, 0)^T$	$(0.30, 0.22)^T$	$(0.50, 0.37)^T$	$(1.19, 0.88)^T$	$(3.26, 2.41)^T$
8	$(0, 0)^T$	$(0.30, 0.22)^T$	$(0.50, 0.37)^T$	$(1.19, 0.88)^T$	$(3.27, 2.42)^T$
9	$(0, 0)^T$	$(0.30, 0.22)^T$	$(0.50, 0.37)^T$	$(1.20, 0.88)^T$	$(3.27, 2.42)^T$

5.2. Effect of the robustness level on the future parameter estimates

Fig. 7 shows the maximum, average and minimum improvement in the performance of Dual Multi-stage NMPC over Adaptive Multi-stage NMPC for different choices of the value β which determines the over-approximation factor $\delta \boldsymbol{d}_k^j$. The value of the over-approximation factor $\delta \boldsymbol{d}_k^j$ for different choices of β obtained along the prediction horizon for the nominal scenario (H^{nom}, k^{nom}) is shown in Table 3. It increases with an increase in the confidence level β, which in turn results in a decrease in the performance of Dual Multi-stage NMPC 2. The best performance is achieved for $\beta = 0$ which coincides with Dual Multi-stage NMPC 1 (i.e., $\delta \boldsymbol{d}_k^j = (0,\ 0)^T$). The constraints are satisfied for all choices of the tuning parameter β in this case. Even if $\delta \boldsymbol{d}_k^j \ge 1$, the confidence region of the uncertain parameter considered in the scenario tree of the Dual Multi-stage NMPC 2 is not bigger than the confidence region of the uncertain parameter obtained at the previous stage. This is realized using constraints (11m) and (11n) in the optimization problem. The value of β is a tuning parameter of Dual Multi-stage NMPC 2 and it is a decision of the user to select particular value of β based on simulation studies for a concrete problem at hand.

Table 4
Comparison of the average performance improvement in % of Multi-stage NMPC, Adaptive Multi-stage NMPC and Dual Multi-stage NMPC 2 using 25 different realizations of the uncertain parameters for different lengths of the prediction horizon.

N_p	Multi-stage vs. Adaptive Multi-stage	Multi-stage vs. Dual Multi-stage 2	Adaptive Multi-stage vs. Dual Multi-stage 2
8	56.02	68.50	7.99
9	57.45	70.96	8.58
10	57.99	71.98	8.86
11	57.98	72.20	8.99
12	57.90	72.27	9.10
13	57.87	72.24	9.11
14–15	57.81	72.20	9.12

Table 5
Comparison of the average performance improvement in % of Multi-stage NMPC, Adaptive Multi-stage NMPC and Dual Multi-stage NMPC 2 using 25 different realizations of the uncertain parameters for different lengths of the robust horizon.

N_r	Multi-stage vs. Adaptive Multi-stage	Multi-stage vs. Dual Multi-stage 2	Adaptive Multi-stage vs. Dual Multi-stage 2
1	57.99	71.98	8.86
2	61.36	76.00	9.07

5.3. Effect of the length of the prediction horizon

The length of the prediction horizon is one of the most important tuning parameters of model predictive controllers. For a dual controller, the value of N_p plays an important role as it directly influences the cautiousness of the controller. Table 4 shows the comparison between the average performance of Multi-stage, Adaptive Multi-stage and Dual Multi-stage NMPC 2 for different values of N_p. It can be seen that higher values of N_p result in larger performance gains of the dual controller over Multi-stage NMPC and Adaptive Multi-stage NMPC. There is no significant improvement in the performance gain of Dual Multi-stage NMPC 2 over Adaptive Multi-stage NMPC if the length of the prediction horizon is larger than 14.

5.4. Effect of the length of the robust horizon

Table 5 shows the comparison between the average performance improvements obtained using Multi-stage, Adaptive Multi-stage and Dual Multi-stage NMPC 2 for different values of the robust horizon N_r. The number of moles of product C produced reduces with the increase in the robust horizon. The cautiousness of Multi-stage NMPC increases with the increase in the robust horizon because $N_r = 1$ implies that the future uncertainty is constant after the first-time step whereas $N_r = 2$ implies that the uncertain parameters vary over the next two steps. The increase in the knowledge about the uncertain parameters when using Dual Multi-stage NMPC results in an increase in the performance improvement obtained over other multi-stage NMPC schemes.

The computational complexity of a multi-stage NMPC scheme increases exponentially w.r.t. the length of the robust horizon. Dual Multi-stage NMPC 2 with robust horizon equal to 2 solves an optimization problem with 71,136 decision variables and 70,617 constraints whereas an optimization problem with 8712 decision variables and 8625 constraints is solved if the robust horizon is equal to 1.

5.5. Effect of the measurement noise

The performance of the Multi-stage NMPC schemes for different standard deviations of the measurement error is studied in this subsection. The comparison between the maximum, average and

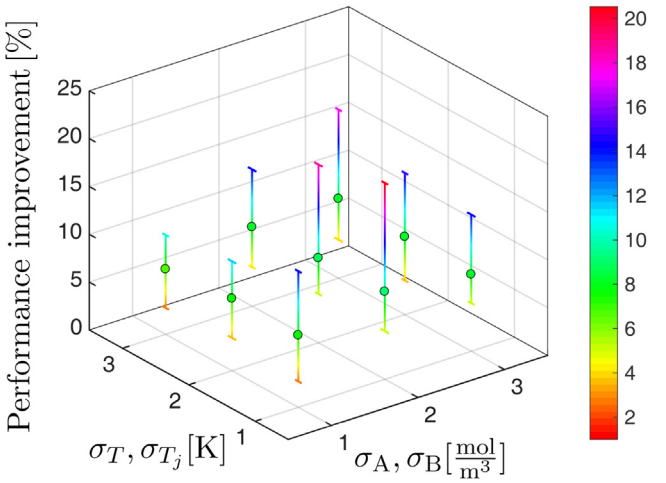


Fig. 8. Comparison between the maximum, average and minimum performance improvements of the Dual Multi-stage NMPC 2 over Adaptive Multi-stage NMPC using 25 different realizations of the uncertain parameters for different standard deviations of measurement errors.

minimum performance improvements of Dual Multi-stage NMPC 2 over Adaptive Multi-stage NMPC is shown in Fig. 8 for one particular realization of the measurement error. The performance improvement of Dual Multi-stage NMPC 2 w.r.t. Adaptive Multi-stage NMPC is approx. 8% on the average with cases where this improvement drops to approx. 3% and rises to approx. 21% for different standard deviations of the measurement errors. In general, the performance improvement increases with the increase in the standard deviations of the measurement errors.

It can be seen from the results presented in this section that the proposed Dual Multi-stage NMPC scheme performs better than the other robust multi-stage controllers. The average improvement in the performance obtained while using Dual Multi-stage NMPC 1 over Adaptive Multi-stage NMPC is around 22% and the maximum improvement obtained is around 40%. The average performance gain of Dual Multi-stage NMPC 2 ($\beta = 0.68$) w.r.t. the Adaptive Multi-stage NMPC is around 9% and the maximum performance gain is around 19%. Despite the approximation introduced by the description of the future uncertainty as an ellipsoid and the approximation of the center of the ellipsoid to calculate the future confidence region, no constraint violations were observed in any of the simulations.

The price to pay for the increased performance is a higher computational effort. The Multi-stage NMPC and the Adaptive Multi-stage NMPC schemes can be solved on the average in 1 s while dual multi-stage NMPC 1 and 2 need around 10 s because of the simultaneous computation of the parametric sensitivities and the increased number of constraints and degrees of freedom.

6. Conclusion

This paper proposes a robust dual NMPC approach that is based on a multi-stage (scenario-based) formulation. The multi-stage approach reduces the conservatism of the robust control actions since it considers that the future control inputs can be adapted to the future observations and the adaptive and dual versions also consider the future reduction of the range of the uncertainty, in a passive manner (adaptive) and in an active manner (dual). It is not necessary to decide a-priori on the relative importance of the control goal and the gain of information about the system. The proposed Dual Multi-stage NMPC approach is applied to a semi-batch reactor example and the results show the advantages of the method with respect to other robust control approaches based on multi-stage NMPC. The price to pay for incorporating dual-control principles into the robust NMPC framework is the increased nonlinearity and complexity of the optimization problems and consequently an increased computation time. Our future work will address this aspect by proposing more sophisticated approaches to obtain the parametric sensitivity information using first-order Taylor approximations based on the evolution of the states along the scenario tree.

Instead of relying on the approximation of the parameter covariance matrix using the Fisher information matrix, other methods with guarantees can be used to identify the bounds on the uncertain parameters [56]. However, this requires additional assumptions on the future control inputs and their variations.

Acknowledgments

The research leading to these results has received funding from the European Commission under grant agreement number 291458 (ERC Advanced Investigator Grant MOBOCON). The authors also acknowledge the contribution of German Academic Exchange Service (DAAD) and The Ministry of Education, Science, Research and Sport of the Slovak Republic under the Exchange involving projects (PPP) project “Reliable and Real-time Feasible Estimation and Control of Chemical Plants”. RP acknowledges the contribution of Slovak Research and Development Agency under the project APVV 15-0007, the contribution of the Scientific Grant Agency of the Slovak Republic under the grant 1/0004/17 and the Research & Development Operational Programme for the project University Scientific Park STU in Bratislava, ITMS 26240220084, supported by the Research 7 Development Operational Programme funded by the ERDF.

References

[1] J. Richalet, A. Rault, J. Testud, J. Papon, Model predictive heuristic control: applications to industrial processes, Automatica 14 (5) (1978) 413–428, http://dx.doi.org/10.1016/0005-1098(78)90001-8.
[2] C.R. Cutler, B.L. Ramaker, Dynamic matrix control: a computer control algorithm, Joint Autom. Control Conf. 17 (1980) 72, http://dx.doi.org/10.1109/JACC.1980.4232009.
[3] C.E. García, D.M. Prett, M. Morari, Model predictive control: theory and practice: a survey, Automatica 25 (3) (1989) 335–348, http://dx.doi.org/10.1016/0005-1098(89)90002-2.
[4] A. Bemporad, M. Morari, Robust model predictive control: a survey, in: Robustness in Identification and Control, Springer London, London, 1999, pp. 207–226, http://dx.doi.org/10.1007/BFb0109870.
[5] J.H. Lee, Model predictive control: review of the three decades of development, Int. J. Control Autom. Syst. 9 (3) (2011) 415, http://dx.doi.org/10.1007/s12555-011-0300-6.
[6] A.A. Jalali, V. Nadimi, A survey on robust model predictive control from 1999–2006, in: Proceedings of the International Conference on Computational Intelligence for Modelling Control and Automation and International Conference on Intelligent Agents Web Technologies and International Commerce, CIMCA'06, IEEE Computer Society, Washington, DC, USA, 207, 2006, http://dx.doi.org/10.1109/CIMCA.2006.29.
[7] P.J. Campo, M. Morari, Robust model predictive control, American Control Conference (1987) 1021–1026.
[8] D. Mayne, M. Seron, S. Raković, Robust model predictive control of constrained linear systems with bounded disturbances, Automatica 41 (2) (2005) 219–224, http://dx.doi.org/10.1016/j.automatica.2004.08.019.
[9] S. Yu, C. Maier, H. Chen, F. Allgöwer, Tube MPC scheme based on robust control invariant set with application to Lipschitz nonlinear systems, Syst. Control Lett. 62 (2) (2013) 194–200, http://dx.doi.org/10.1016/j.sysconle.2012.11.004.
[10] S. Lucia, T. Finkler, S. Engell, Multi-stage nonlinear model predictive control applied to a semi-batch polymerization reactor under uncertainty, J. Process Control 23 (9) (2013) 1306–1319, http://dx.doi.org/10.1016/j.jprocont.2013.08.008.
[11] H. Witsenhausen, A minimax control problem for sampled linear systems, IEEE Trans. Autom. Control 13 (1) (1968) 5–21, http://dx.doi.org/10.1109/TAC.1968.1098788.
[12] P.O.M. Scokaert, D.Q. Mayne, Min–max feedback model predictive control for constrained linear systems, IEEE Trans. Autom. Control 43 (8) (1998) 1136–1142, http://dx.doi.org/10.1109/9.704989.

[13] J. Lee, Z. Yu, Worst-case formulations of model predictive control for systems with bounded parameters, Automatica 33 (5) (1997) 763–781, http://dx.doi.org/10.1016/S0005-1098(96)00255-5.
[14] D. Mayne, Control of constrained dynamic systems, Eur. J. Control 7 (2) (2001) 87–99, http://dx.doi.org/10.3166/ejc.7.87-99.
[15] D.Q. Mayne, E.C. Kerrigan, Tube-based robust nonlinear model predictive control, IFAC Proc. Vol. 40 (12) (2007) 36–41, http://dx.doi.org/10.3182/20070822-3-ZA-2920.00006 (7th IFA, Symposium on Nonlinear Control Systems).
[16] J.B. Rawlings, D.Q. Mayne, Model Predictive Control: Theory and Design, Nob Hill Pub, 2009.
[17] D.Q. Mayne, E.C. Kerrigan, P. Falugi, Robust model predictive control: advantages and disadvantages of tube-based methods, IFAC Proc. Vol. 44 (1) (2011) 191–196, http://dx.doi.org/10.3182/20110828-6-IT-1002.01893 (18th IFA, World Congress).
[18] S.V. Raković, B. Kouvaritakis, M. Cannon, C. Panos, R. Findeisen, Fully parameterized tube MPC, IFAC Proc. Vol. 44 (1) (2011) 197–202, http://dx.doi.org/10.3182/20110828-6-IT-1002.03110 (18th IFA, World Congress).
[19] B. Kouvaritakis, M. Cannon, Model Predictive Control: Classical, Robust and Stochastic, Springer, 2016.
[20] K. Åström, B. Wittenmark, Problems of identification and control, J. Math. Anal. Appl. 34 (1) (1971) 90–113, http://dx.doi.org/10.1016/0022-247X(71)90161-2.
[21] B. Wittenmark, Adaptive dual control methods: an overview, 5th IFAC symposium on Adaptive Systems in Control and Signal Processing (1995) 67–72.
[22] V. Adetola, D. DeHaan, M. Guay, Adaptive model predictive control for constrained nonlinear systems, Syst. Control Lett. 58 (5) (2009) 320–326, http://dx.doi.org/10.1016/j.sysconle.2008.12.002.
[23] P. Swarnkar, S.K. Jain, R. Nema, Adaptive control schemes for improving the control system dynamics: a review, IETE Tech. Rev. 31 (1) (2014) 17–33, http://dx.doi.org/10.1080/02564602.2014.890838.
[24] A.A. Feldbaum, Dual control theory I, Autom. Remote Control 21 (1960) 874–880.
[25] K. Åström, A. Helmersson, Dual control of an integrator with unknown gain, Comput. Math. Appl. 12 (6) (1986) 653–662, http://dx.doi.org/10.1016/0898-1221(86)90052-0.
[26] N. Filatov, H. Unbehauen, Adaptive Dual Control: Theory and Applications, Lecture Notes in Control and Information Sciences, Springer Berlin Heidelberg, 2004.
[27] N.M. Filatov, H. Unbehauen, Survey of adaptive dual control methods, IEEE Proc. Control Theory Appl. 147 (1) (2000) 118–128, http://dx.doi.org/10.1049/ip-cta:20000107.
[28] T.A.N. Heirung, B.E. Ydstie, B. Foss, An MPC approach to dual control, IFAC Proc. Vol. 46 (32) (2013) 69–74, http://dx.doi.org/10.3182/20131218-3-IN-2045.00151.
[29] H. La, A. Potschka, J. Schlöder, H. Bock, Dual control and information gain in controlling uncertain processes, IFAC Symp. Dynam. Control Process Syst. Incl. Biosyst. 49 (7) (2016) 139–144, http://dx.doi.org/10.1016/j.ifacol.2016.07.230.
[30] D.P. Bertsekas, Dynamic Programming and Stochastic Control, Vol. 125 of Mathematics in Science and Engineering, Academic Press, New York, San Francisco, 1976.
[31] S. Thangavel, S. Lucia, R. Paulen, S. Engell, Towards dual robust nonlinear model predictive control: a multi-stage approach, in: American Control Conference, ACC 2015, Chicago, IL, USA, July 1–3, 2015, 2015, pp. 428–433, http://dx.doi.org/10.1109/ACC.2015.7170773.
[32] K.G. Hanssen, B. Foss, Scenario based implicit dual model predictive control, IFAC Proc. Vol. 48 (23) (2015) 416–421, http://dx.doi.org/10.1016/j.ifacol.2015.11.314.
[33] S. Thangavel, R. Paulen, S. Engell, S. Lucia, Robust nonlinear model predictive control with reduction of uncertainty via dual control, 21st International Conference on Process Control (PC) (2017) 48–53, http://dx.doi.org/10.1109/PC.2017.7976187.
[34] Y. Bar-Shalom, E. Tse, Concepts and methods in stochastic control, Control Dynam. Syst. Adv. Theory Appl. 12 (1976) 99–172.
[35] D.S. Bayard, M. Eslami, Implicit dual control for general stochastic systems, Optim. Control Appl. Methods 6 (3) (1985) 265–279, http://dx.doi.org/10.1002/oca.4660060307.
[36] J.H. Lee, N.L. Ricker, Extended Kalman filter based nonlinear model predictive control, Ind. Eng. Chem. Res. 33 (6) (1994) 1530–1541, http://dx.doi.org/10.1021/ie00030a013.
[37] C.V. Rao, J.B. Rawlings, Nonlinear Moving Horizon State Estimation, Birkhäuser Basel, Basel, 2000, pp. 45–69.
[38] S. Lucia, J.A. Andersson, H. Brandt, M. Diehl, S. Engell, Handling uncertainty in economic nonlinear model predictive control: a comparative case study, J. Process Control 24 (8) (2014) 1247–1259, http://dx.doi.org/10.1016/j.jprocont.2014.05.008.
[39] S. Lucia, S. Engell, Control of towing kites under uncertainty using robust economic nonlinear model predictive control, European Control Conference (ECC) (2014) 1158–1163, http://dx.doi.org/10.1109/ECC.2014.6862335.
[40] S. Lucia, R. Paulen, S. Engell, Multi-stage nonlinear model predictive control with verified robust constraint satisfaction, Proceedings of the 53rd IEEE Conference on Decision and Control (2014) 2816–2821, http://dx.doi.org/10.1109/CDC.2014.7039821.
[41] D.M. de la Penad, A. Bemporad, T. Alamo, Stochastic programming applied to model predictive control, Proceedings of the 44th IEEE Conference on Decision and Control (2005) 1361–1366, http://dx.doi.org/10.1109/CDC.2005.1582348.
[42] D. Bernardini, A. Bemporad, Scenario-based model predictive control of stochastic constrained linear systems, Proceedings of the 48h IEEE Conference on Decision and Control (CDC) held jointly with the 28th Chinese Control Conference (2009) 6333–6338, http://dx.doi.org/10.1109/CDC.2009.5399917.
[43] S. Lucia, Robust multi-stage nonlinear model predictive control (Dr.-Ing. Dissertation), Shaker Verlag, TU Dortmund, 2015.
[44] D. Mayne, J. Rawlings, C. Rao, P. Scokaert, Constrained model predictive control: stability and optimality, Automatica 36 (6) (2000) 789–814, http://dx.doi.org/10.1016/S0005-1098(99)00214-9.
[45] G. Franceschini, S. Macchietto, Model-based design of experiments for parameter precision: state of the art, Chem. Eng. Sci. 63 (19) (2008) 4846–4872, http://dx.doi.org/10.1016/j.ces.2007.11.034.
[46] S. Lucia, R. Paulen, Robust nonlinear model predictive control with reduction of uncertainty via robust optimal experiment design, IFAC Proc. Vol. 47 (3) (2014) 1904–1909, http://dx.doi.org/10.3182/20140824-6-ZA-1003.02332.
[47] E. Bradford, L. Imsland, Economic stochastic model predictive control using the unscented Kalman filter, in: 10th IFAC International Symposium on Advanced Control of Chemical Processes, Shenyang, Liaoning, China, July 25–27, 2018, 2018.
[48] M. Milanese, A. Vicino, Optimal estimation theory for dynamic systems with set membership uncertainty: an overview, Automatica 27 (6) (1991) 997–1009.
[49] E.-W. Bai, R. Tempo, H. Cho, Membership set estimators: size, optimal inputs, complexity and relations with least squares, IEEE Trans. Circuits Syst. I: Fundam. Theory Appl. 42 (5) (1995) 266–277.
[50] A.R. Gottu Mukkula, R. Paulen, Model-based design of optimal experiments for nonlinear systems in the context of guaranteed parameter estimation, Comput. Chem. Eng. 99 (2017) 198–213.
[51] S. Thangavel, M. Aboelnour, S. Lucia, R. Paulen, S. Engell, Robust dual multi-stage NMPC using guaranteed parameter estimation, in: Preprints of the 6th IFAC Conference on Nonlinear Model Predictive Control, IFAC, Madison, Wisconsin, USA, 2018, pp. 74–79.
[52] J. Andersson, J. Åkesson, M. Diehl, CasADi: a symbolic package for automatic differentiation and optimal control, in: Recent Advances in Algorithmic Differentiation, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 297–307.
[53] H. Pirnay, R. López-Negrete, L.T. Biegler, Optimal sensitivity based on ipopt, Math. Program. Comput. 4 (4) (2012) 307–331, http://dx.doi.org/10.1007/s12532-012-0043-2.
[54] L. Ljung (Ed.), System Identification: Theory for the User, second ed., Prentice Hall PTR, Upper Saddle River, NJ, USA, 1999.
[55] A. Wächter, T.L. Biegler, On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming, Math. Program. 106 (1) (2006) 25–57, http://dx.doi.org/10.1007/s10107-004-0559-y.
[56] R. Paulen, M.E. Villanueva, B. Chachuat, Guaranteed parameter estimation of non-linear dynamic systems using high-order bounding techniques with domain and CPU-time reduction strategies, IMA J. Math. Control Inf. 33 (3) (2016) 563–587, http://dx.doi.org/10.1093/imamci/dnu055.

Received: 15 June 2019 | Revised: 7 February 2021 | Accepted: 19 February 2021

DOI: 10.1002/rnc.5486

RESEARCH ARTICLE

WILEY

Tube-enhanced multi-stage model predictive control for flexible robust control of constrained linear systems with additive and parametric uncertainties

Sankaranarayanan Subramanian[1] | **Sergio Lucia**[2] | **Radoslav Paulen**[3] | **Sebastian Engell**[1]

[1]Process Dynamics and Operations Group, TU Dortmund, Dortmund, Germany

[2]Laboratory of Process Automation Systems, TU Dortmund, Dortmund, Germany

[3]Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, Bratislava, Slovakia

Correspondence
Sankaranarayanan Subramanian, Process Dynamics and Operations Group, TU Dortmund, Dortmund, Germany.
Email: sankaranarayanan.subramanian@tu-dortmund.de

Funding information
Agentúra na Podporu Výskumu a Vývoja, Grant/Award Number: APVV 15-0007; European Commission, Grant/Award Number: 291458 (MOBOCON)

Abstract

The trade-off between optimality and complexity has been one of the most important challenges in the field of robust model predictive control (MPC). To address the challenge, we propose a flexible robust MPC scheme by synergizing the multi-stage and tube-based MPC approaches. The key idea is to exploit the nonconservatism of the multi-stage MPC and the simplicity of the tube-based MPC. The proposed scheme provides two options for the user to determine the trade-off depending on the application: the choice of the robust horizon and the classification of the uncertainties. Beyond the robust horizon, the branching of the scenario-tree employed in multi-stage MPC is avoided with the help of tubes. The growth of the problem size with respect to the number of uncertainties is reduced by handling *small* uncertainties via an invariant tube that can be computed offline. This results in linear growth of the problem size beyond the robust horizon and no growth of the problem size concerning small magnitude uncertainties. The proposed approach helps to achieve a desired trade-off between optimality and complexity compared to existing robust MPC approaches. We show that the proposed approach is robustly asymptotically stable. Its advantages are demonstrated for a CSTR example.

KEYWORDS

constrained systems, predictive control, robust control, stability, uncertain systems

1 | INTRODUCTION

Robust model predictive control (MPC) schemes address the presence of uncertainties in the model with the goal to achieve constraint satisfaction and closed-loop stability. It is desirable that the robust MPC schemes are computationally cheap and nonconservative. The dual goal of nonconservatism and low complexity is a key challenge that is being actively researched in the field of robust MPC, and often a trade-off is needed. Open-loop min-max MPC was one of the earliest robust MPC schemes proposed.[1] In this approach, the worst-case cost is minimized while satisfying the constraints for all realizations of the uncertainty. The scheme, however, does not account for the presence of feedback in the predictions

and predicts a single control input at every stage. Because of the lack of recourse, the robustness comes at the cost of a significant loss of performance.

Feedback min-max MPC models the presence of feedback information explicitly in the predictions and thus reduces the conservatism of the open-loop schemes.[2] A general feedback min-max MPC optimizes the worst-case value of the cost function over a sequence of control policies, leading to infinite-dimensional optimization problems. One possibility to formulate a feedback MPC method with a finite-dimensional optimization problem is to consider a tree-structure to represent the evolution of the uncertainty[3] because, for each predicted state at every stage, the possibility is considered to adapt the inputs in the predictions. The tree structure grows exponentially with respect to the length of the prediction horizon, making the approach inapplicable in practice for long prediction horizons.

Other related robust MPC approaches optimize the expected value of the cost function[4] or a weighted sum of all the predicted scenarios, as done in multi-stage MPC.[5] The weights of the multi-stage MPC are tuning parameters that provide additional degrees of freedom to improve the closed-loop performance compared to a feedback min-max MPC scheme.

An alternative to the representation of feedback via a scenario tree consists of restricting the optimization to control policies with a fixed structure, linear policies,[6] or affine policies.[7-10] Tube-based MPC is one of the most discussed robust MPC approaches in the literature that usually considers an affine parameterization of the feedback policies.[10,11] It was shown in Reference 11 that the problem size can be kept the same as that of nominal MPC if the feedback gain is chosen offline and kept constant in the predictions. However, this comes at the cost of performance loss. Tube-based MPC approaches that relax the structure of the control policy or that predict the tube online (as opposed to an invariant tube) can improve the performance as shown in References 12-15. The performance advantages come at the cost of an increase in computational complexity with respect to the length of the prediction horizon. To handle parametric uncertainties, tube-based MPC based on Farkas' lemma was proposed.[16,17] The complexity of the approach grows linearly with respect to the length of the prediction horizon. Advanced tube-based schemes such as in References 13,14,16, and 17 use contractive sets for the prediction of tubes online. The number of inequalities and the number of vertices that characterize the tube can increase rapidly with respect to the dimension of the states and this makes the approach difficult to implement for high dimensional systems. If low complexity tubes are employed as proposed in References 18 and 19, the schemes can be highly conservative.

The aim of this article is to propose a novel scheme to achieve the dual goal of low computational cost and low conservatism. To achieve this goal, we combined the multi-stage and the tube-based MPC approaches by the classification of uncertainties in Reference 20. The multi-stage MPC is employed to handle *significant* uncertainties, and the tube-based MPC is used to handle *small* magnitude disturbances. We extend the scheme proposed in Reference 20 in this contribution in such a way that the rapid increase in problem complexity in multi-stage MPC is addressed both with respect to the number of uncertainties and with respect to the length of the prediction horizon. The proposed scheme gives the user two options that determine the trade-off depending on the requirements of an application. The two options are the choice of the robust horizon in multi-stage MPC and the classification of the uncertainties. The key aspects of the proposed approach are as follows:

1. The branching of the scenario tree is stopped beyond a certain prediction step called robust horizon. An affine feedback policy is employed beyond the robust horizon with the help of tubes to achieve robust constraint satisfaction and recursive feasibility guarantees.
2. Different scenarios are predicted in the proposed framework for any choice of robust horizon greater than or equal to 1. The resulting increase in the number of degrees of freedom enables the employment of low complexity tubes for high dimensional systems without a significant loss of performance when compared to standard tube-based schemes.
3. In addition, the growth of the problem size with respect to the number of uncertainties is reduced by formulating an invariant tube for small disturbances by making use of the ideas proposed in Reference 20.

We investigate in detail the theoretical properties of the proposed approach and demonstrate that the proposed approach is robustly asymptotically stable. We present the resulting tube-enhanced multi-stage MPC scheme as a convex optimization problem that is solved at every time step. This is achieved by employing the tube-based formulations from References 13 and 16 and the multi-stage formulation from Reference 20. The advantages of the scheme are demonstrated for a CSTR example.

2 | PRELIMINARIES

We study discrete-time linear dynamical systems of the form:

$$x^+ = Ax + Bu + w, \tag{1}$$

where $x \in \mathbb{R}^{n_x}$ represents the state, $u \in \mathbb{R}^{n_u}$ represents the input, $w \in \mathsf{W} \subset \mathbb{R}^{n_x}$ denotes additive disturbances, the matrix $A \in \mathbb{R}^{n_x \times n_x}$ represents the uncertain system matrix, and $B \in \mathbb{R}^{n_x \times n_u}$ denotes the uncertain input matrix of the controlled system. The system matrix A and the input matrix B are contained in a convex polytope and can be represented as $(A, B) \in \text{conv}(\{(A_i,\ B_i),\ \forall i \in \Gamma_p\})$, where conv() denotes the convex-hull operator and $\Gamma_p := \{1,\ \dots,\ n_p\}$. We assume that there exists a feedback gain K that is stabilizing for all $(A, B) \in \text{conv}(\{(A_i,\ B_i),\ \forall i \in \Gamma_p\})$. W is assumed to be a convex polytope with the origin in its interior and is characterized by n_w vertices. The bounds of the additive disturbances can be defined in terms of vertices of the set as $\mathsf{W} := \{w | w \in \text{conv}(\{w_l,\ \forall l \in \Gamma_w\})\}$, where $\Gamma_w = \{1,\ \dots,\ n_w\}$. The following definitions of invariant sets adapted from References 21 and 22 will be used throughout this article.

Definition 1. A set S is said to be robust positively invariant (RPI) for the system $x^+ = (A_i + B_i K)x + w$, if $\forall x \in \mathsf{S}$, $x^+ \in \mathsf{S},\ \forall w \in \mathsf{W}, \forall i \in \Gamma_p$.

Definition 2. A set S_{min} is said to be the minimal robust positively invariant set (mRPI) if S_{min} is contained in every closed robust positively invariant set.

Definition 3. A set S_{max} is said to be the maximal robust positively invariant set (MRPI) if S_{max} contains every closed robust positively invariant set.

2.1 | Multi-stage MPC

The robustness of multi-stage MPC is achieved by modeling the future evolution of the system by a scenario tree as shown in Figure 1.

Each branch of the tree denotes a realization of the uncertainties. Each node denotes a predicted state at the corresponding point in time. If all the extreme values of the uncertainties are realized in the predictions, the predicted states form the convex hull of all the possible trajectories in the future until the end of the prediction horizon. Realizations of the uncertainties that are not extreme can also be included to improve the resulting closed-loop performance. The tree branches until the end of the prediction horizon for each predicted node. The availability of feedback information in the predictions is explicitly modeled in the tree structure without restricting the structure of the feedback policy. This makes

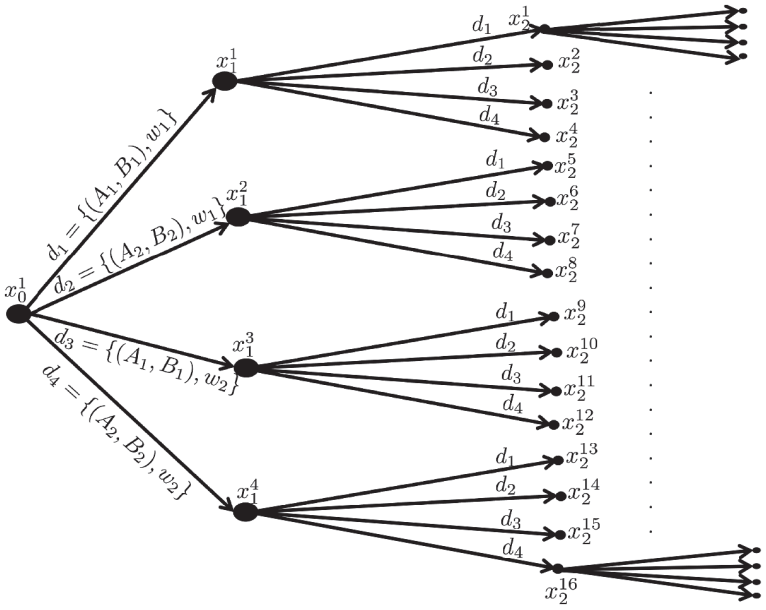


FIGURE 1 Scenario tree representation of the uncertainty evolution for multi-stage MPC for the system with additive and parametric uncertainties. The vertex matrices are given by $\{(A_1, B_1), (A_2, B_2)\}$ and the vertices of the additive disturbances are given by w_1 and w_2

the approach less conservative than those which assume linear or affine feedback policies, but exponentially complex.[23] The optimization problem that is solved at every time step is given as follows:

$$\min_{u_k^j, \forall (j,k) \in I_{[\![0,\ N_p-1]\!]}} \quad \sum_{k=0}^{N_p-1} \sum_{j=1}^{n_d{}^k} \omega_k^j \ell(x_k^j, u_k^j) + \sum_{j=1}^{n_d{}^{N_p}} V_f(x_{N_p}^j) \tag{2a}$$

subject to:

$$x_{k+1}^c = A_i x_k^j + B_i u_k^j + w_l, \qquad \forall\ (j,k) \in I_{[\![0,\ N_p-1]\!]},\ \forall i \in \Gamma_p,\ \forall l \in \Gamma_w, \tag{2b}$$

$$x_k^j \in \mathbb{X},\ u_k^j \in \mathbb{U}, \qquad \forall\ (j,k) \in I_{[\![0,\ N_p-1]\!]}, \tag{2c}$$

$$x_{N_p}^j \in \mathbb{X}_f, \qquad \forall\ (j,N_p) \in I_{N_p}, \tag{2d}$$

where the set of all indices (j,k) in the scenario tree is denoted as I and the set of indices occurring from a stage k_1 until a certain stage k_2 is denoted by $I_{[\![k_1,\ k_2]\!]}$, where $0 \le k_1 \le N_p$ and $k_1 \le k_2 \le N_p$. Also, the set of indices occurring at a stage at k is denoted by $I_k \triangleq I_{[\![k,\ k]\!]}$, where $0 \le k \le N_p$. $I_{[\![k_1,\ k_2]\!]} \triangleq \emptyset$, if $k_2 < k_1$. The number of branches at every predicted node is given by $n_d = n_p \times n_w$. The uncertainty realization corresponding to each branch of the scenario tree is denoted by $d_r, r \in \{1, \ \dots\ , n_d\}$, where d_r represents a particular combined realization of the parametric uncertainties (A_i, B_i), and of the additive disturbances w_l. That is, $d_r = \{(A_i, B_i), w_l\}, r = (l-1)n_w + i,\ i \in \Gamma_p,\ l \in \Gamma_w$. Each state x_{k+1}^c predicted at the time step $k+1$ is the child node in the scenario tree obtained from the parent node x_k^j, the input u_k^j, and the realization of the uncertainty d_r. The mapping of the index of a child node from the parent node across each prediction step is given by the relationship $c = n_d(j-1) + r$. Each node in the scenario tree has a nonnegative weight associated with it $\omega_k^j, \forall (j,k) \in I$. The weighted sum of the stage costs $\ell(x,u)$ along the prediction horizon and the terminal penalty function $V_f(x)$ constitute the overall objective function.

The state and the input bounds, and the bounds on the terminal state are enforced via (2c) using polytopic sets $\mathbb{X}$ and $\mathbb{U}$, and $\mathbb{X}_f$, respectively.

The control input for a particular node must be the same for all the branches to enforce the causality of the control policy, that is, $u_k^j = u_k^l$ if $x_k^j = x_k^l$ for all $(j,k), (l,k) \in I_{[\![0,\ N_p-1]\!]}$. However, the future inputs at different nodes can be different as measurement information will be available at the next stages, that is, u_k^j can be different from u_k^l if $x_k^j \ne x_k^l$ for all $(j,k), (l,k) \in I_{[\![0,\ N_p-1]\!]}$. The optimal input at the fist prediction step $u_0^{1*}(x)$ obtained by solving the optimization problem (2) is applied to the plant. The terminal region $\mathbb{X}_f$ is chosen as the maximal RPI set for a stabilizing control law $K_f x$. The problem size grows rapidly with respect to the number of uncertainties and the length of the prediction horizon. Therefore we investigate solutions of reduced complexity that approximately realize the performance of the multi-stage scheme. This will be discussed in detail in the rest of this article.

In the linear case considered here, since all the extreme realizations of the uncertainties are used in the predictions, the scenario tree predicts the reachable set of state trajectories. Every node of the scenario tree denotes the vertices of the polytope that forms the reachable set of the system for the predicted control policy.

3 | TUBE-ENHANCED MULTI-STAGE MPC

The problem defined in (2) suffers from rapid growth with respect to the number of realizations of the uncertainties and the length of the prediction horizon. We propose to employ two kinds of tubes to deal with the growth in problem complexity as described below:

1. An invariant tube using an affine feedback policy is employed to handle small-magnitude disturbances.
   - The invariant tube is obtained offline and hence, the complexity of the optimization problem does not grow with respect to the number of small disturbances.
   - The invariant tube is employed only for small disturbances and hence the method does not introduce a large conservatism.

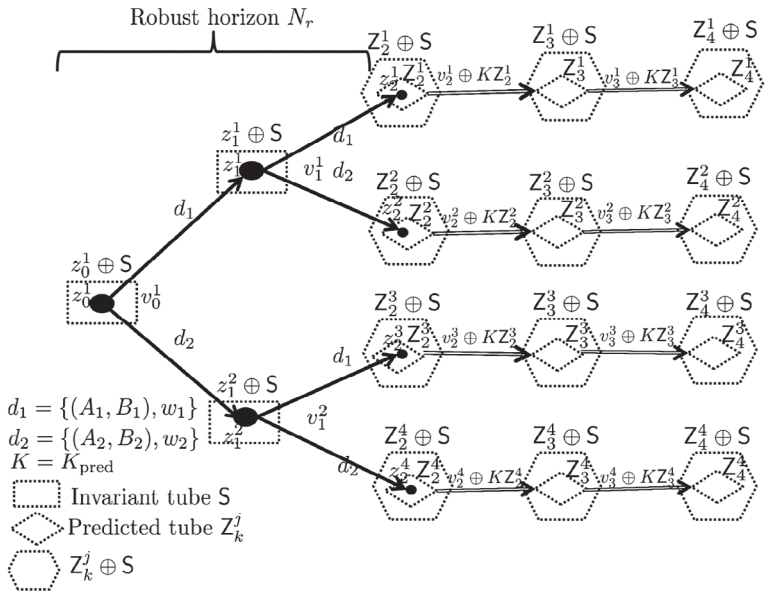


FIGURE 2 Scenario tree representation of the evolution of the uncertainties for tube-enhanced multi-stage MPC with robust horizon $N_r = 2$

2. Different tubes for each scenario are introduced to handle the significant uncertainties after a predefined horizon (robust horizon) in the prediction, instead of further branching of the scenario tree (see Figure 2).

- The problem complexity grows only linearly with respect to the prediction horizon beyond the robust horizon.
- The formulation optimizes for different feed-forward terms of the predicted tubes at every stage that belong to different scenarios beyond the robust horizon (in addition to modeling full recourse until the robust horizon) and hence the approach is less conservative when compared to a pure tube-based scheme. Low complexity tubes can also be employed for less conservatism making the approach applicable to high dimensional systems.

The following subsections will elaborate the key points discussed above to obtain an improved trade-off between optimality and complexity.

3.1 | Handling small disturbances using an invariant tube

Because of the multiplicative nature of the parametric uncertainties, they will have a large influence on the state trajectory far away from the origin. Therefore, we classify all parametric uncertainties as large magnitude uncertainties. The additive disturbances can be large or small depending on the application. To reduce the computational complexity, the set of additive disturbances W is decomposed into two polytopes $\overline{\mathsf{W}}$ and $\underline{\mathsf{W}}$ that contain the origin in their interiors such that $\mathsf{W} \subseteq \overline{\mathsf{W}} \oplus \underline{\mathsf{W}}$, where $\overline{\mathsf{W}}, \ \underline{\mathsf{W}} \subset \mathbb{R}^{n_x}$.

The set $\overline{\mathsf{W}}$ denotes large magnitude disturbances and the set $\underline{\mathsf{W}}$ denotes small magnitude uncertainties. It is recommended that the uncertain set W is decomposed such that the number of vertices of the large uncertainties $\overline{\mathsf{W}}$ is small while $\underline{\mathsf{W}}$ is of smaller volume. The model that accounts for large uncertainties is defined as follows:

$$z^+ = A_i z + B_i v + w, \tag{3}$$

for all $i \in \Gamma_p$ and for all $w \in \overline{\mathsf{W}}$, where $z \in \mathbb{R}^{n_x}$ is the state and $v \in \mathbb{R}^{n_u}$ is the input of the model (3). The number of vertices of the additive uncertainty set $\overline{\mathsf{W}}$ is denoted by $n_{\overline{w}}$ and the set $\overline{\mathsf{W}}$ is defined as $\overline{\mathsf{W}} := \{w | w \in \text{conv}(\{w_l, \ \forall l \in \Gamma_{\overline{w}}\})\}$, where $\Gamma_{\overline{w}} = \{1, \ \dots, \ n_{\overline{w}}\}$. The large uncertainties are considered in the predictions using the multi-stage approach. To handle small disturbance set $\underline{\mathsf{W}}$, an affine feedback policy is employed as follows:

$$u = v + K_{\text{inv}}(x - z), \tag{4}$$

where K_{inv} denotes the feedback gain associated with the invariant tube and is chosen such that the parameter-varying closed-loop system $\{(A_i + B_i K_{\text{inv}}),\ \forall i \in \Gamma_p\}$ is asymptotically stable. For the system given in (1), the state x and the control law (4), a set S is defined as small disturbance invariant if $(A_i + B_i K_{\text{inv}})\mathsf{S} \oplus \underline{\mathsf{W}} \subseteq \mathsf{S}$ for all $i \in \Gamma_p$. This disturbance invariant set S can be chosen as the convex RPI over-approximation of the minimal RPI set of the model (3). A convex outer-approximation (S) of the minimal RPI set S can be obtained from the algorithm given in Reference 24. This problem is solved offline and hence does not affect the online computation time of the proposed algorithm. In the implementation section, we propose a novel convex optimization problem to over-approximate the minimal RPI set that is disturbance invariant using a linear programming problem. Since the small magnitude uncertainties are handled using an affine feedback policy, they will not be considered in the online optimization problem and hence do not affect the problem complexity.

3.2 | Handling long prediction horizons using predicted tubes

For the large uncertainties considered in the scenario tree, the problem complexity increases exponentially with respect to the prediction horizon N_p. To reduce the problem complexity, the branching of the tree can be stopped beyond a certain prediction step called the robust horizon N_r. Beyond the robust horizon, the affine policies $(v + K_{\text{pred}} z)$ are employed to handle all large uncertainties considered in the scenario tree. Here K_{pred} denotes the feedback gain associated with the tubes predicted online. The dynamics of the system beyond N_r can be described using the following set recursion:

$$\mathsf{Z}^+ \supseteq (A_i + B_i K_{\text{pred}})\mathsf{Z} \oplus \{B_i v\} \oplus \overline{\mathsf{W}}, \tag{5}$$

for all $i \in \Gamma_p$ and Z denotes the tube of states at the current time step and Z^+ denotes the tube at the successor time step. The idea is illustrated in Figure 2 for a robust horizon $N_r = 2$. It can be seen that the tubes replace branches beyond N_r leading to a linear growth of the complexity of the scenario tree with respect to the prediction horizon. The feed-forward terms associated with the affine control law can be different for different scenarios. This can improve the performance of the controller when compared to the tube-based scheme,[16] where the problem formulation considers one feed-forward term per prediction step. The invariant tube is shown on top of the scenario tree for representational purposes. The invariant tube S is used to obtain a suitable back-off that guarantees satisfaction of the original constraints for all possible values of the small magnitude uncertainties $\underline{\mathsf{W}}$ and does not contribute additional costs to the online optimization problem.

Remark 1. The idea of robust horizon has been already proposed for nonlinear systems in Reference 5 by assuming that the uncertainty remains constant beyond the robust horizon. However, no rigorous study on the recursive feasibility and the stability of the closed-loop system was performed. In Reference 25, it was proposed that a terminal set must be employed at the end of robust horizon to achieve stability but this is clearly restrictive. In this work, we propose a rigorous solution using a robust horizon in the multi-stage MPC framework enhanced by a tube-based formulation. We make use of affine control policies beyond the robust horizon and formulate the problem such that the closed-loop system is stable. The proposed formulation is not as restrictive as in Reference 25, the terminal constraints are enforced at the end of the prediction horizon as in conventional MPC schemes and the proposed scheme does not assume that the uncertainties remain constant beyond N_r.

3.3 | Problem formulation

The optimization problem $\mathtt{P}_{N_p}(x)$ that is solved at every time step is given by

$$\min_{z_0^1, v_k^j \forall (j,k) \in I_{[\![0,\ N_p-1]\!]}} \left(\sum_{k=0}^{N_r-1} J_k^{\text{MS}} + \sum_{k=N_r}^{N_p-1} J_k^{\text{tube}} + J_{N_p}^{\text{term}} \right) \tag{6a}$$

subject to:

$$z_{k+1}^c = A_i z_k^j + B_i v_k^j + w_l, \qquad \forall\ (j,k) \in I_{[\![0,\ N_r-1]\!]}, \forall i \in \Gamma_P, \forall l \in \Gamma_{\overline{w}}, \tag{6b}$$

$$z_k^j \in \mathbb{Z},\ v_k^j \in \mathbb{V}, \qquad \forall (j,k) \in I_{[\![0,\ N_r-1]\!]}, \tag{6c}$$

$$z_{N_r}^j \in \mathsf{Z}_{N_r}^j \subseteq \mathbb{Z}, \qquad \forall (j,N_r) \in I_{N_r}, \tag{6d}$$

$$(A_i + B_i K_{\text{pred}})\mathsf{Z}_k^j \oplus \{B_i v_k^j\} \oplus \overline{\mathsf{W}} \subseteq \mathsf{Z}_{k+1}^j, \qquad \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \forall i \in \Gamma_p, \tag{6e}$$

$$\mathsf{Z}_k^j \subseteq \mathbb{Z}, \tilde{z}_k^j \in \mathsf{Z}_k^j,\ \{v_k^j\} \oplus K_{\text{pred}}\mathsf{Z}_k^j \subseteq \mathbb{V}, \qquad \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{6f}$$

$$\mathsf{Z}_{N_p}^j \subseteq \mathbb{Z}_f, \qquad \forall (j,N_p) \in I_{N_p}, \tag{6g}$$

$$x \in \{z_1^0\} \oplus \mathsf{S}, \tag{6h}$$

where

$$J_k^{\text{MS}} = \sum_{j=1}^{{n_d}^k} \omega_k^j \ell(z_k^j, v_k^j), \tag{6i}$$

$$J_k^{\text{tube}} = \max_{\tilde{z}_k^j,\ \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}} \sum_{j=1}^{{n_d}^{N_r}} \omega_k^j \ell(\tilde{z}_k^j, v_k^j + K_{\text{pred}}\tilde{z}_k^j), \tag{6j}$$

$$J_{N_p}^{\text{term}} = \max_{\tilde{z}_{N_p}^j,\ \forall (j,N_p) \in I_{N_p}} \sum_{j=1}^{{n_d}^{N_r}} V_f(\tilde{z}_{N_p}^j). \tag{6k}$$

There are several differences in both the objective and the constraints in (6) compared to that of the standard multi-stage MPC (2). In the optimization problem (6), the objective can be divided into three parts: a multi-stage part J_k^{MS}, a tube-based part J_k^{tube} and the terminal penalty part J_k^{term}. The multi-stage part of the objective function J_k^{MS} is applied until $N_r - 1$, and beyond $N_r - 1$, the tube-based part of the objective function J_k^{tube} is applied. As always, the terminal penalty J_k^{term} is applied at the last prediction step N_p. The multi-stage part of the objective function J_k^{MS} is the same as in (2). The tube-based part of the objective function J_k^{tube} and the terminal part of the objective function J_k^{term} have maximization terms associated with them. For the tube-based part of the optimization problem, the worst-case cost associated with the predicted tubes Z_k^j are obtained at every prediction step. This helps in establishing the optimal value function as a Lyapunov function for the proposed scheme.

The state inside the tube that maximizes the objective function is defined as $\tilde{z}_k^j$ and is constrained by $\tilde{z}_k^j \in \mathsf{Z}_k^j$ as given in (6f). The variables $\tilde{z}_k^j$ are formulated as decision variables for all $(j,k) \in I_{[\![N_r,\ N_p-1]\!]}$. Equations (6e)–(6g) represent the constraints beyond the robust horizon N_r. Equation (6e) guarantees the recursive bounding of the state trajectories for the chosen control law $v + K_{\text{pred}}z$. It can also be seen that the affine term v_k^j can be freely chosen for all $(j,k) \in I_{[\![0,\ N_p-1]\!]}$, that is, v_k^j can be different from v_k^l if $z_k^j \neq z_k^l$ for all $(j,k),(l,k) \in I_{[\![0,\ N_p-1]\!]}$. This can improve the resulting solution even for a robust horizon of $N_r = 1$ when compared to a pure tube-based scheme. Equations (6f)–(6g) denote the state, input, and the terminal constraints. The set of all $x \in \mathbb{X}$ for which there exists a feasible feedback policy is denoted as $\mathtt{X}_{N_p}$. Equation (6d) is formulated at the robust horizon to establish continuity between the predicted scenarios until the robust horizon and the tubes predicted beyond the robust horizon.

The initial state of the scenario tree z_0^1 is a function of the current state x as defined in (6h) and is a decision variable of the optimization problem (6). The optimization problem is solved with tightened constraints $\mathbb{Z} = \mathbb{X} \ominus \mathsf{S}$ and $\mathbb{V} = \mathbb{U} \ominus K_{\text{inv}}\mathsf{S}$ in (6c) and (6f). The number of branches at every node is given by $n_d = n_p \times n_{\overline{w}}$ until the robust horizon N_r. The uncertainty realization corresponding to each branch of the scenario tree for the proposed formulation is given by $d_r = \{(A_i, B_i), w_l\}, r = (l-1)n_{\overline{w}} + i,\ i \in \Gamma_p,\ l \in \Gamma_{\overline{w}}$.

Note that the number of branches can be reduced dramatically if $n_{\overline{w}} \ll n_w$ when compared to the consideration of all uncertainties in a scenario tree. The control input u applied to the system is given by $u = v_0^{1*} + K_{\text{inv}}(x - z_0^{1*}(x))$, where v_0^{1*} is the first element of the optimal control input sequence obtained by solving (6). Since the polytope $\overline{\mathsf{W}}$ has smaller number of vertices compared to W, the problem size solved using (6) is reduced. Despite the reduced complexity when compared to a standard multi-stage problem, the proposed controller can often achieve a performance comparable to that of multi-stage MPC for a fraction of the computational complexity by the choice of N_r and $\overline{\mathsf{W}}$.

Remark 2. The proposed scheme is flexible and includes options to further improve it which we do not analyze in this work. Two of the possible modifications are listed below:

1. The feedback gain is denoted as a constant K_{pred} for the predicted tubes. This is done only to simplify the presentation. The feedback gains can be different for different scenarios. The only necessary condition is that the gain must be stabilizing and can be chosen freely for different scenarios to improve the performance of the closed-loop.
2. The number of tubes at the robust horizon is formulated equal to the number of nodes predicted until that stage using the scenario tree in (6). The constraint (6d) represents the continuity equation. However, different nodes can be bundled together in one tube and the number of tubes can be smaller than the number of nodes predicted until N_r by modifying the continuity constraint (6d). This can help to reduce problem complexity further.

3.4 | Stabilizing objective function and choice of weights

Since a persisting disturbance $w \in \mathsf{W}$ is assumed, convergence to the origin cannot be established. Instead, as described in Reference 26, a robust positively invariant set T will be shown to be asymptotically stable using the proposed robust MPC scheme. To achieve stability, we assume that the terminal set, the proposed stage cost and the terminal penalty function satisfy the following properties:

1. The terminal set $\mathbb{Z}_f := \mathsf{T}$ is a robust positively invariant set for a stabilizing control law $K_f z$.
2. The stage cost $\ell(z, K_f z) = 0$ and the terminal penalty $V_f(z) = 0$, $\forall z \in \mathsf{T}$.

A stage cost with these properties as proposed in Reference 26 is given as follows:

$$\ell(z, v) = \min_{y \in \mathbb{Z}_f} ||Q(z - y)||_p + ||R(v - K_f z)||_p, \tag{7}$$

where Q and R are positive semi-definite matrices. The terminal penalty function can be simply set to zero, that is, $V_f(z) = 0,\ \forall z \in \mathbb{R}^{n_x}$. The choice of the stage cost is different from the nominal cost $\ell_{\text{nom}}(z, v) = ||Qz||_p + ||Rv||_p$ which is generally used in the case of nominal MPC. The nominal stage cost ℓ_{nom} penalizes the distance to the origin and the control effort depending on the choice of tuning matrices Q and R. The stage cost $\ell(z, v)$ penalizes the distance to the set $\mathbb{Z}_f$ and the deviations from the control law $K_f z$. The terminal gain K_f is chosen equal to the gain of the predicted tubes K_{pred}. The gain K_f can be chosen freely when there is no tube-based part of the scheme. This is further discussed in Section 4.4 and in Section 5.

Each uncertainty realization (branch) has a fixed and bounded positive weight $\overline{\omega}_r > 0$ associated with it for all $r \in \{1, \dots, n_d\}$. Appropriate weights can be chosen depending on the applications. However, to establish stability the weights must follow certain rules which are formalized in the following assumption. The weights associated with the particular realization of the uncertainty will be assigned to the nodes that result from them in the predictions. For example, if z_1^1 is realized because of (A_1, B_1) and w_1, the weight associated with the node z_1^1 will be defined as $\omega_1^1 = \overline{\omega}_1$ (the corresponding weight of $\{(A_1, B_1), w_1\}$). The requirement for the choice of weights associated with each node in the scenario tree is given in the following assumption.

Assumption 1. The weight of the root node ω_0^1 is smaller than or equal to the minimum of all the weights (i.e., $\omega_0^1 \leq \min\{\overline{\omega}_0, \overline{\omega}_1, \dots, \overline{\omega}_{n_d}\}$) and it must be positive $\omega_0^1 > 0$. The weights ω_k^j associated with the other nodes z_k^j are equal to the weights associated with the realization of the uncertainty from which they are obtained for all $(j, k) \in I_{[\![1,\ N_r - 1]\!]}$. The weights associated with tubes are chosen as $\omega_k^j = n_d^{k - N_r} \omega_{\text{tube}},\ \forall (j, k) \in I_{[\![N_r,\ N_p - 1]\!]}$ where $\omega_{\text{tube}} \geq \max\{\overline{\omega}_0, \overline{\omega}_1, \dots, \overline{\omega}_{n_d}\}$ and is bounded.

Remark 3. The choice of the weights is important in establishing the stability properties of the proposed approach. The weights affect the objective function and thus the value function. If the weights are chosen as per Assumption 1 stability can be proven, see Lemma 4. In the multi-stage part of the scheme, the weights are chosen the same in each stage for the same realization of the uncertainty. Then feasible values of the stage costs for the succeeding step can be obtained by the convex combination of the stage costs that are realized at the current time step. Since the branching stops beyond the

robust horizon, the weights must be updated to account for the receding horizon implementation of the MPC scheme. Hence the weights of the tube-based part of the scheme are employed as $\omega_k^j = n_d^{k-N_r}\omega_{\text{tube}},\ \forall (j,k) \in I_{\llbracket N_r,\ N_p-1 \rrbracket}$.

Assumption 2. The stage cost $\ell(z,v)$ and the terminal penalty $V_f(z)$ are convex and positive definite functions and satisfy the following relationships:

1. $\ell(z,v) \geq c'|z|_{\mathbb{Z}_f},\ \forall z \in \mathbb{Z} \setminus \mathbb{Z}_f$, $\ell(z, K_f z) = 0,\ \forall z \in \mathbb{Z}_f$, $V_f(z) = 0,\ \forall z \in \mathbb{Z}_f$, where $|z|_{\mathbb{Z}_f} := \min_{y \in \mathbb{Z}_f} \|z - y\|_p$ and c' is a positive constant.
2. $\max_{\tilde{z} \in \mathsf{Z}} \ell(\tilde{z}, v + K_{\text{pred}}\tilde{z}) \geq c''|\tilde{z}|_{\mathbb{Z}_f},\ \forall \mathsf{Z} \subseteq \mathbb{Z} \setminus \mathbb{Z}_f$, $\max_{\tilde{z} \in \mathsf{Z}} \ell(\tilde{z}, K_f\tilde{z}) = 0,\ \forall \mathsf{Z} \subseteq \mathbb{Z}_f$, $\max_{\tilde{z} \in \mathsf{Z}} V_f(\tilde{z}) = 0, \forall \mathsf{Z} \subseteq \mathbb{Z}_f$, where c'' is a positive constant.

4 | IMPLEMENTATION DETAILS

We employ the tube-based MPC approach based on Farkas' Lemma as proposed in Reference 16 to account for the disturbances after the robust horizon. The proposed approach works for any choice of robust horizon and in principle can be chosen equal to $N_r = 0$ in which case, the scheme reduces to the approach in Reference 16 if $\underline{\mathsf{W}} = \{0\}$.

The tightened state constraint set $\mathbb{Z}$ is defined as $\mathbb{Z} := \{z | Fz \leq \mathbf{1}\}$, where F is a $n_c \times n_x$ dimensional matrix, n_c denotes the number of state constraints, and $\mathbf{1}$ denotes a vector with all elements 1 of an appropriate dimension. The tightened input constraint set $\mathbb{V}$ is defined as $\mathbb{V} := \{v | Gv \leq \mathbf{1}\}$, where G is a $n_m \times n_u$ matrix and n_m denotes the number of input constraints. In the following, we discuss three types of tubes: general complexity tube, homothetic tube and low complexity tube. We show that how the proposed formulation (6) can be implemented as convex optimization problem. In addition, we also discuss the subtleties in the reformulation with respect to the proposed approach.

4.1 | General complexity tube

The complexity of the tubes that are employed beyond the robust horizon can be fixed and defined as

$$\mathsf{Z} := \{z | Tz \leq \tau\},$$

where $T \in \mathbb{R}^{n_r \times n_x}$ is fixed for all prediction steps and $\tau \in \mathbb{R}^{n_r}$ is a decision variable that is chosen online. Here n_r denotes the number of inequalities that describes the tube Z and it is typically larger than $2n_x$. The matrix T is chosen such that the set $\Lambda := \{z | Tz \leq \mathbf{1}\}$ is λ-contractive for $\lambda \in [\lambda_{\min}, 1)$, where $\lambda_{\min}$ is the joint spectral radius of the closed-loop uncertain system matrices for the chosen feedback gain K_{pred}. The details on algorithm for obtaining the largest λ-contractive set can be found in Reference 22 (pp. 171-184). One can also use off-the-shelf toolboxes such as Multi-Parametric Toolbox[27] to obtain λ-contractive sets.

Now, let us have a look at the problem formulation (6), where the Equations (6e)–(6f) are in the form of set operations. Farkas' lemma can be employed to convert them from set operations to a set of linear equalities and inequalities. This reformulation will help us formulate the optimization problem as a convex optimization problem that does not require any set operations online. Equation (6e) bounds the error dynamics recursively as $(A_i + B_i K_{\text{pred}})\mathsf{Z}_k^j \oplus \mathsf{W} \subseteq \mathsf{Z}_{k+1}^j$ for all $i \in \Gamma_p$. Let Z_k^j be represented as

$$\mathsf{Z}_k^j = \{z | Tz \leq \tau_k^j\}, \tag{8}$$

where the variables $\tau_k^j,\ \forall (j,k) \in I_{\llbracket N_r,\ N_p \rrbracket}$ can be formulated as decision variables in the optimization problem. Since an affine control law $v + K_{\text{pred}}z$ is employed beyond the robust horizon N_r, $z^+ \in \{(A_i + B_i K_{\text{pred}})z + B_i v\} \oplus \overline{\mathsf{W}}$ for a given $i \in \Gamma_p$ and the tube at the next prediction step that bounds all the trajectories for the predicted input and the arbitrary realizations of the uncertainties for all $(j,k) \in I_{\llbracket N_r,\ N_p-1 \rrbracket}$ is given by

$$\mathsf{Z}_{k+1}^j := \{z | T((A_i + B_i K_{\text{pred}})z + B_i v_k^j + w_l) \leq \tau_{k+1}^j,\ \forall i \in \Gamma_p\ \forall l \in \Gamma_{\overline{w}},\ \forall z \in \mathsf{Z}_k^j\}. \tag{9}$$

To achieve the set recursion, Farkas' lemma is employed and it is given below:[16,22]

Lemma 1. *Given two nonempty sets* $\mathsf{X}_1 := \{x|T_1x \leq \tau_1\}$, $\mathsf{X}_2 := \{x|T_2x \leq \tau_2\}$, $\mathsf{X}_1 \subseteq \mathsf{X}_2$ *holds iff there exists a nonnegative matrix P that satisfies the equality* $PT_1 = T_2$ *and the inequality* $P\tau_1 \leq \tau_2$.

Using Lemma 1, we can employ nonnegative matrices $P_i,\ \forall i \in \Gamma_p$, and $(A_i + B_iK_{\text{pred}})\mathsf{Z}_k^j \oplus \overline{\mathsf{W}} \subseteq \mathsf{Z}_{k+1}^j$ holds for all $i \in \Gamma_p$ iff:

$$P_iT = T(A_i + B_iK_{\text{pred}}), \qquad \forall i \in \Gamma_p, \tag{10a}$$

$$P_i\tau_k^j + TB_iv_k^j + Tw_l \leq \tau_{k+1}^j, \qquad \forall i \in \Gamma_p,\ \forall l \in \Gamma_{\overline{w}},\ (j,k) \in I_{[\![N_r,\ N_p-1]\!]}. \tag{10b}$$

The set (9) is reformulated using linear equalities and inequalities as shown in (10). Similarly, the set operations (6f)–(6g) can be reformulated using Lemma 1. The state and input constraints can be formulated using nonnegative matrices P_x, P_u as follows:

$$P_xT = F, \tag{11a}$$

$$P_x\tau_k^j \leq \mathbf{1}, \qquad \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{11b}$$

$$P_uT = GK_{\text{pred}}, \tag{11c}$$

$$Gv_k^j + P_u\tau_k^j \leq \mathbf{1}, \qquad \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{11d}$$

The choice of nonnegative matrices $P_i,\ \forall i \in \Gamma_p,\ P_x,\ P_u$ if obtained online, results in a nonconvex optimization problem. Hence as proposed in Reference 16, the nonnegative matrices can be obtained offline such that the equality constraints are satisfied. This reduces the computational load of the online problem and convexifies it. The matrices $P_i,\ \forall i \in \Gamma_p$ can be obtained by solving the following problem for a given i:

$$P_i = \arg\min_{\tilde{P}_i} ||\tilde{P}_i||_\infty, \text{s.t.}\tilde{P}_iT = T(A_i + B_iK_{\text{pred}}). \tag{12}$$

Similar to (12), the matrices P_x, P_u can be obtained by formulating linear programming problems that satisfy the equality constraints (11a), and (11c). The terminal constraints are implemented as follows:

$$P_i\tau_{N_p}^j + Tw_l \leq \tau_{N_p}^j, \quad \forall i \in \Gamma_p,\ \forall l \in \Gamma_{\overline{w}}, \forall (j,N_p) \in I_{N_p}, \tag{13}$$

where the terminal set is defined as

$$\mathbb{Z}_f \triangleq \{z|Tz \leq \tau, P_iT = T(A_i + B_iK_{\text{pred}}), P_i\tau + Tw_l \leq \tau, \forall i \in \Gamma_p,\ \forall l \in \Gamma_{\overline{w}}\}, \tag{14}$$

and $\mathbb{Z}_f \subseteq \mathbb{Z}$ and $K_f\mathbb{Z}_f \subseteq \mathbb{V}, K_f = K_{\text{pred}}$ hold. The terminal gain K_f is chosen as the gain of the predicted gain to simplify the implementation and the discussion that follows. Further discussions on the computation of the terminal set can be found in Section 4.4. The minimal RPI set can be obtained using the methods discussed in Reference 24. Another possible over-approximation of the mRPI set can be obtained as $\mathsf{S} := \{z|\hat{T}z \leq \tau\}$. Here the matrix $\hat{T}$ is chosen such that $\{z|\hat{T}z \leq \mathbf{1}\}$ is λ-contractive for the chosen feedback gain K_{inv}. The right hand side of the inequality that defines S, τ can be obtained as follows:

$$\min_{\tau} ||\tau||_p \tag{15a}$$

subject to:

$$\hat{P}_i\tau + \hat{T}w_l \leq \tau, \forall i \in \Gamma_p,\ \forall l \in \Gamma_{\overline{w}}. \tag{15b}$$

The objective can be chosen as 1-norm or ∞-norm so that the optimization problem is an LP. The LP (15) guarantees that the set $S := \{z | \hat{T}z \leq \tau\}$ is robust positively invariant. The nonnegative matrices $\hat{P}_i$, for all $i \in \Gamma_p$ can be chosen as discussed in (11). Since $\hat{T}$ and $\hat{P}_i$, $\forall i \in \Gamma_p$ are fixed, the set S can be a conservative over-approximation. The advantage however is that the LP (15) can be solved much faster compared to the algorithm given in Reference 24.

In addition to the reformulation of the set operations in the constraints, the objective function can be simplified by removing the maximization part with the help of slack variables as shown in References 3 and 26. If the stage cost (7) is used with $p = 1$, $\min_{v \in \mathbb{V}} \max_{\tilde{z}} \ell(\tilde{z}, v + K_{\text{pred}}\tilde{z})$ can be obtained as follows:

$$\min_{v \in \mathbb{V}} \max_{\tilde{z} \in \mathsf{Z}} \ell(\tilde{z}, v + K_{\text{pred}}\tilde{z}) = \min_{v, \tilde{z}, y, \mu, \eta, \gamma} \gamma \tag{16a}$$

subject to:

$$-\mu \leq Q(\tilde{z} - y) \leq \mu, \ \forall \tilde{z} \in \mathsf{Z}, y \in \mathbb{Z}_f, \tag{16b}$$

$$-\eta \leq Rv + K_{\text{pred}}\tilde{z} - K_f\tilde{z} \leq \eta, \ \forall \tilde{z} \in \mathsf{Z}, \tag{16c}$$

$$\mathbf{1}^T\mu + \mathbf{1}^T\eta \leq \gamma. \tag{16d}$$

Note that $\tilde{z}$ is not known a priori and has been added as a decision variable in (16). The constraint (16c) simplifies to $-\eta \leq Rv \leq \eta$ because the terminal gain K_f is chosen equal to the gain employed in the predicted tubes (i.e., $K_f = K_{\text{pred}}$). The constraints of the state are infinite dimensional. However, a simplification is possible by reformulating the bounds on the state objective (16b) with the help of a nonnegative matrix P_Q as discussed earlier and the constraint (16b) can be satisfied for any given $\tilde{z} \in \mathsf{Z}$ if all $z \in \mathsf{Z}$ satisfy (16b) as proposed in Reference 28. Using Lemma 1, we can define a nonnegative matrix P_Q such that the following equations hold to satisfy the constraints (16b) and the constraints (16b)–(16d) can be rewritten as follows for all $(j,k) \in I_{[\![N_r,\ N_p-1]\!]}$.:

$$P_Q T = Q, \tag{17a}$$

$$-\mu_k^j \leq P_Q\tau_k^j - Qy_k^j \leq \mu_k^j, \qquad y_k^j \in \mathbb{Z}_f, \ \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{17b}$$

$$-\eta_k^j \leq Rv_k^j \leq \eta_k^j, \qquad \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{17c}$$

$$\mathbf{1}^T\mu_k^j + \mathbf{1}^T\eta_k^j \leq \gamma_k^j, \qquad \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}. \tag{17d}$$

The equality constraint (17a) can be obtained offline and the inequality constraints (17b) and (17c) can be included as the constraints in the optimization problem online. However, this reformulation of the inequality constraints does not provide a tight upper bound which means that the obtained cost is not the same as the one obtained using the original inner maximization in (36a). The formulation, however, retains the theoretical properties of recursive feasibility and stability of the original formulation (6). Combining all the reformulations, the resulting optimization problem $\mathtt{P}_{N_p}^G(x)$ can be formulated as

$$\min_{z_0^1 v_k^j y_k^j \ jkI_{0\ N_p1}\rrbracket \tau_k^j, \gamma_k^j, \eta_k^j, \mu_k^j, \ \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}} \sum_{k=0}^{N_r-1} \sum_{j=1}^{n_d^k} \omega_k^j \ell(z_k^j, v_k^j) + \sum_{k=N_r}^{N_p-1} \sum_{j=1}^{n_d^{N_r}} \omega_k^j \gamma_k^j \tag{18}$$

subject to:

(6b), (6c), (6h), (10b), (11b), (11d), (13), (17b), (17c), (17d), $Tz_{N_r}^j \leq \tau_{N_r}^j, \ \forall (j, N_r) \in I_{N_r}$.

In (18), it can be seen that the set operations are replaced by inequality constraints with the help of the nonnegative matrices $P_i, \forall i \in \Gamma_p$, P_x, P_u, P_Q that are obtained offline. The constraints (17b)–(17d) bound the objective from above using the slack variables $\gamma_k^j \in I_{[\![N_r,\ N_p-1]\!]}$. The set of all $x \in \mathbb{X}$ for which there exists a feasible solution for the optimization problem (18) is denoted as $\mathtt{X}_{N_p}^G$. The optimal value function obtained by solving the optimization problem (18) is denoted as $V_{N_p}^{G*}(x)$. The stability properties of the implementation (18) is proven in Section 5.

4.2 | Homothetic tube

Since the tubes are characterized using inequalities in the general complexity tube, explicit values of the vertices are not available online. Hence, the tight upper bound for the stage costs of the tubes are difficult to obtain. By formulating the predicted tubes as homothetic tubes (i.e., by fixing the shape of the sets and only varying the scaling variable), the characterizations of the vertices can be obtained as follows:

$$\mathcal{Z}_k^j = \{z | T(z - \hat{z}_k^j) \leq \alpha_k^j \mathbf{1}\}, \tag{19a}$$

$$= \hat{z}_k^j \oplus \alpha_k^j \Lambda, \tag{19b}$$

$$= \hat{z}_k^j \oplus \alpha_k^j \text{conv}\{\vee_1, \vee_2, \ \dots\ , \vee_{n_v}\}, \tag{19c}$$

where $\alpha_k^j \geq 0$ is a nonnegative scalar and the λ-contractive set Λ is characterized by the vertices $\vee_1, \vee_2, \ \dots\ , \vee_{n_v}$, where n_v denotes the number of vertices. The constraints for the homothetic tube can be rewritten as follows using Lemma 1.

$$P_i(\alpha_k^j \mathbf{1} + T\hat{z}_k^j) + TB_i v_k^j + Tw_l \leq T\hat{z}_{k+1}^j + \alpha_{k+1}^j \mathbf{1}, \qquad \forall i \in \Gamma_p, \ \forall l \in \Gamma_{\overline{w}}, \ (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{20a}$$

$$P_x(\alpha_k^j \mathbf{1} + T\hat{z}_k^j) \leq \mathbf{1}, \qquad \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{20b}$$

$$Gv_k^j + P_u(\alpha_k^j \mathbf{1} + T\hat{z}_k^j) \leq \mathbf{1}, \qquad \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{20c}$$

$$P_i(\alpha_{N_p}^j \mathbf{1} + T\hat{z}_{N_p}^j) + Tw_l \leq T\hat{z}_{N_p}^j + \alpha_{N_p}^j \mathbf{1}, \qquad \forall i \in \Gamma_p, \ \forall l \in \Gamma_{\overline{w}}, \forall (j,N_p) \in I_{N_p}. \tag{20d}$$

Note that the terminal feedback gain is chosen as the gain of the predicted tubes K_{pred} as in the general complexity tube case. The resulting terminal set is defined as

$$\mathbb{Z}_f \triangleq \{z | T(z - \hat{z}) \leq \alpha \mathbf{1}, P_i T = T(A_i + B_i K_{\text{pred}}), P_i(T\hat{z} + \alpha \mathbf{1}) + Tw_l \leq T\hat{z} + \alpha \mathbf{1}, \forall i \in \Gamma_p, \ \forall l \in \Gamma_{\overline{w}}\}. \tag{21}$$

The homothetic tube enables us to formulate the tube and obtain a tight upper bound on the extreme stage costs of the tubes as follows:

$$\sum_{k=N_r}^{N_p} J_k^{\text{tube}} = \min_{\alpha_k^j, \gamma_k^j, \eta_k^j, \mu_k^j \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}} \sum_{k=N_r}^{N_p} \sum_{j=1}^{n_d^{N_r}} \omega_k^j \gamma_k^j \tag{22a}$$

subject to

$$-\mu_k^j \leq Q(\hat{z}_k^j + \alpha \vee_r - y_k^{j,r}) \leq \mu_k^j, \qquad \forall r \in \{1, \ \dots\ , \ n_v\}, (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{22b}$$

$$-\eta_k^j \leq Rv_k^j \leq \eta_k^j, \qquad \forall r \in \{1, \ \dots\ , \ n_v\}, \ (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{22c}$$

$$y_k^{j,r} \in \mathbb{Z}_f, \qquad \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{22d}$$

$$\mathbf{1}^T \mu_k^j + \mathbf{1}^T \eta_k^j \leq \gamma_k^j \qquad \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{22e}$$

where (j,r) denotes the indices associated with the vertices of the tube $\mathcal{Z}_k^j$ for all $(j,k) \in I_{[\![N_r,\ N_p-1]\!]}$. The optimization formulation in the case of homothetic tubes $\mathbb{P}_{N_p}^H(x)$ results as follows:

$$\min_{z_0^1 v_k^j y_k^j \ jkI_{0\ N_p 1]\!]} \alpha_k^j, \gamma_k^j, \eta_k^j, \mu_k^j \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}} \sum_{k=0}^{N_r-1} \sum_{j=1}^{n_d^k} \omega_k^j \ell(z_k^j, v_k^j) + \sum_{k=N_r}^{N_p-1} \sum_{j=1}^{n_d^{N_r}} \omega_k^j \gamma_k^j \tag{23}$$

subject to:

(6b), (6c), (6h), (20a)–(20d), (22b)–(22e), $T\mathcal{Z}^j_{N_r} \leq \alpha^j_{N_r}\mathbf{1},\ \forall (j, N_r) \in I_{N_r}$.

As in the general complexity tube case, the nonnegative matrices $P_i, \forall i \in \Gamma_p,\ P_x,\ P_u$ are obtained offline as defined in (10), (11a), (11c). The set of all $x \in \mathbb{X}$ for which there exists a feasible solution for the optimization problem (18) is denoted as $\mathrm{X}^H_{N_p}$. The value function obtained by solving the optimization problem (23) at every time step is denoted as $V^{H*}_{N_p}(x)$.

Remark 4. The homothetic tubes can also be implemented using the vertices of the tube as proposed in Reference 13. However, the formulation proposed here using the approach in Reference 16 can be advantageous in terms of reduced computational complexity for a small conservatism. For example, a low complexity tube for a system of n_x dimensions require $2n_x$ inequalities only, while 2^{n_x} vertices are required to represent the same tube (an exponential increase against a linear increase). By restricting the vertices to get a tight upper bound on the stage cost and not computing the reachable sets, applications to high dimensional systems can be achieved. This is the motivation for the proposed formulation of the tube-enhanced multi-stage MPC.

4.3 | Low complexity tube

To reduce the computational complexity of the scheme, low complexity tubes can be employed. A polytopic tube for the low complexity tube for a given $(j, k) \in I_{[\![N_r,\ N_p-1]\!]}$ can be defined as follows:

$$\mathcal{Z}^j_k = \{z | \underline{\tau}^j_k \leq \overline{T}z \leq \overline{\tau}^j_k\}, \tag{24}$$

where $\overline{T} \in \mathbb{R}^{n_x \times n_x}$ and $\underline{\tau}^j_k, \overline{\tau}^j_k \in \mathbb{R}^{n_x}$. Notice that the matrix $\overline{T}$ is a square matrix with n_x rows and columns. Because of the reduced number of inequalities describing the tube, the complexity can be reduced significantly. In addition, there exists efficient implementations of the tubes (See[7,29]) to improve the online computational time. The low complexity tubes can, however, lead to conservatism of the resulting scheme. In the proposed formulation, because of the presence of more degrees of freedom, the conservatism of low complexity tubes can be mitigated to a certain extent while reducing the computational time. A more detailed discussion is presented in Section 6.

4.4 | Computation of feedback gains and invariant sets

The feedback gains K_{inv} and K_{pred} should be chosen offline such that the joint spectral radii of the systems $\{(A_i + B_iK_{\text{inv}}), i \in \Gamma_p\}$ and $\{(A_i + B_iK_{\text{pred}}), i \in \Gamma_p\}$ are in the unit circle of the complex plane. They can be chosen such that they optimize a performance measure: for example, the feedback gain K_{inv} can be chosen such that the volume of the invariant tube S is as small as possible and the prediction gain K_{pred} can be chosen as the optimal gain for the linear quadratic regulator either in the nominal or in the worst-case. As we point out in Remark 2, instead of choosing a prediction gain, one can choose different gains for different scenarios. However, a systematic way to obtain multiple prediction gains is an issue for further investigations. If $N_r = N_p$, the terminal gain K_f can be chosen to maximize the volume of the terminal set to improve the volume of the feasible domain for a chosen prediction horizon N_p.

The implementation of the terminal ingredients is simplified and the complexity is reduced by using (14) and (21) because these formulations can be implemented directly in the optimization problem without computing additional sets. As pointed in References 16 and 29, this simplifies the implementation but can lead to a decrease in the feasible domain (for a fixed prediction horizon). In References 16 and 29, mode 2 dynamics are introduced to find a trade-off between improved feasible domain and computational complexity. Another option is to choose the terminal set as large as possible as described in Reference 29 (pp. 216-219). Since the proposed scheme offers additional feed-forward terms, the proposed implementation of the terminal set is not as restrictive as in the standard tube-based MPC schemes. The disturbance invariant set S can be obtained by solving the optimization problem defined in (15).

5 | RECURSIVE FEASIBILITY AND STABILITY PROPERTIES

We formulate the fundamental assumptions to establish the theoretical properties of the proposed scheme as follows.

Assumption 3. A convex compact disturbance invariant polytopic set S is available for the system (3) if $\underline{\mathsf{W}} \neq \{0\}$, where $\mathsf{S} \subset \mathbb{X}$ and $K_{\text{inv}}\mathsf{S} \subset \mathbb{U}$. If $\underline{\mathsf{W}} = \{0\}$, $\mathsf{S} = \{0\}$.

Assumption 3 is required to obtain a nonempty feasible domain. The proposed formulation (6) offers flexibility to decompose the given uncertainty set W. Only the disturbances considered in $\underline{\mathsf{W}}$ is used to build the set S. Hence, satisfying the assumptions is always possible because $\underline{\mathsf{W}} = \{0\}$ is always a possible choice. If $\underline{\mathsf{W}} = \{0\}$, there is no tightening of the constraints required, but this can result in increased computational complexity.

Assumption 4. An RPI polytopic set $\mathbb{Z}_f \subseteq \mathbb{X} \ominus \mathsf{S}$ that contains the origin is available for the system (3) for the feedback gain $K_f = K_{\text{pred}}$ that satisfies $K_f\mathbb{Z}_f \subseteq \mathbb{U} \ominus K_{\text{inv}}\mathsf{S}$ such that for all $z \in \mathsf{Z} \subseteq \mathbb{Z}_f$, $(A_i + B_iK_f)\mathsf{Z} \oplus \overline{W} \subseteq \mathsf{Z}^+ \subseteq \mathbb{Z}_f,\ \forall i \in \Gamma_p$ holds, where Z and Z^+ represent the employed tubes (that over-approximates the reachable set) in the optimization problem (6).

If $N_r = N_p$, there is no tube-based part in the predictions. Hence it is sufficient that the terminal set is robustly invariant with respect to the system (3) for a terminal feedback gain K_f. In this case, the tube Z in Assumption 4 is defined as a singleton $\mathsf{Z} = \{z\}$ and $\mathsf{Z}^+ = \{(A_i + B_iKf)z + w_l, i \in \Gamma_p, l \in \Gamma_{\overline{w}}\}$. However , if $N_r < N_p$, the tube-based part of the scheme requires that the set recursion (5) employed by the tube is also robustly invariant, that is, it is also necessary that the terminal set in robust positively invariant with the employed tubes that over-approximate the reachable sets of the system at every time step for the feedback gain K_{pred}.

Remark 5. We propose to keep the terminal feedback gain K_f the same as that of gain of the predicted tubes. If K_f is chosen different from K_{pred}, then the requirement in Assumption 4 should be modified as follows to guarantee recursive feasibility: An RPI polytopic set $\mathbb{Z}_f \subseteq \mathbb{X} \ominus \mathsf{S}$ that contains the origin is available for the system (3) for the control law $K_{\text{pred}}\mathbb{Z}_f \subseteq \mathbb{U} \ominus K_{\text{inv}}\mathsf{S}$ and $K_f\mathbb{Z}_f \subseteq \mathbb{U} \ominus K_{\text{inv}}\mathsf{S}$ such that for all $\mathsf{Z} \subseteq \mathbb{Z}_f$, $(A_i + B_iK_{\text{pred}})\mathsf{Z} \oplus \overline{W} \subseteq \mathsf{Z}^+ \subseteq \mathbb{Z}_f,\ \forall i \in \Gamma_p$ and $(A_i + B_iK_f)\mathsf{Z}_f \oplus \overline{W} \subseteq \mathbb{Z}_f,\ \forall i \in \Gamma_p$ hold. This leads to additional complexity in the terminal set and it is not clear if it leads to advantages in terms of performance. Keeping $K_f = K_{\text{pred}}$ simplifies the requirement and is consistent with the tube-based schemes proposed in References 16 and 29.

Lemma 2. *Suppose Assumptions 3 and 4 hold and $x \in \mathtt{X}_{N_p}$ such that $\mathtt{P}_{N_p}(x)$ (6) has a feasible solution, then $\mathtt{P}_{N_p}(x^+)$ is feasible for all $x^+ \in \text{conv}(\{A_ix + B_iu\}) \oplus \mathsf{W},\ \forall i \in \Gamma_p$ if the control input applied to the system (1) follows the control policy $u = v + K_{\text{inv}}(x - z)$.*

Proof. Let the sequence of optimal control inputs obtained by solving the problem (6) be defined as

$$\mathbf{v^*} = \{v_0^{1*},\ \ \dots,\ v_{N_r}^{j*},\ v_{N_r+1}^{j*} + K_{\text{pred}}z_{N_r+1}^{j*},\ \ \dots,\ v_{N_p-1}^{j*} + K_{\text{pred}}z_{N_p-1}^{j*},\ \forall (j,k) \in I_{[\![0,\ N_p-1]\!]}\}. \tag{25}$$

The root node of the scenario tree is a decision variable as defined in (6g). Let the optimal value be denoted as $z = z_0^{1*}$. The first element in the input sequence sequence $v = v_0^{1*}$ and the optimal root node $z = z_0^{1*}$ are used in the control law $u = v + K_{\text{inv}}(x - z)$, where x is the current state of the plant. The input u is then applied to the plant. The plant evolves from the current state x to the state x^+ for the applied input u and the realizations of the uncertainties $w \in \mathsf{W}$ and the system matrices $(A, B) \in \text{conv}\{(\{A_i, B_i,\ \forall i \in \Gamma_p\})\}$. x^+ satisfies the constraints because, the additive disturbances $w \in \overline{\mathsf{W}}$ and the vertex matrices $\{(A_i, B_i),\ \forall i \in \Gamma_p\}$ are explicitly considered in the scenario tree in the predictions and the invariant set S accounts for the disturbances $w \in \underline{\mathsf{W}}$. At the next time step, the optimization problem (6) is solved again for the realized state x^+. Since S is invariant with respect to the additive disturbances $w \in \underline{\mathsf{W}}$, there exists a $z^+ \in \text{conv}(\{z_1^{j*},\ \forall (j,1) \in I_1\}) \subseteq \mathbb{Z} = \mathbb{X} \ominus \mathsf{S}$ from Assumption 3. There exists a feasible input sequence for the optimization problem $\mathtt{P}_{N_p}(x^+)$ that is in the convex hull of inputs predicted in the previous time step for the optimization problem $\mathtt{P}_{N_p}(x)$ for all prediction steps until $N_p - 1$. For the last prediction step, there exists a control law K_fz for all $z \in \mathbb{Z}_f$ from Assumption 4. A feasible input sequence for the next time step can be obtained as the convex combination of the predicted inputs as follows:

$$\mathbf{v}(x^+) = \left\{ \sum_{(j,1)\in I_1} \lambda_1^j v_1^{j*},\ \dots,\ \sum_{(j,N_r)\in I_{N_r}} \lambda_{N_r}^j v_{N_r}^{j*},\ \sum_{(j,N_r+1)\in I_{N_r+1}} \lambda_{N_r+1}^j v_{N_r+1}^{j*} + K_{\text{pred}}z_{N_r+1}^{j*},\ \dots,\ \sum_{(j,N_p)\in I_{N_p}} \lambda_{N_p}^{j*} K_f z_{N_p}^{j*} \right\}, \tag{26}$$

where λ_k^j denote the associated convex weights for all $(j,k) \in I$ ($\lambda_k^j \geq 0$, and $\sum_{j=1}^{n_d} \lambda_k^j = 1$). Since there exists a feasible root node z^+ and a feasible input sequence $\mathbf{v}(x^+)$, problem $\mathtt{P}_{N_p}(x^+)$ is feasible for all $x^+ \in \{A_i x + B_i u\} \oplus \mathsf{W}$ for all $i \in \Gamma_p$ if $\mathtt{P}_{N_p}(x)$ is feasible. ∎

We now show that the recursve feasibility property is retained if the convex optimization problems (18) and (23) related to the formulation (6) are solved.

Corollary 1. *Suppose Assumptions 3 and 4 hold and $x \in \mathtt{X}_{N_p}$ such that $\mathtt{P}^{\mathrm{G}}_{N_p}(x)$ (18) has a feasible solution for the tube $\mathsf{Z}_k^j = \{z | Tz \leq \tau_k^j\}, \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}$, then $\mathtt{P}^{\mathrm{G}}_{N_p}(x^+)$ is feasible for all $x^+ \in \mathrm{conv}(\{A_i x + B_i u\}) \oplus \mathsf{W},\ \forall i \in \Gamma_p$ if the control input applied to the system (1) follows the control policy $u = v + K_{\mathrm{inv}}(x - z)$.*

Proof. The constraints of the optimization problems (6) and (18) can be compared one to one. The constraints (6b), (6c), (6h) are retained in optimization problem (18). The remaining constraints are direct results of Lemma 1 which establishes sufficient conditions for the set recursion and guaranteeing that a set is a subset of another. Hence, the feasibility arguments discussed in Lemma 2 directly applies to the formulation (18). Hence, $\mathtt{P}^{\mathrm{G}}_{N_p}(x^+)$ is feasible for all $x^+ \in \mathrm{conv}(\{A_i x + B_i u\}) \oplus \mathsf{W},\ \forall i \in \Gamma_p$ if the control input applied to the system (1) follows the control policy $u = v + K_{\mathrm{inv}}(x - z)$. ∎

Corollary 2. *Suppose Assumptions 3 and 4 hold and $x \in \mathtt{X}_{N_p}$ such that $\mathtt{P}^{\mathrm{H}}_{N_p}(x)$ (23) has a feasible solution for the tube $\mathsf{Z}_k^j = \{z | T(z - \hat{z}_k^j) \leq \alpha_k^j \mathbf{1}\}, \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}$, then $\mathtt{P}^{\mathrm{H}}_{N_p}(x^+)$ is feasible for all $x^+ \in \mathrm{conv}(\{A_i x + B_i u\}) \oplus \mathsf{W},\ \forall i \in \Gamma_p$ if the control input applied to the system (1) follows the control policy $u = v + K_{\mathrm{inv}}(x - z)$.*

Proof. The same arguments in Corollary 1 directly apply here as well. ∎

Lemma 3. *If Assumptions 1–4 hold, then $V^*_{N_p+1}(x) \leq V^*_{N_p}(x),\ \forall x \in \mathtt{X}_{N_p}$, where $V^*_{N_p}(x)$ is the optimal value function of (6) with the length of the prediction horizon N_p.*

Proof. Since the stage costs $\ell(z, K_f z) = 0$ and $\max_{\tilde{z} \in \mathsf{Z}} \ell(z, K_f z) = 0$ in the terminal set and the control law $K_f z$ is feasible, the additional prediction step does not add any cost to the optimal value function $V^*_{N_p+1}(x)$ and hence $V^*_{N_p+1}(x) \leq V^*_{N_p}(x),\ \forall x \in \mathtt{X}_{N_p}$. ∎

Lemma 4. *If Assumptions 1–4 hold and $\mathtt{X}_{N_p}$ is compact, then the optimal value function fulfills the following properties*

$$V^*_{N_p}(x) = 0,\ \forall x \in \mathbb{Z}_f \oplus \mathsf{S}, \tag{27}$$

$$V^*_{N_p}(x) \geq c_1 |z|_{\mathbb{Z}_f},\ \forall x \in \mathtt{X}_{N_p}, \tag{28}$$

$$V^*_{N_p}(x^+) \leq V^*_{N_p}(x) - c_1 |z|_{\mathbb{Z}_f},\ \forall x \in \mathtt{X}_{N_p}, \tag{29}$$

where c_1 is a positive constant.

Proof. Since $x \in \mathbb{Z}_f \oplus \mathsf{S}$ implies that $z_0^1(x) \in \mathbb{Z}_f$ is a feasible point, the stage cost $\ell(z, K_f z) = 0$ from Assumption 2 and the control law $K_f z$ keeps the state z in the terminal set $\mathbb{Z}_f$ from Assumption 4. Since $V_f(z) = 0$ and all the stage costs are zero $V^*_{N_p}(x) = 0$ for all $x \in \mathbb{Z}_f \oplus \mathsf{S}$. This proves (27).

From Assumption 2, $\ell(z, v) \geq k|z|_{\mathbb{Z}_f}$ and the optimal value function satisfies the property $V^*_{N_p}(x) \geq \omega_0^1 \ell(z_0^{1*}, v_0^{1*})$. Since $\omega_0^1 \geq 0$, we have $V^*_{N_p}(x) \geq c_1 |z|_{\mathbb{Z}_f}$ for all $z \in \mathbb{Z} \setminus \mathbb{Z}_f$, where $c_1 = k\omega_0^1$ a positive constant. Note that $V^*_{N_p}(x) = 0$ for all $x \in \mathbb{Z}_f \oplus \mathsf{S}$ and $|z|_{\mathbb{Z}_f} = 0$ for all $z \in \mathbb{Z}_f$. Hence $V^*_{N_p}(x) \geq c_1 |z|_{\mathbb{Z}_f}$ for all $x \in \mathtt{X}_{N_p}$.

In the following, the descent property of the optimal value function will be proven. This property establishes the optimal value function as a Lyapunov function and is an important contribution of this article.

As shown in Lemma 1, the initial state $z_0^{1*}(x)$ and the control policy given in (26) are feasible for all $x \in \mathtt{X}_{N_p}$. Since $z^+ \in \mathrm{conv}(\{z_1^{j*},\ \forall (j,1) \in I\})$, the optimal value function $V^*_{N_p}(x)$ and a feasible value function at the next time step ($V_{N_p}(x^+)$)

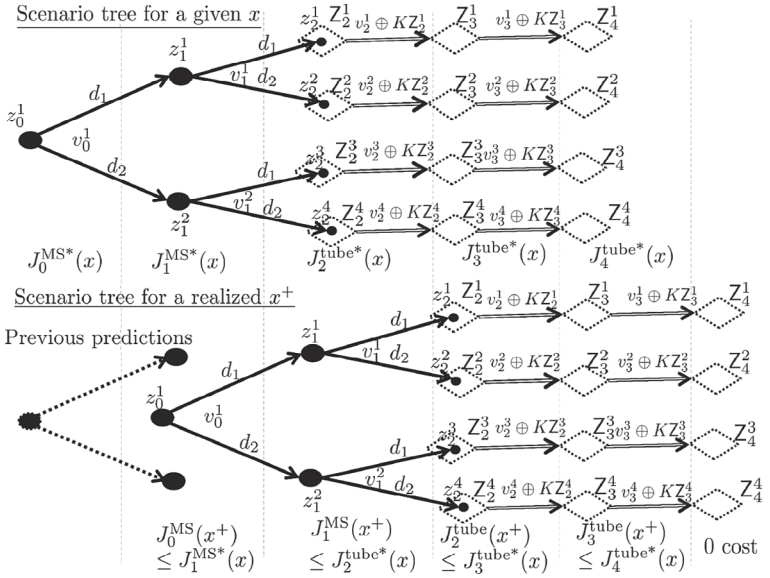


FIGURE 3 Illustration of the descent property of the proposed scheme

can be written as:

$$V_{N_p}^*(x) = \sum_{k=0}^{N_r-1} J_k^{\mathrm{MS}*}(x) + J_{N_r}^{\mathrm{tube}*}(x) + \sum_{k=N_r+1}^{N_p-1} J_k^{\mathrm{tube}*}(x), \tag{30a}$$

$$V_{N_p}(x^+) = \sum_{k=0}^{N_r-1} J_k^{\mathrm{MS}}(x^+) + \sum_{k=N_r}^{N_p-1} J_k^{\mathrm{tube}}(x^+). \tag{30b}$$

The proof will be accomplished in two parts. Initially the multi-stage part of the value function $V_{N_p}(x^+)$ will be compared to the optimal value function $V_{N_p}^*(x)$ obtained one step before. Because of the receding horizon nature of the MPC, the comparison of the multi-stage part of $V_{N_p}(x^+)$ will be performed with the multi-stage part and the one-step tube-based part $(J_k^{\mathrm{MS}*}(x) + J_{N_r}^{\mathrm{tube}*}(x))$ of the optimal value function $V_{N_p}^*(x)$. It will be shown that $\sum_{k=0}^{N_r-1} J_k^{\mathrm{MS}}(x^+) \le \sum_{k=0}^{N_r-1} J_k^{\mathrm{MS}*}(x) + J_{N_r}^{\mathrm{tube}*}(x) - c_1|z|_{\mathbb{Z}_f}$. The multi-stage part of the value function $\sum_{k=0}^{N_r-1} J_k^{\mathrm{MS}}(x^+)$ can be written as (refer to Figure 3):

$$\begin{aligned}\sum_{k=0}^{N_r-1} J_k^{\mathrm{MS}}(x^+) = \omega_0^1 \ell\left(\sum_{j=1}^{n_d} \lambda_1^j z_1^{j*}, \sum_{j=1}^{n_d} \lambda_1^j v_1^{j*}\right) + \ \dots \ + \sum_{j=1}^{n_d^{N_r-2}} \omega_{N_r-2}^j \ \ell\left(\sum_{j=1}^{n_d} \lambda_{N_r-1}^j z_{N_r}^{j*}, \sum_{j=1}^{n_d} \lambda_{N_r-1}^j v_{N_r}^{j*}\right) + \\ + \sum_{j=1}^{n_d^{N_r-1}} \omega_{N_r-1}^j \ \ell\left(\sum_{j=1}^{n_d} \lambda_{N_r}^j z_{N_r+1}^{j+}, \sum_{j=1}^{n_d} \lambda_{N_r}^j (v_{N_r}^{j*} + K_{\mathrm{pred}} z_{N_r}^{j+})\right),\end{aligned} \tag{31}$$

where $z_{N_r}^{j+} \in \mathrm{conv}(\{Z_{N_r}^j(x)\},\ \forall (j, N_r) \in I_{N_r})$. Since the proposed stage cost is convex from Assumption 2, the following inequality holds for the stage cost of the root node of the tree at x^+.

$$\ell\left(\sum_{j=1}^{n_d} \lambda_1^j z_1^{j*}, \sum_{j=1}^{n_d} \lambda_1^j v_1^{j*}\right) \le \sum_{j=1}^{n_d} \lambda_1^j \ell(z_1^{j*}, v_1^{j*}). \tag{32}$$

Multiplying both sides with ω_0^1, we get

$$\omega_0^1 \ell\left(\sum_{j=1}^{n_d} \lambda_1^j z_1^{j*}, \sum_{j=1}^{n_d} \lambda_1^j v_1^{j*}\right) \le \omega_0^1 \sum_{j=1}^{n_d} \lambda_1^j \ell(z_1^{j*}, v_1^{j*}). \tag{33}$$

From Assumption 1, $\omega_0^1 \leq \min\{\omega_1, \ \dots \ , \ \omega_p\}$, so $\omega_0^1 \sum_{j=1}^{n_d} \lambda_1^j \ell(z_1^{j*}, v_1^{j*}) \leq \sum_{j=1}^{n_d} \lambda_1^j \omega_1^j \ell(z_1^{j*}, v_1^{j*})$ and $\sum_{j=1}^{n_d} \lambda_1^j = 1$, which results in:

$$\omega_0^1 \ell\left(\sum_{j=1}^{n_d} \lambda_1^j z_1^{j*}, \sum_{j=1}^{n_d} \lambda_1^j v_1^{j*}\right) \leq \sum_{j=1}^{n_d} \omega_1^j \lambda_1^j \ell(z_1^{j*}, v_1^{j*}). \tag{34}$$

Similarly the stage costs for the prediction steps until $N_r - 2$ of the value function $V_{N_p}(x^+)$ can be compared for all $(j,k) \in I_{[\![0,\ N_r-2]\!]}$ by grouping the nodes that are established for a particular realization of the uncertainty $i \in \Gamma_p$ and $l \in \Gamma_{\overline{w}}$ together. Note that, if $N_r = 0$, there is no multi-stage part in the objective. If $N_r = 1$ only the root node belongs to the multi-stage part and it will be compared with the tube-based part (first-stage) of the optimal value function in the previous time-step $V_{N_p}^*(x)$ as will be discussed later. If $N_r = 2$, the root node will be compared to the multi-stage part as in (34) and for the comparison of the further stages onwards one can skip to (36). The following relationship for the case is discussed whenever $N_r > 2$ until $N_r - 2$ of the value function $V_{N_p}(x^+)$. Since the weights ω_k^j associated with each node z_k^j are the same as the weights associated with a realized uncertainty until $N_r - 1$ of the problem $\mathtt{P}_{N_p}(x^+)$, we have

$$\omega_k^j \ell\left(\sum_{j=1}^{n_d} \lambda_k^j z_k^{j*}, \sum_{j=1}^{n_d} \lambda_k^j v_k^{j*}\right) \leq \sum_{j=1}^{n_d} \omega_{k+1}^j \lambda_k^j \ell(z_k^{j*}, v_k^{j*}). \tag{35}$$

The stage cost at the prediction step $N_r - 1$ of the problem $\mathtt{P}_{N_p}(x^+)$ can be compared with the tube-based part of the optimal value function $V_{N_p}^*(x)$. Again using convexity, the following relationship can be established.

$$\ell\left(\sum_{j=1}^{n_d} \lambda_{N_r}^j z_{N_r}^{j+}, \sum_{j=1}^{n_d} \lambda_{N_r}^j (v_{N_r}^{j*} + K_{\text{pred}} z_{N_r}^{j+})\right) \leq \ell(\tilde{z}_{N_r}^{j*}, v_k^{j*} + K_{\text{pred}} \tilde{z}_{N_r}^{j*}), \tag{36}$$

where $z_{N_r}^{j+}$ belongs to the convex hull of the tubes $Z_{N_r}^{j*}$ predicted in the problem $\mathtt{P}_{N_p}(x)$. Note that the stage cost of the tube-based part of the optimal value function $V_{N_p}^*(x)$ contains the maximization objective, the optimal states $\tilde{z}_{N_r}^{j*}$ give the worst-case cost for all $(j, N_r) \in I_{N_r}$. Since the weight associated with $J_{N_r}^{\text{tube}*}(x)$ is greater than the weights associated with all the realizations of the uncertainty from Assumption 1, we have

$$\omega_{N_r-1}^j \ell\left(\sum_{j=1}^{n_d} \lambda_{N_r}^j z_{N_r}^{j+}, \sum_{j=1}^{n_d} \lambda_{N_r}^j (v_{N_r}^{j*} + K_{\text{pred}} z_{N_r}^{j+})\right) \leq \sum_{j=1}^{n_d} \omega_{\text{tube}} \lambda_{N_r}^j \ell(\tilde{z}_{N_r}^{j*}, v_k^{j*} + K_{\text{pred}} \tilde{z}_{N_r}^{j*}), \tag{37}$$

Substituting (34), (35), (37) in (31), we get

$$\sum_{k=0}^{N_r-1} J_k^{\text{MS}}(x^+) \leq \sum_{k=1}^{N_r-1} J_k^{\text{MS}}(x) + J_{N_r}^{\text{tube}*}(x), \tag{38}$$

$$\leq \sum_{k=0}^{N_r-1} J_k^{\text{MS}}(x) + J_{N_r}^{\text{tube}*}(x) - \omega_0^1 \ell(z_0^{1*}, v_0^{1*}). \tag{39}$$

The tube-based part of the optimal value function $V_{N_p}(x^+)$ can be compared with the tube-based part of $V_{N_p}^*(x)$. Since the predicted tubes of the problem $\mathtt{P}_{N_p}(x^+)$ from the prediction step N_r until $N_p - 2$ belong to the convex hull of $\mathtt{P}_{N_p}(x)$, we have

$$\sum_{k=N_r}^{N_p-2} J_k^{\text{tube}}(x^+) \leq \sum_{k=N_r+1}^{N_p-1} J_k^{\text{tube}*}(x). \tag{40}$$

The stage costs at the prediction step $N_p - 1$ are $\ell(\sum_{j=1}^{n_d} \lambda_{N_p}^j \tilde{z}_{N_p}^{j*}(x), \ \sum_{j=1}^{n_d} \lambda_{N_p}^j K_f \tilde{z}_{N_p}^{j*}) = 0$ by definition. Hence the following relationship holds:

$$\sum_{k=N_r}^{N_p-1} J_k^{\text{tube}}(x^+) \leq \sum_{k=N_r+1}^{N_p-1} J_k^{\text{tube}*}(x). \tag{41}$$

From (39) and (41), the value function $V_{N_p}(x^+)$ can then be written in terms of $V_{N_p}(x)$ as follows:

$$V_{N_p}(x^+) \leq V^*_{N_p}(x) - \omega^1_0 \ell(z^{1*}_0, v^{1*}_0). \tag{42}$$

Since $V^*_{N_p}(x^+) \leq V_{N_p}(x^+)$, we have

$$V^*_{N_p}(x^+) \leq V^*_{N_p}(x) - \omega^1_0 \ell(z^{1*}_0, v^{1*}_0), \tag{43}$$

$$V^*_{N_p}(x^+) \leq V^*_{N_p}(x) - c_1|z|_{\mathbb{Z}_f}, \forall x \in \mathrm{X}_{N_p}. \tag{44}$$

This proves (29) and with this Lemma 3 is established. ∎

Remark 6. The proof of descent (29) is valid for all values of robust horizon in the range $N_r \in [0, N_p]$. If $N_r = 0$, then the scheme is simplified into a tube-based MPC scheme enhanced by an invariant tube. If $N_r = N_p$, the scheme simplifies into a multi-stage MPC scheme enhanced by an invariant tube. The proof was established for a generic case where both the multi-stage and tube components are present. If one of the components is absent, it can be shown that the proof still holds by removing the corresponding elements in the proof.

Lemma 5. *If Assumptions 1–4 hold and $\mathrm{X}^G_{N_p}$ is compact, the optimal value function $V^{G*}_{N_p}(x)$ of the optimization problem (18) fulfills the following properties for the system (1)*

$$V^{G*}_{N_p}(x) = 0, \ \forall x \in \mathbb{Z}_f \oplus \mathsf{S}, \tag{45}$$

$$V^{G*}_{N_p}(x) \geq c_2|z|_{\mathbb{Z}_f}, \ \forall x \in \mathrm{X}^G_{N_p}, \tag{46}$$

$$V^{G*}_{N_p}(x^+) \leq V^{G*}_{N_p}(x) - c_2|z|_{\mathbb{Z}_f}, \ \forall x \in \mathrm{X}^G_{N_p}, \tag{47}$$

where c_2 is a positive constant.

Proof. The proof can be found in Appendix A. ∎

Lemma 6. *If Assumptions 1–4 hold and $\mathrm{X}^H_{N_p}$ is compact, the optimal value function $V^{H*}_{N_p}(x)$ fulfills the following properties*

$$V^{H*}_{N_p}(x) = 0, \ \forall x \in \mathbb{Z}_f \oplus \mathsf{S}, \tag{48}$$

$$V^{H*}_{N_p}(x) \geq c_3|z|_{\mathbb{Z}_f}, \ \forall x \in \mathrm{X}^H_{N_p}, \tag{49}$$

$$V^{H*}_{N_p}(x^+) \leq V^{H*}_{N_p}(x) - c_3|z|_{\mathbb{Z}_f}, \ \forall x \in \mathrm{X}^H_{N_p}, \tag{50}$$

where c_3 is a positive constant.

Proof. The proof is given in Appendix B. ∎

The terminal set $\mathbb{Z}_f \oplus \mathsf{S}$ is asymptotically stable for the proposed scheme as it is shown in the following theorem.

Theorem 1. *Suppose the Assumptions 1-4 are satisfied, then the set $\mathbb{Z}_f \oplus \mathsf{S}$ is robustly asymptotically stable for the controlled uncertain system given in (1) using the proposed scheme (6) with the implementation (18) or (23).*

Proof. Since the optimal value function is established as a Lyapunov function in Lemmas 4–6 with respect to the terminal set $\mathbb{Z}_f$, the state z of (3) converges to the terminal set asymptotically. The state of the system (1) satisfies the property $x \in \{z\} \oplus \mathsf{S}$ and converges robustly asymptotically to the set $\mathbb{Z}_f \oplus \mathsf{S}$. ∎

Finite time reachability of the terminal set and robust asymptotic stability of the minimal RPI set can be proven for the proposed scheme using a dual mode control policy as proposed in Reference 26. The required conditions are formalized in the following assumption.

Assumption 5. There exists a dual mode control policy that is employed as follows:

$$u(x) = \begin{cases} K_f x, & \text{if } x \in \mathbb{X}_{max}, \\ v + K_{\text{inv}}(x - z), & \text{otherwise}, \end{cases} \tag{51}$$

where $\mathbb{X}_{max}$ is an RPI set for the asymptotically stabilizing control law $u = K_f x$ for the system (1) that satisfies the conditions $\mathbb{Z}_f \oplus \mathsf{S} \subset \mathbb{X}_{max}$ and $\mathbb{X}_{max} \subseteq \mathbb{X}$ and v and z are optimal solutions of the proposed scheme (6) with an asymptotically stabilizing feedback gain K_{inv}.

Theorem 2. *Suppose the Assumptions 1–5 are satisfied, the minimal RPI set S_{min} of the system (1) is robustly asymptotically stable for the controlled uncertain system defined in (1) employed using the dual mode control policy (51).*

Proof. As $\mathbb{Z}_f \oplus \mathsf{S} \subset \mathbb{X}_{max}$ and $\mathbb{Z}_f \oplus \mathsf{S}$ is robustly asymptotically stable from Theorem 1, the state enters $\mathbb{X}_{max}$ in finitely many time steps. Since the control policy is switched to $K_f x$ when $x \in \mathbb{X}_{max}$ and that the control law $K_f x$ is asymptotically stabilizing for the system (1), the state x converges to the minimal RPI set asymptotically. Hence for the uncertain system defined in (1) controlled using the dual mode control policy (51), the minimal RPI set S_{min} for the control law $u = K_f x$ is robustly asymptotically stable. ∎

Corollary 3. *Suppose the Assumptions 1–5 are satisfied, the set $\mathbb{Z}_f \oplus \mathsf{S}$ can be reached in finite time steps if $\mathsf{S}_{min} \subset \mathbb{Z}_f \oplus \mathsf{S}$ holds.*

Proof. This follows directly from the proof of Theorem 2. If the minimal RPI set S_{min} is contained in $\mathbb{Z}_f \oplus \mathsf{S}$ and S_{min} is robustly asymptotically stable, the state reaches $\mathbb{Z}_f \oplus \mathsf{S}$ in finite time steps. ∎

In the proposed approach, a multi-stage MPC solution is computed on the scenario tree for the large uncertainties with recourse, that is, a tree of future inputs depends on the realization of the uncertainty. The affine feedback is added "on top" to robustify the solution against the small disturbances. The feedback gain K_{inv} is fixed only for small disturbances and the degrees of freedom are increased using the multi-stage approach for large uncertainties resulting in an improved trade-off between optimality and complexity. Also, if the robust horizon is chosen as $N_r \geq 1$, we have different feed-forward terms at each stage in the predictions. This results in a scheme with the following advantages when compared to multi-stage MPC and tube-based MPC independently:

1. The growth in problem complexity is reduced when compared to a pure multi-stage approach because the small uncertainties are not considered in the scenario tree.
2. The structurally relaxed recourse which is modeled in the prediction for the realizations of the large uncertainties until robust horizon reduces the conservatism compared to pure tube-based MPC.
3. The choice of a robust horizon on the one hand limits the rapid growth of the scenario tree and on the other hand, provides increased degrees of freedom when compared to a standard tube-based MPC resulting in an improved trade-off.
4. The use of low complexity tubes is possible for less conservatism because of increased degrees of freedom in the form of feed-forward terms beyond the robust horizon. This enables the application of the approach to high dimensional systems.

6 | CASE STUDY

The example considered in this article is a continuous stirred-tank reactor (CSTR) with a reaction scheme that is adapted from Reference 30. Two chemical reactions take place in the reactor:

$$\begin{aligned} \mathrm{A} &\rightarrow \mathrm{B} \rightarrow \mathrm{C} \\ 2\mathrm{A} &\rightarrow \mathrm{D} \end{aligned}$$

TABLE 1 Details of the different types of tubes studied in this work

Tube type	Implemented optimization problem	Tube complexity	Recursive feasibility	Stability
general complexity tube tube	(18)	18 inequalities and 44 vertices	YES	YES
Homothetic tube	(23)	18 inequalities and 44 vertices	YES	YES
Low complexity tube	(23) with 8 row T matirx	8 inequalities and 16 vertices	YES	YES

The linearized discrete-time model has the form given in (1), where $x = [\Delta C_a,\ \Delta C_b,\ \Delta T_R, \Delta T_J]^T$ and $u = \Delta F$. ΔC_a and ΔC_b denotes the deviations of concentration of component A and B in mol/l , ΔT_R and ΔT_J denote the change in the temperature of the reactor and in the jacket temperature in °C with respect to the equilibrium point. The input is the deviation of the feed from the equilibrium in l/h. The uncertain model is given as:

$$A_{\text{unc}}(d_1, d_2, d_3, d_4) = \begin{pmatrix} 0.3 + d_1 & -0.09 & -0.01 & 0 \\ 0.2 & 0.29 + d_2 & 0.002 & 0 \\ d_3 & d_4 & 1.10 & 0.15 \\ 0.05 & 0.07 & 0.13 & 0.68 \end{pmatrix}.$$

The system vertex matrices are

$$\begin{aligned} A_1 &= A_{\text{unc}}(0.1, 0.1, 0.33, 0.26), \\ A_2 &= A_{\text{unc}}(-0.1, -0.1, -0.33, -0.26), \\ A_3 &= A_{\text{unc}}(0.1, -0.1, 0.33, -0.26), \\ A_4 &= A_{\text{unc}}(-0.1, 0.1, -0.33, 0.26). \end{aligned}$$

The input vertex matrices are $B_1 = B_2 = B_3 = B_4 = [0.1, -0.05, 0.8, 0.1]^T$. The additive uncertainty bounds on all states are $w \in \mathsf{W} := \{w|\ ||w||_\infty \leq 0.1\}$. The state constraints are $x \in \mathbb{X} := \{x|[-5, -5, -3, -5]^T \leq x \leq [5, 5, 3, 5]^T\}$ and the input bounds are given by $u \in \mathbb{U} := \{u|\ |u| \leq 2\}$. The control task is to take the system to a bounded set around the origin while respecting the state and input constraints. The Q matrix of the stage cost is chosen as an identity matrix and the R matrix is chosen as 0.01. The length of the prediction horizon of is chosen as $N_p = 5$.

6.1 | Details of the simulation study

The tube-based part of the proposed scheme was implemented with three different types of tubes and different simulation studies were performed. The types of the tubes studied are given in Table 1.

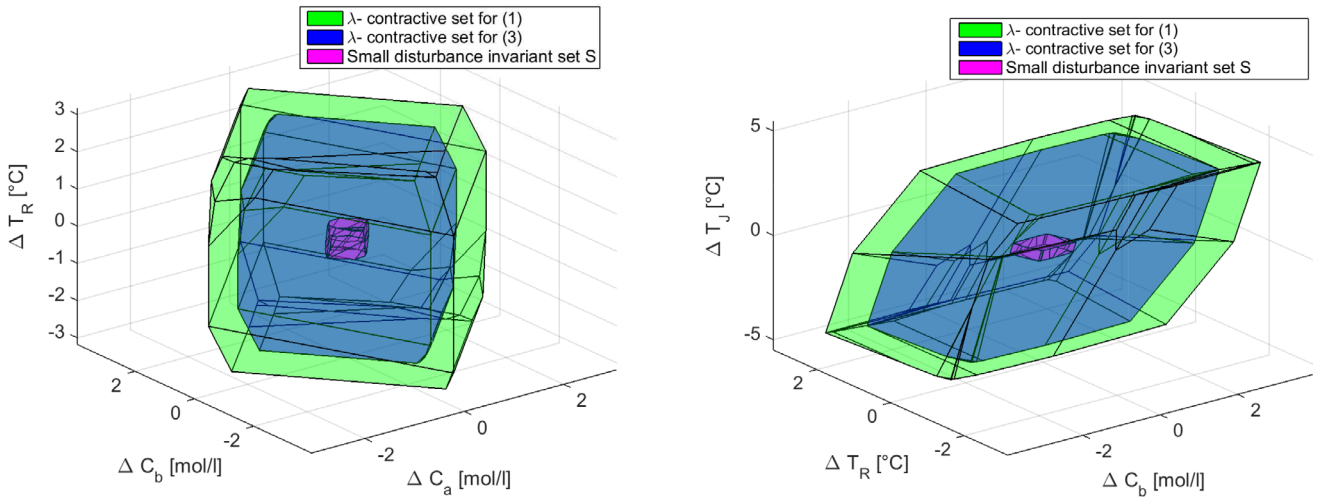
The additive disturbances are considered as small disturbances and an offline invariant tube S was obtained. The set S is obtained by using (15) using the λ-contractive set for the value of $\lambda = 0.68$ obtained for the system (1) and it is contained in the box given by $\mathsf{S} \subseteq \{x|[-0.4088, -0.5670, -0.3936, -0.3518]^T \leq x \leq [0.4088, 0.5670, 0.3936, 0.3518]^T\}$ for the LQ-optimal feedback gain $K_{\text{inv}} = K_{\text{pred}} = K_f = K = [-0.0493, -0.0004, -1.3330, -0.3485]$. In Table 2, the volumes and the computation times of the proposed approach for different robust horizons and the tube-based MPC approach are given. In the following, we investigate all aspects of the proposed approach.

6.2 | The effect of the invariant tube

If the pure multi-stage MPC is applied, it gives rise to 64 branches per node in the scenario tree resulting in more than 100 million scenarios. The additive disturbances therefore are removed from the multi-stage part and are formulated in the (invariant) tube-based part of the scheme. Hence, $\overline{\mathsf{W}} = \{0\}$ and $\underline{\mathsf{W}} = \mathsf{W}$ for the studied example. This results in four branches at every node which is a dramatic reduction when compared to the 64 branches required in the case of a full scenario tree for all the uncertainties.

TABLE 2 Quantitative comparison of feasible domains obtained using the proposed tube-enhanced multi-stage (TEMS) MPC scheme for varying robust horizons and the tube-based MPC

Tube type	Parameters	Tube MPC (without the invariant tube)	TEMS MPC with $N_r = 0$	TEMS MPC with $N_r = 1$	TEMS MPC with $N_r = 2$	TEMS MPC with $N_r = 3$	TEMS MPC with $N_r = 4$	TEMS MPC with $N_r = 5$
General complexity	Volume	1197.1	1110.7	4007.6	4392.7	4570.9	4574.6	4574.6
	Comp. time (s)	6.55	0.15	0.45	1.27	3.54	8.65	1.2
Homothetic	Volume	1065.2	1001.0	3820.3	4319.5	4536.6	4574.2	as
	Comp. time (s)	5.04	0.20	0.73	2.3	7.03	14.18	above
Low complexity	Volume	96.02	96.02	1415.9	3563.2	4411.3	4545.9	as
	Comp. time (s)	1.1	0.07	0.28	0.82	2.8	6.5	above



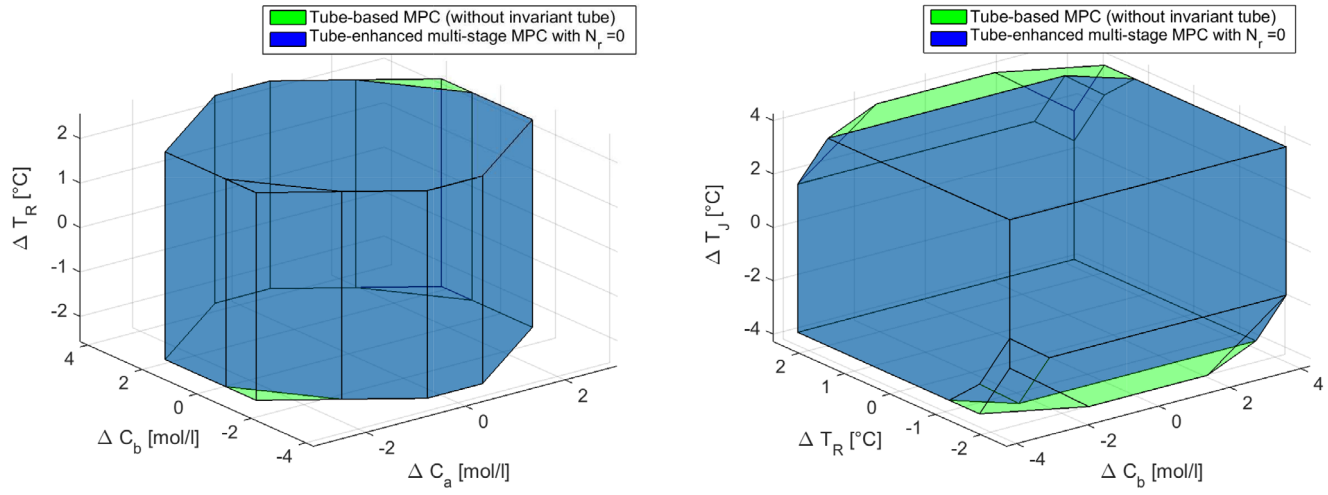
(A) Projection on ΔC_a, ΔC_b and ΔT_R.

(B) Projection on ΔC_b, ΔT_R and ΔT_J.

FIGURE 4 Plot of λ-contractive sets and the disturbance invariant set S [Colour figure can be viewed at wileyonlinelibrary.com]

For comparison purposes, tube-based MPC was implemented for the three different types of tubes considered in the proposed framework. The λ-contractive set resulted in 32 inequalities for the system (1) compared to 18 inequalities for the system (3). The contractive sets were computed using the multiparametric toolbox.[27] Since the additive disturbances are not considered in (3), the resulting complexity of the λ-contractive set is different. The 3D projections of the λ-contractive sets and the small disturbances invariant set S are shown in Figure 4.

If the tube-based MPC[16] is employed without the classification of uncertainties proposed in this article, the tube at every time step is characterized by 32 inequalities in the case of homothetic and general complexity tube tubes and the number of constraints increases with the number of additive and parametric uncertainties. In the studied example, we must consider four vertex matrices and 16 vertices of the additive disturbance set. Hence, to formulate the propagation of tubes, $32 \times 4 \times 16 = 2048$ constraints are required per prediction step. In contrast, for the proposed scheme with robust horizon $N_r = 0$, only $18 \times 4 = 72$ constraints are required to formulate the propagation of the tubes. This is because the small uncertainties are not considered both in the computation of contractive sets and in the online problem. Instead, a suitable back-off is obtained by making use of the disturbance invariant set S. If the low complexity tube is employed, the complexity of the tube-based MPC scheme without the invariant tube is 16 times larger than the proposed scheme with $N_r = 0$ because of the 16 vertices of the additive disturbance bounds.



(A) Projection on ΔC_a, ΔC_b and ΔT_R.

(B) Projection on ΔC_b, ΔT_R and ΔT_J.

FIGURE 5 Comparison of the feasible domains obtained by the tube-based NMPC (without the invariant tube S) and the proposed scheme with $N_r = 0$ [Colour figure can be viewed at wileyonlinelibrary.com]

The feasible domains of the tube-based MPC scheme (without the invariant tube) and the proposed tube-enhanced multi-stage MPC scheme with $N_r = 0$ obtained using the general complexity tube predicted tubes are shown in Figure 5. It can be seen that the feasible region of the tube-based MPC scheme that does not employ the invariant tube is larger than the feasible region of the proposed scheme with $N_r = 0$. The volume of the feasible domain of tube-based MPC scheme (without the invariant tube) is 1197.1 whereas the proposed scheme with $N_r = 0$ results in a volume of 1110.7 (approx. 7% smaller). However, the computation time of the proposed approach with $N_r = 0$ is only 0.15 seconds compared to the other scheme which has a computation time of 6.55 seconds. If the homothetic tubes are employed, the feasible domain of the proposed scheme is 6% smaller while the computation time of the proposed approach is approximately 96% smaller (for exact values refer to Table 2). In the case of low complexity tube, there is no difference in the volumes of feasible domains observed.

With this example, it can be seen that the invariant set can introduce a certain conservatism in the closed-loop. However, it reduces significantly the computational complexity and improves the computation time of the approach largely (up to 98% reduction). Thanks to the use an invariant tube for small uncertainties, an important reduction in computation time can be expected at the cost of only minor additional conservativeness.

6.3 | The effect of the robust horizon

If the proposed tube-enhanced multi-stage approach is employed to handle large uncertainties, the number of scenarios considered by the optimization problem increases with the length of robust horizon. Total number of scenarios in the problem is determined by N_r. For $N_r = 1$, the problem has only four scenarios and for $N_r = 5$, the problem has $4^5 = 1024$ scenarios. First, we discuss the effect of the types of the predicted tubes that are employed and then summarize the observations.

6.3.1 | General complexity tube

The scheme with a robust horizon $N_r = 4$ has the same feasible domain as the scheme with full robust horizon. When the robust horizon decreases further, the volume of the feasible domain decreases monotonically. The comparison of the feasible domains of the proposed scheme with different robust horizons that uses general complexity tubes are shown in Figure 6. The proposed scheme with $N_r = 1$ has a volume that is approximately 12% smaller than the volume of the full horizon case. The volumes of the feasible domains of the tube-based MPC implemented without the invariant tube

and the proposed scheme with $N_r = 0$ are significantly smaller (approx. 74% smaller and 76% smaller). In the tube-based MPC scheme, the optimization problem has one feed-forward term per prediction step as degrees of freedom. Whereas, in the cased of $N_r = 1$, there are four feed-forward terms optimized at every stage. This improves the degrees of freedom of the controller and results in an improved performance. The polytopes are plotted with the help of the Multi-Parametric Toolbox.[27]

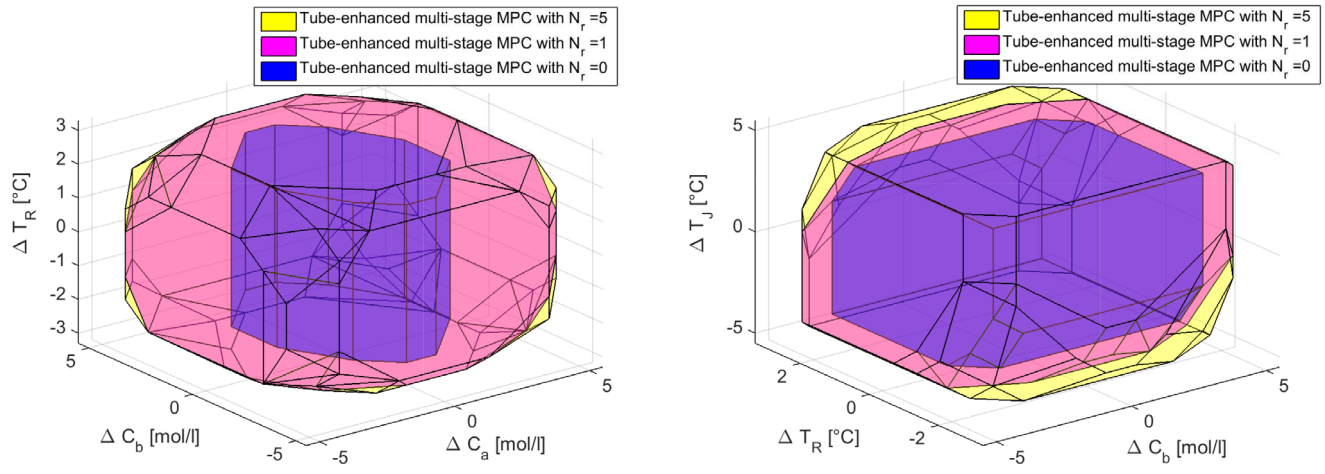
The computation times of the scheme with different robust horizons however does not show a uniform trend. The computation times of robust horizons $N_r = 0$ and $N_r = 1$ are smaller than the full robust horizon. However, the schemes with the robust horizons 2 to 4 have computation times larger than the scheme with full robust horizon. This is because of the difference in complexity associated with the tree and the tube. The scheme with $N_r = 2$ has 16 scenarios. However, from the second prediction step, the propagation of tubes is characterized by 72 inequalities and this is formulated for all 16 scenarios requiring $72 \times 16 = 1152$ inequalities after the second prediction step. This leads to an increased computational effort when compared to a full tree with $4 \times 4 = 16$ equality constraints per node (though exponentially increasing every stage).

6.3.2 | Homothetic tube

The trends in the volumes of the feasible domain and the computation times are similar to the general complexity tube case, if the homothetic tubes are employed in the predictions. However, the scheme results in an increased conservatism compared to the general complexity tube case. This is expected because the shape of the predicted sets is restricted. The proposed scheme with $N_r = 1$ has a volume of the feasible domain that is smaller by approximately 17% in this case compared to the full robust horizon case. The feasible domains of the proposed scheme implemented with homothetic tubes are given in Figure 7. There is also an increase in computation times when compared to the general complexity tube case for a fixed horizon. This is because of the increase in the number of constraints due to 44 vertices of the tube considered to obtain the tight upper bound of the stage cost. This leads to a proportional increase in computational cost.

6.3.3 | Low complexity tube

The computation times are smaller in this case as expected because the tube is represented using the minimal number of inequalities. There is however large conservatism as a result. When $N_r = 0$ is employed, the volume of the feasible region is only 96.02 which is more than 10 times smaller than for the homothetic tube case. Similar reductions in the feasible



(A) Projection on ΔC_a, ΔC_b and ΔT_R.

(B) Projection on ΔC_b, ΔT_R and ΔT_J.

FIGURE 6 Comparison of the feasible domains obtained using the proposed scheme that uses general complexity tubes for varying robust horizons [Colour figure can be viewed at wileyonlinelibrary.com]

domains can be observed across all robust horizons. The feasible domains of the proposed scheme implemented with the Low complexity tube are given in Figure 8. An interesting point to note here is that when $N_r = 1$ is applied, the volume of the feasible domain is much smaller for the same robust horizon than when general complexity/homothetic tubes are employed. Despite this reduction, the volume is larger than the best feasible domain observed for the tube-based MPC. This clearly demonstrates the advantages of the proposed approach. For robust horizon $N_r = 2$, the feasible domain is larger and the computational cost is still lower than tube-based MPC. Hence the scheme can be used to control high dimensional systems, where the pure tube-based scheme is either highly conservative (in the case of using a Low complexity tube) or intractable (in the case of a higher complexity tube). The proposed scheme offers an alternative with an improved trade-off between optimality and complexity.

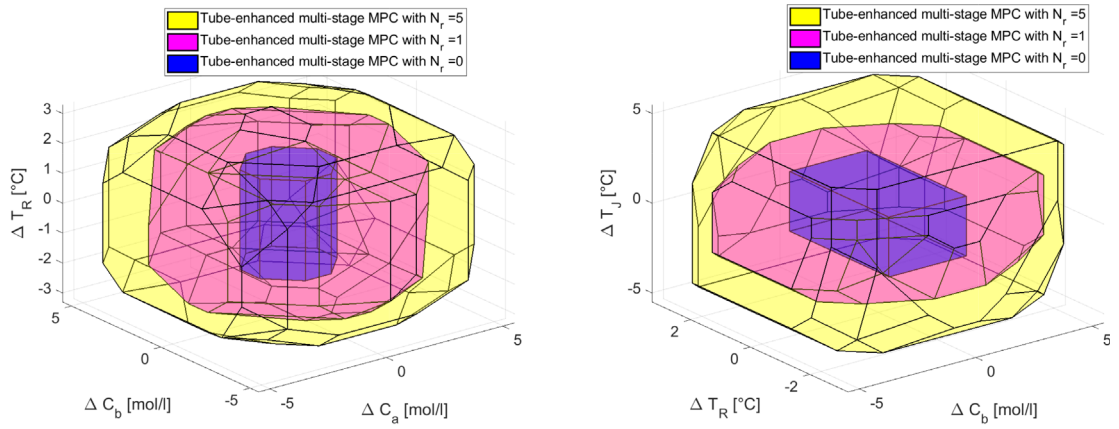
The computation time and the feasible domain for the case $N_r = N_p = 5$ does not depend on the type of tube employed because there is no tube-based part in the predictions of the optimization problems (18) and (23) for this case.

6.3.4 | Summary

Below are the summary of the observations:

1. The volumes of the feasible domains of the proposed scheme increases monotonically with the increase in the robust horizon. For a given robust horizon, the volume of the feasible domain of the proposed scheme that employs the Low complexity tube is lower and that of the general complexity tube is higher. The volume of the scheme that employs a homothetic tube is in between but it is closer to the general complexity tube than to the Low complexity tube.
2. The proposed scheme with $N_r = 1$ has a larger volume of the feasible domain and a smaller computational effort than the tube-based MPC that is employed without the invariant tube. Even the proposed scheme with $N_r = 1$ that employs the low complexity tube has a larger volume than the tube-based MPC that employs a general complexity tube.
3. Though the proposed scheme with Low complexity tube shows conservatism with respect to the scheme with full robust horizon, it gives the best computation times. In addition, the proposed scheme employed with Low complexity tubes with $N_r \geq 1$ shows better performance than the tube-based MPC scheme with complex/general complexity tubes.
4. The scheme offers flexibility with respect to the choice of tube and robust horizon. For example, the proposed scheme that employs a homothetic tube with $N_r = 1$ has a feasible volume and computation times comparable to that of the proposed scheme with Low complexity tube with $N_r = 2$. The scheme offers a wide variety of options to achieve a desired trade-off between optimality and complexity than the existing robust MPC schemes.

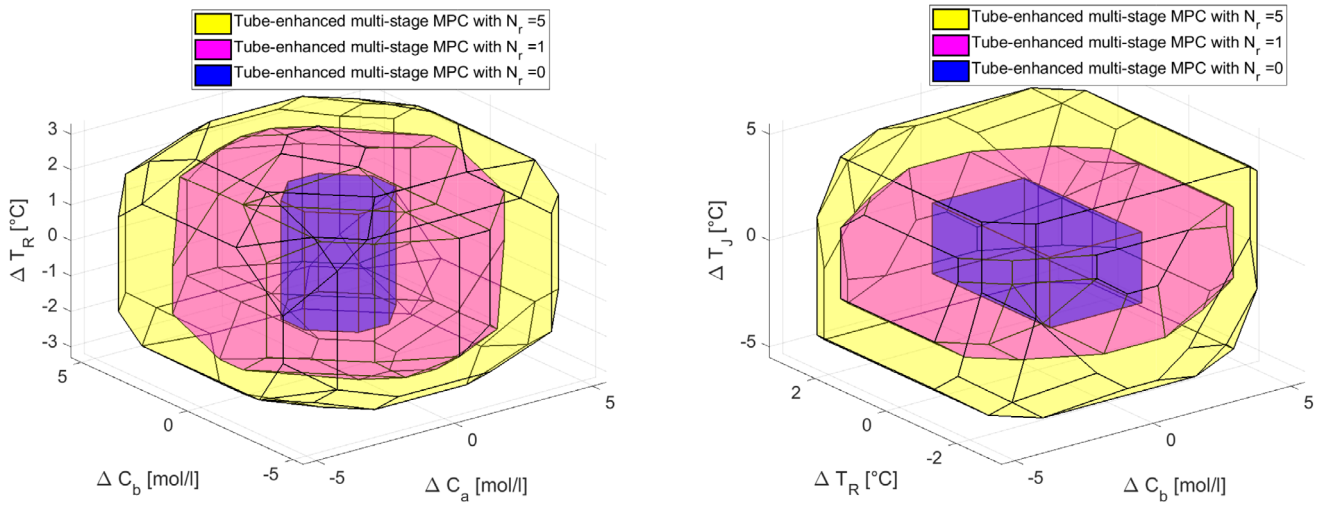
For this example, $N_r = 1$ using a homothetic/general complexity tube, $N_r = 2$ using Low complexity tube and $N_r = 5$ are possible choices to obtain good trade-offs overall in optimality and computational complexity. The scheme with $N_r = 5$



(a) Projection on ΔC_a, ΔC_b and ΔT_R.

(b) Projection on ΔC_b, ΔT_R and ΔT_J.

FIGURE 7 Comparison of the feasible domains obtained using the proposed scheme that uses homothetic tubes for varying robust horizons [Colour figure can be viewed at wileyonlinelibrary.com]



(A) Projection on ΔC_a, ΔC_b and ΔT_R.

(B) Projection on ΔC_b, ΔT_R and ΔT_J.

FIGURE 8 Comparison of the feasible domains obtained using the proposed scheme that uses low complexity homothetic tubes for varying robust horizons [Colour figure can be viewed at wileyonlinelibrary.com]

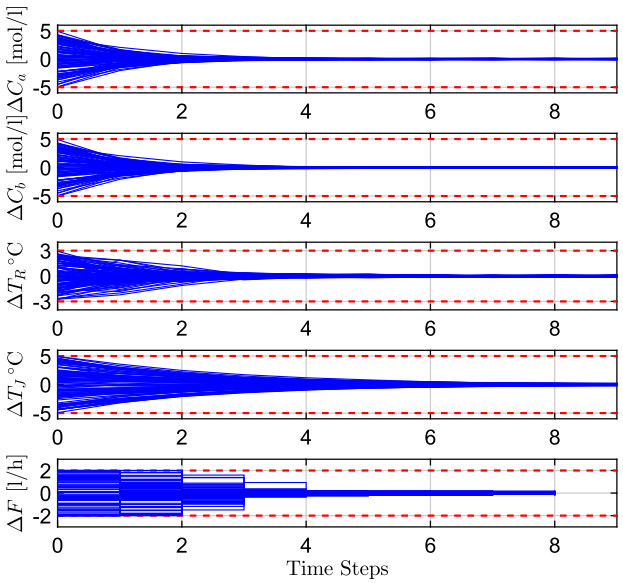


FIGURE 9 Trajectories of ΔC_a, ΔC_b, ΔT_R, ΔF obtained using 100 simulation runs [Colour figure can be viewed at wileyonlinelibrary.com]

gives the best performance for a reasonable computational time. Though the volume of the feasible domain of $N_r = 1$ using general complexity tube is approximately 12% smaller, the computation time is reduced by 62%. Hence $N_r = 1$ using general complexity/homothetic tube is a good choice for this example. The closed-loop state and input trajectories of the proposed scheme employed using a general complexity tube with $N_r = 1$ for random initial conditions and random realizations of the uncertainties for 100 simulation runs is shown in Figure 9.

6.3.5 | The effect of the prediction horizon

The growth of the problem complexity with respect to the prediction horizon is analyzed by comparing the average computation times of the schemes with different robust horizons. The results are plotted in Figure 10. It can be seen that the computational cost increases exponentially if a full robust horizon is used and is significantly larger than for the other robust horizons considered for $N_p = 8$. The computation times of the schemes with $N_r < N_p$ grows linearly in complexity with respect to the prediction horizon. However, the slope is seen increasing when the robust horizon increases. The

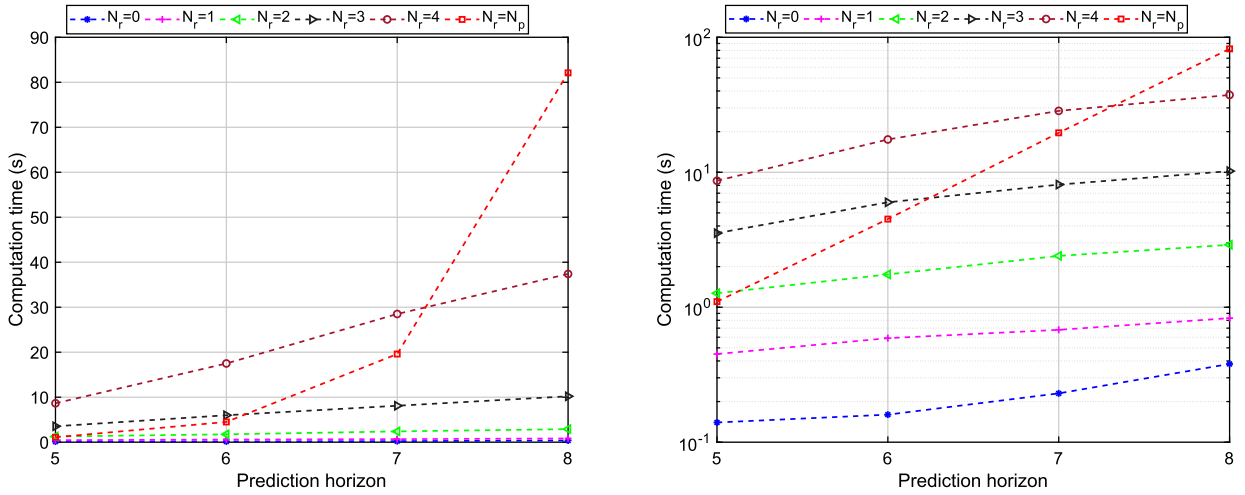


FIGURE 10 Trend of computation times of the proposed scheme for varying different robust horizons [Colour figure can be viewed at wileyonlinelibrary.com]

proposed scheme with robust horizon $N_r = 1$ has a computation time of less than one second and the scheme with $N_r = 2$ has a computation time of less than 3 seconds. The variations in the times are much smaller than in the full robust horizon case. As the horizon grows larger, the computational advantages of the proposed scheme increase.

7 | CONCLUSION

In this article, we have shown that the combination of multi-stage and tube-based MPC schemes offers a flexible framework to manage the trade-off between the performance and computational complexity of the robust scheme. The proposed method uses a tube-based method to handle uncertainties that are small or occur far in the prediction (after the robust horizon), while the multi-stage approach handles the significant and immediate uncertainties to increase the performance. The stability of the scheme for linear systems with parametric and additive disturbances was demonstrated for any choice of robust horizon, including the pure multi-stage case. Simulation results show that the proposed method provides flexibility to obtain a good trade-off between complexity and performance.

ACKNOWLEDGMENTS

The research leading to these results has received funding from the European Commission under grant agreement number 291458 (MOBOCON). RP acknowledges the contribution of the Slovak Research and Development Agency under the project APVV 15-0007.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Sankaranarayanan Subramanian https://orcid.org/0000-0002-9688-2575
Sergio Lucia https://orcid.org/0000-0002-3347-5593

REFERENCES

1. Campo PJ, Morari M. Robust model predictive control. Paper presented at: Proceedings of the American Control Conference, Minneapolis, MN; 1987:1021-1026.
2. Lee JH, Yu ZH. Worst-case formulations of model predictive control for systems with bounded parameters. *Automatica*. 1997;33(5):763-781.
3. Scokaert POM, Mayne DQ. Min-max feedback model predictive control for constrained linear systems. *IEEE Trans Autom Control*. 1998;43(8):1136-1142.

4. Bernardini D, Bemporad A. Scenario-based model predictive control of stochastic constrained linear systems. Paper presented at: Proceedings of the 48th IEEE Conference on Decision and Control, Shanghai, China; 2009; 2009:6333-6338.
5. Lucia S, Finkler T, Engell S. Multi-stage nonlinear model predictive control applied to a semi-batch polymerization reactor under uncertainty. *J Process Control*. 2013;23:1306-1319.
6. Kothare MV, Balakrishnan V, Morari M. Robust constrained model predictive control using linear matrix inequalities. *Automatica*. 1996;32(10):1361-1379.
7. Lee YI, Kouvaritakis B. Robust receding horizon predictive control for systems with uncertain dynamics and input saturation. *Automatica*. 2000;36(10):1497-1504.
8. Chisci L, Rossiter JA, Zappa G. Systems with persistent disturbances: predictive control with restricted constraints. *Automatica*. 2001;37(7):1019-1028.
9. Löfberg J. Approximations of closed-loop minimax MPC. Paper presented at: Proceedings of the 42nd IEEE Conference on Decision and Control, Maui, HI; 2003:1438-1442.
10. Langson W, Chryssochoos I, Rakovi SV, Mayne DQ. Robust model predictive control using tubes. *Automatica*. 2004;40(1):125-133.
11. Mayne DQ, Seron MM, Rakovic SV. Robust model predictive control of constrained linear systems with bounded disturbances. *Automatica*. 2005;41:219-224.
12. Rakovic SV, Kouvaritakis B, Cannon M, Panos C, Findeisen R. Parameterized tube model predictive control. *IEEE Trans Autom Control*. 2012;57(11):2746-2761.
13. Raković SV, Kouvaritakis B, Findeisen R, Cannon M. Homothetic tube model predictive control. *Automatica*. 2012;48(8):1631-1638.
14. Raković SV, Levine WS, Açıkmeşe B. Elastic tube model predictive control. Paper presented at: Proceedings of the American Control Conference (ACC), Boston, MA; 2016:3594-3599; IEEE.
15. Villanueva ME, Quirynen R, Diehl M, Chachuat B, Houska B. Robust MPC via min-max differential inequalities. *Automatica*. 2017;77:311-321.
16. Fleming J, Kouvaritakis B, Cannon M. Robust tube MPC for linear systems with multiplicative uncertainty. *IEEE Trans Autom Control*. 2015;60(4):1087-1092.
17. Muñoz-Carpintero D, Cannon M, Kouvaritakis B. Robust MPC strategy with optimized polytopic dynamics for linear systems with additive and multiplicative uncertainty. *Syst Control Lett*. 2015;81(Suppl C):34-41.
18. Lee YI, Kouvaritakis B. A linear programming approach to constrained robust predictive control. *IEEE Trans Autom Control*. 2000 Sep;45(9):1765-1770.
19. Blanco TB, Cannon M, Moor BD. On efficient computation of low-complexity controlled invariant sets for uncertain linear systems. *Int J Control*. 2010;83(7):1339-1346.
20. Subramanian S, Lucia S, Birjandi SAB, Paulen R, Engell S. A combined multi-stage and tube-based mpc scheme for constrained linear systems. Paper presented at: Proceedings of the 6th IFAC Conference on Nonlinear Model Predictive Control, Madison, Wisconsin; 2018:577-582.
21. Rawlings JB, Mayne DQ. *Model Predictive Control Theory and Design*. Nob Hill Publishing; 2009.
22. Blanchini F, Miani S. *Set-theoretic Methods in Control*. New York, NY: Springer; 2008.
23. Hadjiyiannis MJ, Goulart PJ, Kuhn D. An efficient method to estimate the suboptimality of affine controllers. *IEEE Trans Autom Control*. 2011;56:2841-2853.
24. Kouramas KI, Rakovic SV, Kerrigan EC, Allwright JC, Mayne DQ. On the minimal robust positively invariant set for linear difference inclusions. Paper presented at: Proceedings of the 44th IEEE Conference on Decision and Control, Seville, Spain; 2005:2296-2301.
25. Lucia S. *Robust Multi-stage Nonlinear Model Predictive Control*. Düren, Germany: Shaker; 2014.
26. Kerrigan EC, Maciejowski JM. Feedback min-max model predictive control using a single linear program: robust stability and the explicit solution. *Int J Robust Nonlinear Control*. 2004;14(4):395-413.
27. Herceg M, Kvasnica M, Jones CN, Morari M. Multi-parametric toolbox 3.0. Paper presented at: Proceedings of the European Control Conference; 2013:502-510; Zürich, Switzerland. http://control.ee.ethz.ch/~mpt.
28. Lu X, Cannon M. Robust adaptive tube model predictive control. Paper presented at: Proceedings of the 2019 American Control Conference (ACC), Philadelphia; 2019:3695-3701; IEEE.
29. Kouvaritakis B, Cannon M. Model predictive control: classical, robust and stochastic. *Advanced Textbooks in Control and Signal Processing*. Switzerland: Springer International Publishing; 2015.
30. Klatt KU, Engell S. Gain-scheduling trajectory control of a continuous stirred tank reactor. *Comput Chem Eng*. 1998;22:491-502.

How to cite this article: Subramanian S, Lucia S, Paulen R, Engell S. Tube-enhanced multi-stage model predictive control for flexible robust control of constrained linear systems with additive and parametric uncertainties. *Int J Robust Nonlinear Control*. 2021;31:4458–4487. https://doi.org/10.1002/rnc.5486

APPENDIX A. PROOF OF LEMMA 5

Because of the choice of the stage costs, the proofs are (45) and (46) are similar to Lemma 4. In here, we show that the reformulation (18) retains these properties.

To prove (45), it is sufficient to show that the stage costs are 0 for a feasible solution for all $x \in \mathbb{Z}_f \oplus \mathsf{S}$. For all $x \in \mathbb{Z}_f \oplus \mathsf{S}$, it can be seen that $z \in \mathbb{Z}_f$ is a feasible point. The terminal control law $K_f z$ keeps the trajectory of the primary controller in the terminal set $\mathbb{Z}_f$. The stage costs associated with the multi-stage part of the scheme are 0, that is, $\ell(z_k^j, v_k^j) = 0$ for all $z_k^j \in \mathbb{Z}_f$, for all $(j,k) \in I_{[\![0,\ N_r-1]\!]}$. For the tube-based part of the scheme, the stage cost is reformulated as in (17a)–(17c). Equivalently, the stage cost can be represented as

$$\max_{\bar{z}_k^j \in \mathsf{Z}_k^j} \ell(\bar{z}_k^j, v_k^j) = \min_{y_k^j \in \mathbb{Z}_f, \eta_k^j, \mu_k^j, \gamma_k^j, \forall (j,k) \in I_{[\![N_r,\ N_p-1]\!]}} \gamma_k^j \tag{A1a}$$

subject to:

$$-\mu_k^j \le (P_Q \tau_k^j - Q y_k^j) \le \mu_k^j, (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{A1b}$$

$$-\eta_k^j \le R v_k^j \le \eta_k^j, (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{A1c}$$

$$\mathbf{1}^T \mu_k^j + \mathbf{1}^T \eta_k^j \le \gamma_k^j, (j,k) \in I_{[\![N_r,\ N_p-1]\!]}. \tag{A1d}$$

For all $x \in \mathbb{Z}_f \oplus \mathsf{S}$ and $N_r = 0$, $\mathsf{Z}_0^j = \mathbb{Z}_f$ is a feasible set for all $(j,0) \in I_0$. If $N_r \ge 1$, the terminal control law keeps the state in the terminal set and hence $z_{N_r}^j \in \mathbb{Z}_f,\ \forall (j, N_r) \in I_{N_r}$. Hence, $\mathsf{Z}_{N_r}^j = \mathbb{Z}_f$ is a feasible set for all $(j, N_r) \in I_{N_r}$. Since the set $\mathbb{Z}_f$ is robustly positive invariant for the control law $K_f \mathsf{Z}_k^j$, $v_k^j = 0$ is a feasible control law for the tube-based part of the scheme. Substituting $v_k^j = 0$, and $Q = P_Q T$ from (17a), the constraints (A1b) and (A1c) can be rewritten as follows:

$$-\mu_k^j \le P_Q(\tau_k^j - T y_k^j) \le \mu_k^j, (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{A2a}$$

$$-\eta_k^j \le R v_k^j \le \eta_k^j, (j,k) \in I_{[\![N_r,\ N_p-1]\!]}, \tag{A2b}$$

Since $\mathsf{Z}_k^j = \mathbb{Z}_f$ is a feasible solution, there exists a feasible solution $y_k^j = z_k^j$ such that $T z_k^j = \tau_k^j$ holds for any $(j,k) \in I_{[\![N_r,\ N_p-1]\!]}$. Hence, $\mu_k^j = \eta_k^j = \gamma_k^j = 0$ is a feasible solution for all $(j,k) \in I_{[\![N_r,\ N_p-1]\!]}$, if $\mathsf{Z}_k^j \subseteq \mathbb{Z}_f$. This implies that the stage costs remain 0 for all $x \in \mathbb{Z}_f \oplus \mathsf{S}$ for the tube-based part of the scheme in addition to the multi-stage part of the scheme for the formulation (18). Hence $V_{N_p}^{G*}(x) = 0,\ \forall x \in \mathbb{Z}_f \oplus \mathsf{S}$.

The proof of (46) is straight forward as shown in Lemma 4. Because of the choice of Q and R matrices as positive definite and that the tube-based part of the scheme over-approximates the cost, $V_{N_p}^{G*}(x) \ge c_2 |z|_{\mathbb{Z}_f},\ \forall x \in \mathbb{X}_{N_p}^G$, where $c_2 \ge c_1$.

To prove (47), the multi-stage part of the scheme follows the same arguments given in the proof of Lemma 4. For the tube-based part, because of the reformulation (17) and choosing the nonnegative matrices offline, we can only establish a sufficient condition online. Hence $J_k^{\text{tube}*}$ solved using (18) will always over-approximate the true solution obtained using the formulation (6). From this, we see that the following inequality holds:

$$\sum_{k=0}^{N_r-1} J_k^{\text{MS}}(x^+) \le \sum_{k=0}^{N_r-1} J_k^{\text{MS}}(x) + J_{N_r}^{\text{tube}*}(x) - \omega_0^1 \ell(z_0^{1*}, v_0^{1*}). \tag{A3}$$

To compare the tube-based part of the scheme, we must establish the following inequality:

$$\sum_{k=N_r}^{N_p-1} J_k^{\text{tube}}(x^+) \le \sum_{k=N_r+1}^{N_p-1} J_k^{\text{tube}*}(x). \tag{A4}$$

for all $(j,k) \in I_{[\![N_r,\ N_p-1]\!]}$. To compare the components of the stage costs at consecutive time steps, let us look at the constraint (A1b) in the optimization problem (A1). Since the predicted tubes at the next time step can be given as the convex combination of the tubes predicted at the current time step, the following holds for all $(j,k) \in I_{[\![N_r,\ N_p-2]\!]}$.

$$|P_Q\tau_k^j(x^+) - Qy_k^j(x^+)| = \left|P_Q\sum_{j=1}^{n_d^{N_r}}\lambda_{k+1}^j\tau_{k+1}^{j*}(x) - Q\sum_{j=1}^{n_d^{N_r}}\lambda_{k+1}^j y_{k+1}^{j*}(x)\right|, \tag{A5a}$$

$$\leq \sum_{j=1}^{n_d^{N_r}}\lambda_{k+1}^j(|P_Q\tau_{k+1}^{j*}(x) - Qy_{k+1}^{j*}(x)|), \tag{A5b}$$

$$\leq \sum_{j=1}^{n_d^{N_r}}\lambda_{k+1}^j(|\mu_{k+1}^{j*}(x)|), \tag{A5c}$$

where λ_k^j denotes the convex weights associated with the tubes for all $(j,k) \in I_{[\![N_r,\ N_p-1]\!]}$. Minkowski's inequality leads to (A5b) from (A5a) and (A5c) follows from (A1b). Similarly, the following holds for the constraint (A1c) in the optimization problem (A1) for all $(j,k) \in I_{[\![N_r,\ N_p-2]\!]}$.

$$|Rv_k^j(x^+)| = \left|R\sum_{j=1}^{n_d^{N_r}}\lambda_{k+1}^j v_{k+1}^{j*}(x)\right| \tag{A6a}$$

$$\leq \sum_{j=1}^{n_d^{N_r}}\lambda_{k+1}^j(|Rv_{k+1}^{j*}(x)|), \tag{A6b}$$

$$\leq \sum_{j=1}^{n_d^{N_r}}\lambda_{k+1}^j(|\eta_{k+1}^{j*}(x)|), \tag{A6c}$$

Substituting (A5) and (A6) in (A1d), there exist a feasible $\gamma_k^j(x^+)$ for all $(j,k) \in I_{[\![N_r,\ N_p-2]\!]}$ such that the following holds:

$$\gamma_k^j(x^+) \leq \sum_{j=1}^{n_d^{N_r}}\lambda_{k+1}^j\gamma_{k+1}^{j*}(x) \tag{A7}$$

Combining (A7) with Assumption 1, we have

$$\sum_{j=1}^{n_d^{N_r}}\omega_k^j\gamma_k^j(x^+) \leq \sum_{j=1}^{n_d^{N_r}}\omega_{k+1}^j\sum_{j=1}^{n_d^{N_r}}\lambda_{k+1}^j\gamma_{k+1}^{j*}(x) \tag{A8}$$

for all $k \in \{N_r,\ \ldots, N_p-2\}$. As shown in (A1), the stage costs remain zero for the additional step. Summing up across the horizon for the tube-based part of the scheme, we get,

$$\sum_{k=N_r}^{N_p-1}J_k^{\text{tube}}(x^+) \leq \sum_{k=N_r+1}^{N_p-1}J_k^{\text{tube}*}(x). \tag{A9}$$

This leads to the condition $V_{N_p}^{G*}(x^+) \leq V_{N_p}^{G*}(x) - c_2|z|_{\mathbb{Z}_f},\ \forall x \in \mathrm{X}_{N_p}^G$. This proves (47).

APPENDIX B. PROOF OF LEMMA 6

First, we prove (48). The multi-stage part of the scheme inherits the same properties discussed in Lemma 4 and 5 and the stage cost $\ell(z_k^j, v_k^j) = 0,\ \forall (j,K) \in I_{[\![0,\ N_r-1]\!]}$, if $x \in \mathbb{Z}_f \oplus \mathsf{S}$. If $N_r = 0$, for all $x \in \mathbb{Z}_f \oplus \mathsf{S}$, $\mathsf{Z}_0^j = \mathbb{Z}_f$ is a feasible set for all $(j,0) \in I_0$. Also, if $N_r \geq 1$, if $z_{N_r}^j \in \mathbb{Z}_f,\ \mathsf{Z}_{N_r}^j = \mathbb{Z}_f$ is a feasible set for each $(j,N_r) \in I_{N_r}$. Since the set $\mathbb{Z}_f$ is robustly positive invariant for the control law $K_{\text{pred}}\mathsf{Z}_k^j, v_k^j = 0$ is a feasible control law for the tube-based part of the scheme. All the vertex points of the tube $\hat{z}_k^j + \alpha\vee_r$ is contained in the set $\mathbb{Z}_f$. For each $\hat{z}_k^j + \alpha\vee_r$, there exists a feasible $y_k^{j,r} \in \mathbb{Z}_f$. Hence, the upper

bound of $|Q(\hat{z}_k^j + \alpha \vee_r - y_k^{j,r})|$ is 0 in (22). Also $|Rv_k^j| \leq 0$ is a feasible solution. Hence, $\gamma_k^j = 0, (j,k) \in I_{[\![N_r,\ N_p-1]\!]}$ is feasible in (22). Since the optimal cost is smaller than or equal to the feasible value, the upper bound of the stage cost is 0. From the definition of the terminal set $(A + BK_{\text{pred}})\mathbb{Z}_f \subseteq \mathbb{Z}_f$, implying $\mathsf{Z}_{k+1}^j \subseteq \mathsf{Z}_k^j$. This leads to the case where the stage cost remains 0 until N_p. Hence if $x \in \mathbb{Z}_f \oplus \mathsf{S}$, $V_{N_p}^{H*}(x) = 0$.

The proof of (49) follows directly from the choice of the stage cost. Since Q and R are positive definite, (49) holds.

To prove (50), the proof of the multi-stage part of the scheme follows the same arguments in Lemma 4. Hence the following inequality holds:

$$\sum_{k=0}^{N_r-1} J_k^{\text{MS}}(x^+) \leq \sum_{k=0}^{N_r-1} J_k^{\text{MS}}(x) + J_{N_r}^{\text{tube}*}(x) - \omega_0^1 \ell(z_0^{1*}, v_0^{1*}). \tag{B1}$$

For the tube-based part of the scheme, the tubes at the next step can be represented as the convex combination of tubes predicted at the previous time step. Hence the following relationship holds $(j,k) \in I_{[\![N_r,\ N_p-2]\!]}$.

$$|Q(\hat{z}_k^j(x^+) + \alpha_k^j \vee_r - y_k^{j,r}(x^+))| = \left| Q \sum_{j=1}^{n_d^{N_r}} \lambda_{k+1}^j \hat{z}_{k+1}^{j*}(x) + \sum_{j=1}^{n_d^{N_r}} \lambda_{k+1}^j \alpha_{k+1}^{j*} \vee_r - \sum_{j=1}^{n_d^{N_r}} \lambda_{k+1}^j y_{k+1}^{j,r*}(x) \right|,\ r \in \{1, \dots, n_v\}, \tag{B2a}$$

$$\leq \sum_{j=1}^{n_d^{N_r}} \lambda_{k+1}^j |Q(\hat{z}_{k+1}^{j*}(x) + \alpha_{k+1}^{j*} \vee_r - y_{k+1}^{j,r*}(x))|,\ r \in \{1, \dots, n_v\}, \tag{B2b}$$

$$\leq \sum_{j=1}^{n_d^{N_r}} \lambda_{k+1}^j |\mu_{k+1}^{j*}|,\ \forall r \in \{1, \dots, n_v\}, \tag{B2c}$$

here λ_k^j denotes the convex weights associated with the predicted tubes for all $(j,k) \in I_{[\![N_r,\ N_p-1]\!]}$. Following the same arguments for the inputs, we arrive at the following inequality

$$|Rv_k^j(x^+)| \leq \sum_{j=1}^{n_d^{N_r}} \lambda_{k+1}^j |\eta_{k+1}^{j*}|, \tag{B3}$$

for all $(j,k) \in I_{[\![N_r,\ N_p-2]\!]}$. Combining (B2) and (B3) in (22e), and summing up across the horizon for the tube-based part of the scheme, we get,

$$\sum_{k=N_r}^{N_p-1} J_k^{\text{tube}}(x^+) \leq \sum_{k=N_r+1}^{N_p-1} J_k^{\text{tube}*}(x). \tag{B4}$$

This leads to the condition $V_{N_p}^{H*}(x^+) \leq V_{N_p}^{H*}(x) - c_3|z|_{\mathbb{Z}_f},\ \forall x \in \mathrm{X}_{N_p}^H$. This proves (50).

Market-Based Coordination of Shared Resources in Cyber-physical Production Sites

Simon Wenzel[1,*], Radoslav Paulen[1], Benedikt Beisheim[1,2], Stefan Krämer[2], and Sebastian Engell[1]

DOI: 10.1002/cite.201700007

Large interconnected production sites typically consist of many plants that are physically coupled by networks of shared resources. An optimal site-wide allocation of shared resources is key to a resource-, energy-, and cost-optimal operation. Market-based coordination can find the site-wide optimum with limited data exchange between the subsystems and thus ensures confidentiality. In this contribution, market-based coordination is applied to a simulation study of a petrochemical production site with steam and intermediate products as shared resources.

Keywords: Distributed optimization, Integrated production sites, Market-based coordination, Resource efficiency, Shared-resource allocation

Received: January 20, 2017; *revised:* February 24, 2017; *accepted:* March 16, 2017

1 Introduction

In process industry, large integrated chemical and petrochemical production sites, such as the production site of INEOS in Cologne [1], usually comprise many individual processing plants, which are physically coupled by shared-resource networks through which flows of material and energy are exchanged. In addition, most of the plants are equipped with local computer systems that are used to optimize the operation of the individual plants with different degrees of detail. Together with the control and information systems, a production site is a large-scale cyber-physical system of systems [2, 3].

Common examples for shared resources are electricity, steam on different pressure levels, pressurized air, or intermediate products. For most of the shared resources there is very limited storage and buffer capacity at the production site. Thus, an optimal allocation of the shared resources among the different production plants is crucial for a site-wide resource- and energy-optimal operation, which leads to a reduction of the overall operating costs. From the perspective of the site management, a cost minimization problem needs to be solved that includes the constraint of balanced networks.

If detailed models of every processing plant and its economic performance are available, the optimal operating point of the complete site can be found by a centralized optimizer. In practice, however, distributed solutions are favored for various reasons. One reason is confidentiality with respect to the plants economics, which must be respected if the different plants belong to different business units or different companies that are connected by networks of shared resources [1]. This situation is illustrated in Fig. 1. It can be seen that, in the automation hierarchy, coordination has to be performed in the upper layers of the enterprise resource planning (ERP), the manufacturing execution systems (MES), and the supervisory control and data acquisition (SCADA). The individual production plants as well as resource and energy providers are linked by a common grid. However, there is no central management of production and consumption and the plants have partial autonomy in terms of making their own decisions following individual objectives. In order to achieve site-wide optimality, a coordinator can be implemented to balance the networks. Such distributed schemes preserve the autonomy of the subsystems [4].

In this contribution, it is shown how distributed coordination, based on market theory, can be used to optimize the overall performance of the site while preserving confidentiality of the details of the economics and the behavior of the plants. The mechanism is illustrated for two different coordination schemes in a simulation study based on the petrochemical production site of INEOS in Cologne.

2 Site-Wide Resource Allocation

In the following, first, the optimization of a single production plant will be explained. Afterwards, the linking of the

[1]Simon Wenzel, Dr. Radoslav Paulen, Benedikt Beisheim, Prof. Dr.-Ing. Sebastian Engell
simon.wenzel@bci.tu-dortmund.de
TU Dortmund, Department of Biochemical, Chemical Engineering, Process Dynamics, Operations Group, Emil-Figge-Straße 70, 44227 Dortmund, Germany.

[2]Benedikt Beisheim, Dr.-Ing. Stefan Krämer
INEOS Köln GmbH, Alte Straße 201, 50769 Köln, Germany.

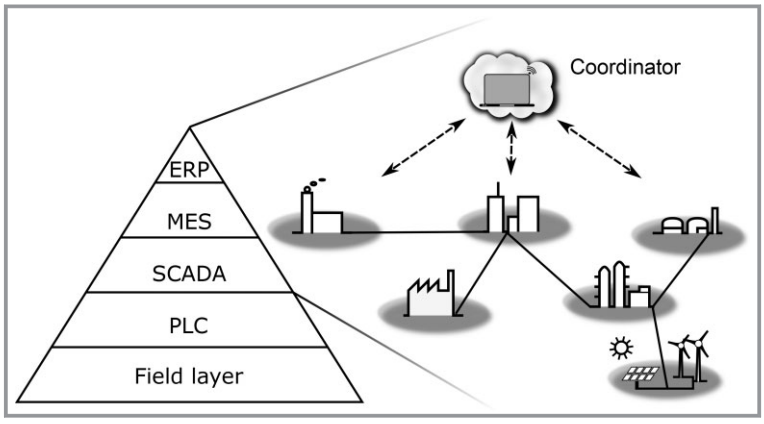


Figure 1. Illustration of the topology of distributed production plants that are coupled by a common distribution grid.

individual plants via shared resources is described, which leads to the formulation of a resource-constrained site-wide optimization problem.

2.1 Optimization of a Single Plant

The operation of an individual production plant within a production site can be optimized by minimizing its operational costs at a given production rate and respecting its individual constraints. The minimization problem for plant *i* can be stated as follows

$$\min_{\mathbf{u}_i \in U_i} J_i(\mathbf{u}_i) \tag{1}$$

where J_i: $R^{n_{u_i}} \rightarrow R$ is an economic cost function and $\mathbf{u}_i \in R^{n_{u_i}}$ is the vector of manipulated variables of the individual plant. The decisions of the plant are restricted by bounds on the manipulated variables, which result from safety regulations, technical constraints, and process limitations. The resulting constraint set is denoted by $U_i \subset R^{n_{u_i}}$. The formulation in Eq. (1) can easily be extended to incorporate equality and inequality constraints on the dependent variables.

2.2 Shared Resources

It is assumed that every plant in the production site is connected to at least one of the shared-resource networks. It is possible that some resources may be shared among only a subset of the plants. The individual production plants can be consumers or producers of shared resources. For instance, a steam cracker plant complex consumes high pressure steam and produces the intermediate fractions ethylene (C2) and propylene (C3). Each production plant is assigned a shared-resource vector $\mathbf{r}_i \in R^{n_r}$ consisting of two parts

$$\mathbf{r}_i(\mathbf{u}_i) = \mathbf{f}_i(\mathbf{u}_i) + \mathbf{r}_{0,i}, \quad \mathbf{f}_i : \mathbb{R}^{n_{u_i}} \rightarrow \mathbb{R}^{n_r} \tag{2}$$

One part depends on the decisions of the plant, $\mathbf{u}_i \in R^{n_{u_i}}$, and can be changed by variation of $\mathbf{u}_i$. The other part represents constant flows over the boundaries of the balance space of the site, $\mathbf{r}_{0,i} \in R^{n_r}$. The shared-resource vector contains an entry for each possible inlet and outlet stream to the shared-resource network that exists at the site. If the plant is not connected to one of the available networks, the respective entry is set to zero. Shared resources that are produced by a plant are sent to the connected shared-resource network and are considered as negative flows. Shared resources that are consumed by a plant are indicated by a positive sign. An example of two processing plants that are coupled by two resource networks is shown in Fig. 2.

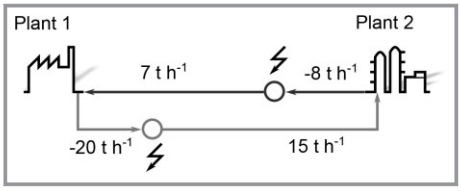


Figure 2. Example for two production plants coupled by two resources. The shared-resource networks are not balanced.

The production and consumption rates of the two production plants are determined by their individual cost optimization if there is no coordination. Then, the produced and consumed amounts of resources most likely do not match. The stationary balance of the networks at a site with *N* production plants can be expressed as

$$\sum_{i=1}^{N} \mathbf{r}_i(\mathbf{u}_i) = \sum_{i=1}^{N} \mathbf{f}_i(\mathbf{u}_i) + \mathbf{r}_{0,i} : \mathbb{R}^{n_{u_1}} \times \mathbb{R}^{n_{u_2}} \times \ldots \times \mathbb{R}^{n_{u_N}} \rightarrow \mathbb{R}^{n_r} \tag{3}$$

If the sum of the produced amounts of resources, the consumed amounts of resources, and the imported and exported amounts do not equalize, the networks are unbalanced. In practice, this situation has to be avoided since it either causes losses of energy and material, e.g., excess steam has to be vented out, or it results in technical infeasibilities as there is no significant storage. Hence, when optimizing the complete production site, the constraint of balanced networks has to be taken into account. In current practice, procedures for solving site-wide infeasibilities are available. These solutions usually rely on the knowledge of experienced production planners and require iterative adjustments of the production plans to react to unforeseen deviations.

2.3 Resource-Constrained Optimization of the Site

The site-wide cost optimization has to consider all individual plants and the balance of the shared-resource networks.

The following resource-constrained cost minimization problem can be formulated

$$\min_{\mathbf{u}_i \in U_i, \forall i} \sum_{i=1}^{N} J_i(\mathbf{u}_i) \quad \text{s.t.} \sum_{i=1}^{N} \mathbf{r}_i(\mathbf{u}_i) = 0 \tag{4}$$

where the network balance in Eq. (4) is referred to as a network or complicating constraint. If the problem is feasible, then the networks are balanced at the optimal solution of Eq. (4) and the total cost of the site is at its minimum.

3 Distributed Coordination

The site-wide optimization problem in Eq. (4) can be solved, if the models, objective functions, and constraint functions of each plant are available to a single optimizer. This precondition, however, is not always satisfied in reality. In the following, a brief overview of distributed techniques is given and the reasons for a distributed optimization and its preconditions are discussed. Afterwards the optimization problem is reformulated to a distributed coordination scheme. Finally, a market-based coordination algorithm to solve the distributed problem is described.

3.1 Distributed Optimization Techniques

Many different distributed coordination schemes have been developed with diverse targets and various applications in mind. For example, coordination algorithms based on non-cooperative and cooperative games are used for the management of the charging of electric vehicles [5], coalitions of individual model predictive controllers (MPCs) can be formed for decentralized control [6], and agents in multi-agent systems can communicate with each other to reach consensus to pursue a common goal [7]. Market-based techniques employ a coordinator that represents the invisible hand of the market. In this approach, central incentives are broadcasted that influence the decisions of the individual subsystems [8, 9].

3.2 Reasons for Distributed Optimization

In general, solving a centralized optimization problem is preferred as, e.g., a centralized solver is able to exploit the structure of the problem. In practice, however, there are various situations in which a solution by a central optimizer cannot be implemented, e.g., when the size of the problem is too large to be solved by a single computer. This is the case when the degree of detail is large and the simulation of the process models of the individual plants is computationally demanding. In addition, centralized optimization is vulnerable to missing information on process models and constraints. Regarding scalability and modularity, adding components or switching components off represents another difficulty since the centralized solution needs to be reconfigured for this case. Last but not least, organizational reasons may prohibit the use of a central computing entity that has full knowledge of every process model, its constraints, and its economic evaluation. Large interconnected production sites or clusters in which different business units or even companies operate in a joint distribution network for shared resources are examples for the need of distributed optimization due to confidentiality reasons [1].

3.3 Preconditions for Distributed Optimization

For the realization of distributed optimization, certain preconditions are necessary. Depending on the chosen distributed optimization algorithm, more or less communication is required either between a central entity and the subsystems or between the distributed systems themselves. The decisions of the individual systems have to depend on the communicated information, otherwise the individual subsystems cannot be steered into a direction that is favorable for the site. Hence, in the problem definition for the optimization of a subsystem, cost information or constraints of neighboring systems or communicated references of a central coordinator have to be considered.

3.4 Market-Based Coordination

A standard approach to distribute an optimization problem of the form given in Eq. (4) is the Lagrangian relaxation. In the Lagrangian relaxation, the constraint functions are multiplied with the so-called Lagrange multipliers and are added to the objective function. In the following, the network constraint is relaxed and added to the objective function. The site-wide optimization problem can thus be formulated as

$$\min_{\mathbf{u}_i \in U_i, \forall i} \sum_{i=1}^{N} J_i(\mathbf{u}_i) + \lambda^{\mathrm{T}} \sum_{i=1}^{N} \mathbf{r}_i(\mathbf{u}_i) \tag{5}$$

which is distributable in the decision variables $\mathbf{u}_i$ and where $\lambda \in R^{n_r}$ is the vector of Lagrange multipliers. For positive values of λ in Eq. (5), the relaxed constraint can be interpreted as an additional cost if a shared resource is consumed and as a reward if a shared resource is produced. For negative prices the situation is reversed. The interpretation of Eq. (5) gives rise to the application of market-based coordination algorithms to compute the optimal solution, which is the set of optimal decisions of the individual plants $\mathbf{u}_i^*$ and the optimal vector of Lagrange multipliers λ^*.

Market-based coordination is inspired by market theory [10]. One example is the coordination of an auction. In the auction, an auctioneer iteratively adjusts prices for traded

goods until the demands match the supplies and the equilibrium price of the market has been found. This process is called tatônnement process [11]. During the auction, the participating agents adjust their produced and consumed amounts of goods according to the prices and announce their planned consumption or production. This procedure can be used to formulate a distributed algorithm to solve the problem given by Eq. (5) in a distributed fashion. In each iteration k, plant i optimizes its own operation based on a given vector of prices λ^k to find its optimal decisions

$$\mathbf{u}_i^{*,k} = \underset{u_i \in U_i}{\arg\min}\, J_i(\mathbf{u}_i) + \lambda^{k,T}\mathbf{r}_i(\mathbf{u}_i) \tag{6}$$

Based on the optimal decisions of the plants, a central coordinator collects the resulting shared-resource vectors for the consumption or the production of the plants

$$\mathbf{r}_i^{*,k}\left(\mathbf{u}_i^{*,k}\right) = \mathbf{f}_i\left(\mathbf{u}_i^{*,k}\right) + \mathbf{r}_{0,i} \tag{7}$$

From Eq. (3), the coordinator evaluates whether the networks are balanced. In case the networks are not balanced, i.e., the demands do not meet the supplies, the price vector has to be updated. The simplest approach is to use a so-called subgradient-based price update, which is defined as follows

$$\lambda^{k+1} = \lambda^k + \mathbf{S}^{k,T}\sum_{i=1}^{N}\mathbf{r}_i^{*,k}\left(\mathbf{u}_i^{*,k}\right) = \lambda^k + \mathbf{S}^{k,T}\mathbf{g}^k \tag{8}$$

where $\mathbf{S}^k \in R^{n_r} \times R^{n_r}$ is a diagonal matrix that determines the update step size and $\mathbf{g}^k \in R^{n_r}$ is a subgradient [12]. A subgradient is a generalization of the gradient of the objective function in Eq. (5) with respect to λ for non-differentiable functions. For the stated problem, it is the value of the constraint in Eq. (4), i.e., the residual of the network balance. The easiest choice to set up the price update step is nonadaptive tatônnement, where $\mathbf{S}^k$ is set to a constant scalar value that is not changed over the index of the iterations. The values that need to be chosen for $\mathbf{S}^k$ require tuning for the specific problem at hand. Without detailed knowledge about the models and the cost functions of the individual subsystems, as it is assumed here, an optimal a priori choice is difficult to find. Due to the lack of knowledge, it is typically started with very small values at the cost of slow rates of convergence. Update steps with varying values of $\mathbf{S}^k$ are referred to as adaptive tatônnement. Different possible choices are discussed, e.g., in [13]. When the prices for the different resources are of different orders of magnitude, an adjustment of the update step according to the average or nominal price vector improves the convergence of the approach [14 – 16].

The market-based algorithm iterates between the updates on the individual plant level and the coordinator as illustrated in Fig. 3. First, the algorithm needs to be initialized with user-defined thresholds for convergence and with an initial price vector for the shared resources. A typical choice for an initial price vector for the shared resources is the vector of current transfer prices, which is available in many sites for billing steam and other resources. The initial price vector is then passed to the subsystem optimizers where the individual plants optimize their cost for the given price vector. The network balance is evaluated using the optimal decisions of the subsystems and a convergence check is performed. If the network balance is satisfied, the algorithm stops. If the balance is not below the threshold, the coordinator performs a price update and communicates the updated price to the subsystems again. The procedure is performed until either convergence is achieved or the maximum number of iterations has been reached. In the course of the price adjustment procedure, the prices can deviate significantly from the nominal transfer prices. Pricing in the public electrical grid, where negative transfer prices are possible, can be considered as an example. Negative prices result from a high excess supply of a particular resource that needs to be balanced by setting high incentives for consumers to increase their demands.

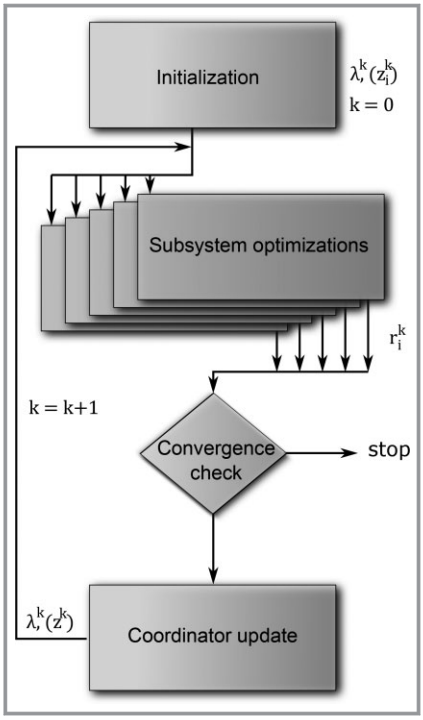


Figure 3. Block diagram of a market-based coordination algorithm.

Note that this approach is of infeasible-path type, i.e., a feasible solution to the problem is found only upon convergence. The advantage of the two-stage coordination scheme using prices is the preservation of confidentiality. The coordinator requires knowledge only about the supplies and demands of the shared-resource vectors to update the prices. The process models, the constraints, and the economics of the individual plants are not revealed. According to [13], the method is guaranteed to converge for a sufficiently small

step size parameter and strictly convex objective functions of the subsystems.

3.5 Extension of the Simple Market-Based Price-Update Scheme

In addition to a simple constraint relaxation in combination with a subgradient-based price update, the individual objective function can be modified to improve the convergence rate of the distributed coordination scheme. To achieve convergence with less iterations and with less assumptions on the mathematical form of the problem stated in Eq. (1), the relaxed objective functions can be augmented with quadratic penalty terms

$$\min_{\mathbf{u}_i \in U_i, \forall i} \sum_{i=1}^{N} J_i(\mathbf{u}_i) + \lambda^{k,T} \sum_{i=1}^{N} \mathbf{r}_i(\mathbf{u}_i) + \frac{\rho^k}{2} \left\| \sum_{i=1}^{N} \mathbf{r}_i(\mathbf{u}_i) \right\|_2^2 \tag{9}$$

where $\rho^k \in R$ is a penalty parameter, which can be adjusted to convexify the objective function such that the above mentioned restriction on J_i can be relaxed. However, the optimization problem in Eq. (9) is not distributable because of the last term, but a reformulation of the constraint can be done where new auxiliary vectors $\mathbf{z}_i$ are introduced. This method is known as the alternating direction method of multipliers (ADMM) [17]

$$\min_{\mathbf{u}_i \in U_i, \forall i} \sum_{i=1}^{N} J_i(\mathbf{u}_i) + \lambda^{k,T} \sum_{i=1}^{N} \mathbf{r}_i(\mathbf{u}_i) + \frac{\rho^k}{2} \sum_{i=1}^{N} \left\| \mathbf{r}_i(\mathbf{u}_i) - \mathbf{z}_i^k \right\|_2^2 \tag{10}$$

The individual plants receive additional vectors $\mathbf{z}_i^k \in R^{n_r}$ together with the prices from the coordinator and solve the following optimization problems in parallel

$$\min_{\mathbf{u}_i \in U_i} J_i(\mathbf{u}_i) + \lambda^{k,T} \mathbf{r}_i(\mathbf{u}_i) + \frac{\rho^k}{2} \left\| \mathbf{r}_i(\mathbf{u}_i) - \mathbf{z}_i^k \right\|_2^2, \quad \forall i \tag{11}$$

where $\mathbf{z}_i^k$ can be interpreted as reference values for the individual plants. On the coordinator level, the prices are updated according to Eq. (8) and the new reference values $\mathbf{z}_i^{k+1}$ are set via the following expression

$$\mathbf{z}_i^{k+1} = \mathbf{r}_i^{k+1} - \mathbf{M}_r^{-1} \sum_{i=1}^{N} \mathbf{r}_i^{*,k+1} \left(\mathbf{u}_i^{*,k+1} \right) \mathbf{M}_z \tag{12}$$

where $\mathbf{M}_r \in R^{n_r} \times R^{n_r}$ is a diagonal matrix used to compute the average shared-resource vector. On the diagonal of $\mathbf{M}_r$ are the numbers of plants connected to a respective network. If all plants are connected to all networks, then each diagonal entry is equal to N. The matrix $\mathbf{M}_z \in R^{n_r} \times R^{n_r}$ is a matrix with zeros and ones, such as only the references for the networks to which a plant is connected are updated. For a fully connected site it is equal to the identity matrix. Note that the reference values $\mathbf{z}_i$ contain information about the overall resource balance of the complete set of plants at the site. However, the details of the individual models and their economic evaluation are still kept undisclosed.

3.6 Implementation Issues

The effort for the computation of the prices and, thus, the runtime of the algorithm are mainly influenced by four factors, the required time for the evaluation of the subsystems for a given price and reference vector, the required time for the update step performed by the coordinator, the communication effort between the subsystems and the coordinator, and the number of iterations (communication rounds) needed to find a solution. Hence, the runtime of the algorithm is case-specific. It can be in the order of seconds, if only simple planning models are involved. However, it can also take hours, if the responses of the subsystems require extensive simulations with large-scale process models. From a practical point of view, it is expected that the difference in the computational effort required by the different coordination algorithms is negligible, since it is small in comparison to the computational effort for the optimizations by the individual optimizers.

Depending on the implementation and the goal of the application, an ad hoc coordination is possible as well as a coordination for the next hours, weeks, or months for medium-term planning. If needed, the individual plants can still react to disturbances on a faster time scale than the coordination. With reactions such as venting or flaring, the operation of the plant can be adjusted, and for fast variations, also the buffer capacities of the networks can be exploited.

4 Shared-Resource Allocation at INEOS in Cologne

The schematic topology of the case study, which covers a part of the site of INEOS in Cologne, is depicted in Fig. 4. The goal of the application of market-based coordination to the case study is to show how four selected shared-resource networks, which are crucial from a site-wide perspective in terms of energy and resource efficiency, can be balanced. In what follows, the site is described in more detail and a scenario is defined for which the site as a whole reacts to price and reference signals that a coordinator announces to steer the individual plants towards site-wide optimal operation.

4.1 Description of the Site

For this simulation study, nine plants are considered that are linked by four shared-resource networks. The 5-bar steam network (s5) connects the power plant with all other processing plants. Thus, this network is the only network

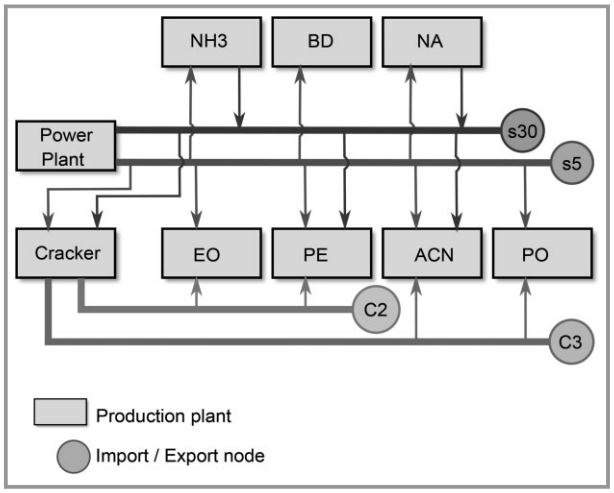


Figure 4. Schematic topology of the case study of INEOS in Köln.

that provides a direct link between all units under consideration. The 30-bar steam network (s30) covers the energy demands of the cracker, the polyethylene (PE) plant, and the acrylonitrile (ACN) plant. The ammonia (NH_3) plant and the nitric acid (NA) plant produce 30-bar steam. The remaining two networks distribute the intermediate products of C2 and C3 from the cracker to downstream processing plants. These two networks connect the smallest subset of the plants in the case study. Import and export of shared resources across the boundaries of the case study are possible and have to be taken into account. These sink and source terms are modeled as an additional plant, which virtually consumes or produces resources. The virtual plant is graphically represented by the circles of the import and export nodes in Fig. 4.

Each plant is represented by a stationary model that is based on the planning models of INEOS in Cologne. The local stationary models are used by optimizers that solve Eq. (6) or Eq. (11), based on given prices of the shared resources and given reference values if ADMM is applied. The objective functions of the plants that are considered for this case study include linear terms that result from the economic evaluation of the performance of the plants and quadratic tracking terms that account for the desired operating points of a plant. The desired operating points result, e.g., from internal contracts or contracts with external suppliers.

4.2 Simulated Scenario

For the simulated scenario, it is assumed that the site is at an optimal operating point with balanced shared-resource networks at the beginning of the simulation. Two changes are assumed to happen at the same time, which influence the operating point of the overall site and result in a new equilibrium price vector for the shared resources. The scenario is shown in Fig. 5. At the event of the change, the exported amount of 30-bar steam over the boundary of the simulated site increases by 20 t h^{-1} and at the same time the capacity of the PE plant drops by 10 %. After the event occurred, the networks are not balanced anymore and the coordination algorithm is triggered. The goal of the coordination is to reduce the imbalance and to steer the plants to a new cost-optimal operating point. The residual of the imbalance is defined as the squared 2-norm of the imbalance of the single networks, where the mass streams are given in t h^{-1}. The solution is considered to be feasible when the residual is below an absolute value of 10^{-2}. In this contribution, the price-update parameter for the subgradient-based method is computed as in [15], while for ADMM the price-update step parameter is adjusted according to [16].

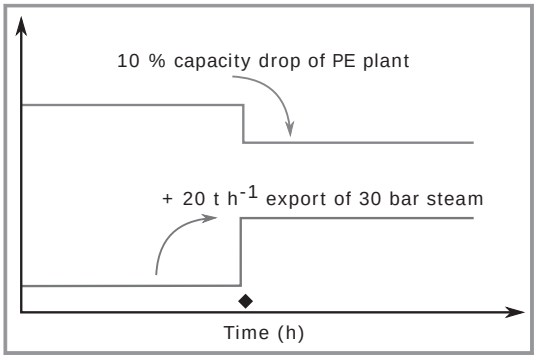


Figure 5. Scenario considered in the simulation study.

4.3 Simulation Results

The simulation results are shown in Figs. 6 – 8. In Fig. 6, the squared 2-norm of the network imbalance is plotted against the number of iterations. Both the subgradient-based price updates and the ADMM price updates are able to perform the coordination task. However, the subgradient-based price updates require a significantly larger number of iterations,

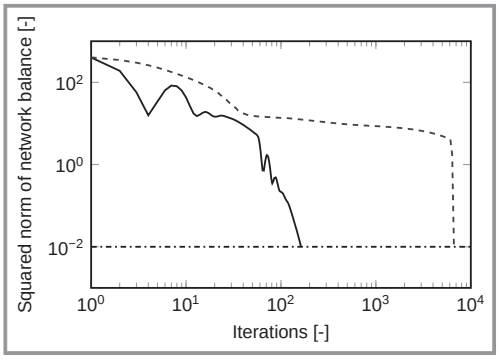


Figure 6. Overall network imbalance against the number of iterations. The dashed lines represent simulation results with subgradient-based updates; the solid line represents the simulation results for ADMM; the dash-dotted line is the threshold for balanced networks.

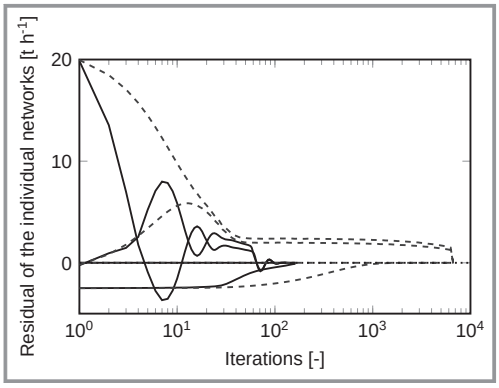


Figure 7. Network imbalance of the four networks against the number of iterations. The dashed lines represent simulation results with subgradient-based updates; the solid lines represent the simulation results for ADMM.

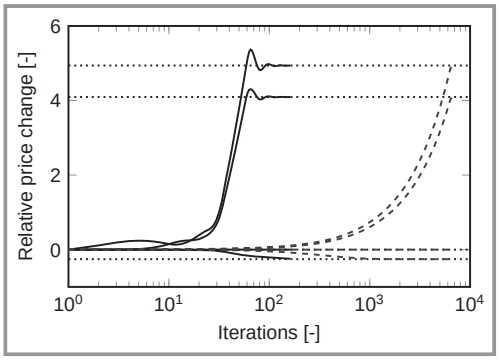


Figure 8. Relative price changes against the number of iterations. The dashed lines represent simulation results with subgradient-based updates; the solid lines represent the simulation results for ADMM; the dotted lines are the centralized solution.

two orders of magnitude larger than for ADMM. An interesting observation is that the residual of the subgradient-based price updates drops suddenly a few iterations before convergence is achieved. The reason for this behavior is that the optima of the production plants show very different sensitivities with respect to the transfer prices. This happens, e.g., when individual plants operate at constraints that determine their optima and, thus, they hardly react to changing transfer prices.

Fig. 7 illustrates that the residual, which is only slowly decreased from iteration 100 in Fig. 6 for the subgradient-based updates, is mainly caused by two networks that require lots of iterations until a balance is achieved. For the coordination with ADMM, the situation is different. The addition of the squared penalty term in Eq. (10) drives all the four networks to a balanced situation in significantly less iterations.

Fig. 8 shows the relative price change for the shared resources in comparison to the initial point over the number of iterations. It can be seen that both coordination schemes converge to the optimal prices that were obtained from the centralized optimization (dotted lines). One price for the shared resources does not change, one price is reduced by 15 %, and two prices increase 4 – 5 times with respect to the initial price. The two increasing prices are those of 5-bar- and 30-bar steam. The drastic changes of the prices show that the individual production plants are rather inflexible in terms of the operating conditions. Only a significant change in the prices causes the individual plants to change their decisions in order to serve the common goal of equilibrated networks.

Tab. 1 shows the reactions of the individual plants to the simulated scenario. Only a subset of the plants contributes to the balancing task. The butadiene (BD), NA, ethylene oxide (EO), ACN, and propylene oxide (PO) plants do not change their operating conditions in this scenario, because their objective functions are not affected by the changing prices such that an adjustment of the operating conditions decreases the overall cost function, or the individual optima are determined by active constraints. For the C2-network, the adjustments required for its rebalancing are only performed by the PE plant, which reduces its capacity, and by the import/export node, where the excess amount of C2 is exported. The increase of 30-bar steam export does not only have a great effect on the 30-bar steam network. It is coupled with the 5-bar steam network such that the resource consumption and production rates of the power plant, the cracker, the NH_3 plant, and the PE plant as well as the import/export node are adjusted. An additional coupling between the shared resources exists in the PE plant, where a reduced amount of C2 consumption leads to a reduced demand of 5-bar steam.

Table 1. Reaction of the production site in the simulated scenario. Upward arrows indicate an increase of production or consumption and downward arrows indicate a decrease; – denote that the shared-resource utilization does not change by an absolute magnitude larger than 0.01; n. c. indicates no connection of the plant to the network.

Plant	s5	s30	C2	C3
Power plant	↓	↑	n. c.	n. c.
Cracker	↑	↓	–	–
NH_3	↓	↓	n. c.	n. c.
BD	–	n. c.	n. c.	n. c.
NA	–	–	n. c.	n. c.
EO	–	n. c.	–	n. c.
PE	↓	–	↓	n. c.
ACN	–	–	n. c.	–
PO	–	n. c.	n. c.	–
Import/export	↓	↑	↑	–

5 Conclusion and Future Work

In this contribution, the application of market-based coordination algorithms to a case study of the integrated petrochemical production site of INEOS in Cologne has been demonstrated. It was shown that both investigated algorithms converge to the centralized solution and are thus able to balance the networks for four shared resources that connect nine individual processing plants. The distributed solutions ensure confidentiality of the economic data of the different plants for both approaches. The degree of autonomy granted to the subsystems, however, differs slightly due to the amount of data that is shared between the central coordinator and the subsystems. For practical applications, the use of the simple subgradient-based price updates is appealing due to its simplicity and due to the necessity to only share prices. But the drawback of this coordination scheme is the slow rate of convergence and thus the large number of iterations until a feasible solution is found. The rate of convergence can be improved by the use of ADMM with additionally shared variables that enable a convexification of the optimization problem. However, this comes at the need of sharing of reference values. It has to be ensured that these reference values are taken into account by the individual optimizers of the subsystems.

For both algorithms it could be demonstrated how the overall production site adjusts its operating conditions to cope with changing conditions inside the production site and with changing demands from the outside. The simulated scenario can be extended to take changing market prices for electricity or changing product and raw material prices into account. Another interesting aspect is the study of scenarios in which the scarcity of resources, due to a partial failure of entities in the network, is handled by market-based coordination.

Future research should aim at a further reduction of the number of iterations that are needed to find the equilibrium price vector, which corresponds to the same number of communication rounds between the coordinator and the individual plants and local optimizations.

The authors gratefully acknowledge the support of the European Commission under the grant agreement numbers 611281 and 723575 (FP7-ICT project DYMASOS (www.dymasos.eu) and H2020-EU project CoPro (www.spire2030.eu/copro)).

Symbols used

$\mathbf{f}$	[t h^{-1}]	function for the mapping of plant inputs to shared resources
$\mathbf{g}$	[t h^{-1}]	subgradient
J	[€ h^{-1}]	objective function
$\mathbf{M}_\mathrm{r}$	[–]	matrix to compute the average resource vector
$\mathbf{M}_\mathrm{z}$	[–]	selection matrix for connected resource networks
N	[–]	number of plants
n_ui	[–]	number of manipulated variables of the ith plant
n_r	[–]	number of shared resources on the site
$\mathbf{r}$	[t h^{-1}]	shared-resource vector
$\mathbf{r}_{0,\mathrm{i}}$	[t h^{-1}]	shared-resource vector for constant import and export
$\mathbf{S}$	[€ h t^{-2}]	diagonal matrix of step size parameter
$\mathbf{u}_\mathrm{i}$	[t h^{-1}]	vector of the manipulated variables of the ith plant
$\mathbf{u}_\mathrm{i}^\star$	[t h^{-1}]	vector of the optimal manipulated variables of the ith plant
U_i	[t h^{-1}]	constraint set of the ith plant
$\mathbf{z}_\mathrm{i}$	[t h^{-1}]	reference value vector for the ith plant

Greek symbols

λ	[€ t^{-1}]	vector of prices for the shared resource
$\lambda^\star$	[€ t^{-1}]	vector of equilibrium prices for the shared resources
ρ	[–]	penalty parameter

Sub- and superscripts

i	plant i
k	iteration k

Abbreviations

ACN	acrylonitrile
ADMM	alternating direction method of multipliers
BD	butadiene
C2	ethylene
C3	propylene
EO	ethylene oxide
ERP	enterprise resource planning
MES	manufacturing execution system
MPC	model-predictive control
NA	nitric acid
PE	polyethylene
PLC	programmable logic control
PO	propylene oxide
s5	5-bar steam network
s30	30-bar steam network
SCADA	supervisory control and data acquisition

References

[1] S. Wenzel, R. Paulen, G. Stojanovski, S. Krämer, B. Beisheim, S. Engell, *Automatisierungstechnik* **2016**, *64 (6)*, 428 – 442. DOI: 10.1515/auto-2016-0003

[2] B. Copigneaux, S. Engell, R. Paulen, M. Reniers, C. Sonntag, H. Thompson, *Proposal of a European Research and Innovation Agenda on Cyber-physical Systems of Systems 2016 – 2025* (Eds: S. Engell, C. Sonntag), TU Dortmund **2016**. www.cpsos.eu/roadmap

[3] S. Engell, R. Paulen, M. A. Reniers, C. Sonntag, H. Thompson, in *Cyber Physical Systems: Design, Modeling, and Evaluation* (Eds: M. Mousavi, C. Berger), Lecture Notes in Computer Science, Vol. 9163, Springer, Cham **2015**. DOI: 10.1007/978-3-319-25141-7_4

[4] N. Gatsis, G. B. Giannakis, *IEEE Trans. Smart Grid* **2013**, *4 (4)*, 1976 – 1987. DOI: 10.1109/TSG.2013.2258179

[5] S. Grammatico, F. Parise, M. Colombino, J. Lygeros, *IEEE Trans. Autom. Control* **2015**, *61 (11)*, 3315 – 3329. DOI: 10.1109/TAC.2015.2513368

[6] F. Fele, J. M. Maestre, S. M. Hashemy, D. M. de la Peña, E. F. Camacho, *J. Process Control* **2014**, *24 (4)*, 314 – 325. DOI: 10.1016/j.jprocont.2014.02.005

[7] Y. Cao, W. Yu, W. Ren, G. Chen, *IEEE Trans. Ind. Inf.* **2013**, *9 (1)*, 427 – 438. DOI: 10.1109/TII.2012.2219061

[8] R. Cheng, J. F. Forbes, W. S. Yip, *J. Process Control* **2007**, *17 (5)*, 429 – 438. DOI: 10.1016/j.jprocont.2006.04.003

[9] R. A. Jose, L. H. Ungar, *AIChE J.* **2000**, *46 (3)*, 575 – 585. DOI: 10.1002/aic.690460316

[10] H. Uzawa, *Econometrica* **1960**, *28 (4)*, 872 – 881. DOI: 10.2307/1907569

[11] D. A. Walker, *J. Polit. Econ.* **1987**, *95 (4)*, 758 – 774. DOI: 10.1086/261484

[12] S. Boyd, L. Vandenberghe, *Convex Optimization*, Cambridge University Press, Cambridge **2004**.

[13] D. P. Bertsekas, *Convex Optimization Theory*, 1st ed., Athena Scientific, Belmont, MA **2009**, Ch. 6.

[14] E. Kutanoglu, S. D. Wu, *IIE Trans.* **1999**, *31 (9)*, 813 – 826. DOI: 10.1080/07408179908969883

[15] H. C. Lau, S. F. Cheng, T. Y. Leong, J. H. Park, Z. Zhao, in *IEEE/WIC/ACM Int. Conf. on Intelligent Agent Technology*, IEEE, Fremont, CA **2007**, 407 – 411. DOI: 10.1109/IAT.2007.41

[16] M. L. Fisher, *Interfaces* **1985**, *15 (2)*, 10 – 21. DOI: 10.1287/inte.15.2.10

[17] S. Boyd, N. Parikh, E. Chu, B. Paleato, J. Eckstein, *Found. Trends Mach. Learn.* **2011**, *3 (1)*, 1 – 122. DOI: 10.1561/2200000016

A collection of the most important citations

R. Martí [35%] – S. Lucia [25%] – D. Sarabia [10%] – R. Paulen [20%] – S. Engell [5%] – C. de Prada [5%]: Improving scenario decomposition algorithms for robust nonlinear model predictive control. *Computers & Chemical Engineering*, vol. 79, pp. 30–45, 2015. [35]

1. Holtorf, F. – Mitsos, A. – Biegler, L.T.: Multistage NMPC with on-line generated scenario trees: Application to a semi-batch polymerization process. *Journal of Process Control*, vol. 80, pp. 167-179, 2019.
2. Krishnamoorthy, D. – Foss, B. – Skogestad, S.: A Primal decomposition algorithm for distributed multistage scenario model predictive control. *Journal of Process Control*, vol. 81, pp. 162-171, 2019.
3. Kouzoupis, D. – Klintberg, E. – Diehl, M. – Gros, S.: A dual Newton strategy for scenario decomposition in robust multistage MPC. *International Journal of Robust and Nonlinear Control*, no. 6, vol. 28, pp. 2340-2355, 2018.

Computers and Chemical Engineering 79 (2015) 30–45

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Improving scenario decomposition algorithms for robust nonlinear model predictive control

Rubén Martí[a,*], Sergio Lucia[b], Daniel Sarabia[c], Radoslav Paulen[b], Sebastian Engell[b], César de Prada[a]

[a] Department of Systems Engineering and Automatic Control, University of Valladolid, c/ Real de Burgos s/n, 47011 Valladolid, Spain
[b] Process Dynamics and Operations Group, Technische Universität Dortmund, Emil-Figge-Str. 70, 44221 Dortmund, Germany
[c] Department of Electromechanical Engineering, Escuela Politécnica Superior, University of Burgos, Avda. Cantabria s/n, Spain

ARTICLE INFO

Article history:
Received 13 December 2014
Received in revised form 29 March 2015
Accepted 24 April 2015
Available online 5 May 2015

Keywords:
Economic model predictive control
Uncertainty
Robust control
Distributed computing
Optimization

ABSTRACT

This paper deals with the efficient computation of solutions of robust nonlinear model predictive control problems that are formulated using multi-stage stochastic programming via the generation of a scenario tree. Such a formulation makes it possible to consider explicitly the concept of recourse, which is inherent to any receding horizon approach, but it results in large-scale optimization problems. One possibility to solve these problems in an efficient manner is to decompose the large-scale optimization problem into several subproblems that are iteratively modified and repeatedly solved until a solution to the original problem is achieved. In this paper we review the most common methods used for such decomposition and apply them to solve robust nonlinear model predictive control problems in a distributed fashion. We also propose a novel method to reduce the number of iterations of the coordination algorithm needed for the decomposition methods to converge. The performance of the different approaches is evaluated in extensive simulation studies of two nonlinear case studies.

Journal of Process Control 80 (2019) 167–179

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

Multistage NMPC with on-line generated scenario trees: Application to a semi-batch polymerization process

Flemming Holtorf[a], Alexander Mitsos[a], Lorenz T. Biegler[b,*]

[a] Aachener Verfahrenstechnik, Lehrstuhl für Systemverfahrenstechnik (AVT.SVT), RWTH Aachen University, Forckenbeckstraße 51, 52074 Aachen, Germany
[b] Department of Chemical Engineering, Carnegie Mellon University, 5000 Forbes Ave, Pittsburgh, PA 15213, USA

ARTICLE INFO

Article history:
Received 29 September 2018
Received in revised form 8 April 2019
Accepted 12 May 2019
Available online 13 June 2019

Keywords:
Scenario-tree generation
Robust control
Adaptive control
Parametric model uncertainty
Economic NMPC

ABSTRACT

We present a multistage NMPC scheme with adaptive on-line scenario-tree generation. The scenario tree is assembled from predictions of worst-case uncertainty realizations that are identified based on a first-order approximation of the process model. The key property of the presented approach is that the size of the resultant optimal control problems does not scale directly with the number of uncertain model parameters. We demonstrate the applicability of the approach with an industrially relevant semi-batch polymerization process under parametric model uncertainty and noisy, incomplete state measurements. By allowing to account explicitly for estimation errors, the presented approach yields increased robustness when compared to nominal NMPC and a standard multistage NMPC scheme. Moreover, we investigate a combination of the presented approach with on-line estimation of uncertain model parameters alongside approximation of their confidence region to reduce the uncertainty range and consequently mitigate unnecessary conservatism. The results show that adaptation of model and uncertainty range yields considerable economic benefits without impairing the attained level of robustness for the considered process.

Multistage NMPC as proposed by Lucia et al. [20] alleviates unnecessary conservatism by accounting for the possibility to take recourse actions in future time steps. To that end, it models the uncertainty propagation in time as a scenario tree. Multistage NMPC has been demonstrated to prevent constraint violations in the presence of model uncertainty with low levels of conservatism. This approach has been used in economic and tracking-type NMPC of several semi-batch polymerization processes [21,20,22,14], a batch bio reactor [19], a hydrodesulphurization plant [26], and in drug dosage design [24]. In these contributions, the scenario trees are commonly constructed by including all combinations of upper and lower bounds of the interval-constrained parameters at the

In order to enable robust shrinking horizon NMPC we seek an on-line solution over N time steps for the closed-loop robust optimal control problem (ROCP). Here we formulate (ROCP) for time step k and the current state (estimate) $\boldsymbol{x}(k)$ in the following way:

$$\min_{\{\boldsymbol{\pi}_l\}_{j=k}^{N} \in \mathcal{U},\ \{\boldsymbol{x}_j\}_{j=k}^{N-1}} \ \max_{\boldsymbol{p}\in\mathcal{P}} \phi(\boldsymbol{x}_N(\boldsymbol{p})) \tag{ROCP}$$

$$\text{s.t.} \quad \boldsymbol{x}_{l+1}(\boldsymbol{p}) = \boldsymbol{f}(\boldsymbol{x}_l(\boldsymbol{p}), \boldsymbol{\pi}_l(\boldsymbol{x}_l(\boldsymbol{p})), \boldsymbol{p}), \quad \forall l \in \{k, \ldots, N-1\}, \quad \forall \boldsymbol{p} \in \mathcal{P} \tag{2}$$

$$\boldsymbol{x}(\boldsymbol{p}) = \boldsymbol{x}(k) \quad \forall \boldsymbol{p} \in \mathcal{P} \tag{3}$$

In addition, scenario-specific back-off margins based on bounds on the reachable set of the underlying dynamic system [41,12] can be introduced to guarantee feasibility as proposed by Lucia et al. [23].

Conflict of interest

The authors have no conflict of interest related to this paper.

Appendix A. Supplementary data

[23] S. Lucia, R. Paulen, S. Engell, Multi-stage nonlinear model predictive control with verified robust constraint satisfaction, in: Proceedings of the 2014 IEEE 53rd Annual Conference on Decision and Control (CDC), IEEE, 2014, pp. 2816–2821.
[24] S. Lucia, M. Schliemann-Bullinger, R. Findeisen, E. Bullinger, A set-based optimal control approach for pharmacokinetic/pharmacodynamic drug dosage design, IFAC-PapersOnLine 49 (7) (2016) 797–802.
[25] D.L. Ma, R.D. Braatz, Worst-case analysis of finite-time control policies, IEEE Trans. Control Syst. Technol. 9 (5) (2001) 766–774.
[26] R. Martí, S. Lucia, D. Sarabia, R. Paulen, S. Engell, C. de Prada, Improving scenario decomposition algorithms for robust nonlinear model predictive control, Comput. Chem. Eng. 79 (2015) 30–45.
[27] A. Mesbah, S. Streif, R. Findeisen, R.D. Braatz, Stochastic nonlinear model

Journal of Process Control 81 (2019) 162–171

Contents lists available at ScienceDirect

Journal of Process Control

journal homepage: www.elsevier.com/locate/jprocont

A Primal decomposition algorithm for distributed multistage scenario model predictive control☆

Dinesh Krishnamoorthy[a,*], Bjarne Foss[b], Sigurd Skogestad[a]

[a] Department of Chemical Engineering, Norwegian University of Science and Technology (NTNU), Trondheim, Norway
[b] Department of Engineering Cybernetics, Norwegian University of Science and Technology (NTNU), Trondheim, Norway

ARTICLE INFO

Article history:
Received 25 September 2018
Received in revised form 5 February 2019
Accepted 8 February 2019
Available online 31 July 2019

Keywords:
Model predictive control
Primal decomposition
Distributed optimization
Uncertainty

ABSTRACT

This paper proposes a primal decomposition algorithm for efficient computation of multistage scenario model predictive control, where the future evolution of uncertainty is represented by a scenario tree. This often results in large-scale optimization problems. Since the different scenarios are only coupled via the so-called non-anticipativity constraints, which ensures that the first control input is the same for all the scenarios, the different scenarios can be decomposed into smaller subproblems, and solved iteratively using a master problem to co-ordinate the subproblems. We review the most common scenario decomposition methods, and argue in favour of primal decomposition algorithms, since it ensures feasibility of the non-anticipativity constraints throughout the iterations, which is crucial for closed-loop implementation. We also propose a novel backtracking algorithm to determine a suitable step length in the master problem that ensures feasibility of the nonlinear constraints. The performance of the proposed approach, and the backtracking algorithm is demonstrated using a CSTR case study.

3. Distributed multistage scenario MPC

The multistage scenario MPC problem (2) consists of S independent MPC problems, except for the non-anticipativity constraints (2e), which couple the different scenarios together. Different decomposition approaches can be used to solve the different scenarios (in parallel), and a master problem can be used to co-ordinate the different subproblems.

3.1. Dual decomposition based approaches

Scenario decomposition using dual decomposition is the most common strategy. Here, each scenario subproblem is solved by relaxing the non-anticipativity constraints, see for example [4], [5] and [8].

The scenario optimization problem (2) can be rewritten as,

$$\begin{aligned}\min_{\mathbf{x}_{k,j},\mathbf{u}_{k,j}} \quad & \sum_{j=1}^{S}\left[\omega_j\sum_{k=0}^{N-1}\mathbf{J}(\mathbf{x}_{k,j},\mathbf{u}_{k,j})\right]+\boldsymbol{\lambda}^T\sum_{j=1}^{S}\bar{\mathbf{E}}_j\mathbf{u}_j\\ \text{s.t.}\quad & \mathbf{x}_{k+1,j}=\mathbf{f}(\mathbf{x}_{k,j},\mathbf{u}_{k,j},\boldsymbol{p}_{k,j})\\ & \mathbf{g}(\mathbf{x}_{k,j},\mathbf{u}_{k,j},\boldsymbol{p}_{k,j})\le 0\\ & \mathbf{x}_{0,j}=\hat{\boldsymbol{x}}\\ & \forall j\in\{1,\ldots,S\},\quad \forall k\in\{0,\ldots,N-1\}\end{aligned}\tag{5}$$

where $\boldsymbol{\lambda}\in\mathbb{R}^q$ is the Lagrange multiplier corresponding to the non-anticipativity constraint (2e). It can be seen that (5) is now additively separable in $\mathbf{x}$ and $\mathbf{u}$ and each jth scenario subproblem can be reformulated as a function of $\boldsymbol{\lambda}$ as shown below,

the subgradient $\sum_{j=1}^{S}\bar{\mathbf{E}}_j\mathbf{u}_j$. The solution to the master problem along the descent direction using a suitable step length α can then be expressed by (8), see [9] and [7]. □

Different forms of augmented Lagrangian decomposition methods were also presented in [5], where an additional quadratic penalty term is added to (6) to improve the convergence properties. However, the additional quadratic penalty terms makes the different subproblems nonseparable in $\mathbf{x}$ and $\mathbf{u}$, and hence cannot be solved in parallel. In such cases, the subproblems must be solved sequentially using the alternating directions method of multipliers (ADMM) approach [10].

However, the main challenge of dual decomposition approach is that, relaxing the non-anticipativity constraints may impede real-time closed-loop implementation. In the receding horizon control framework, at each time step, the first control move is implemented in the plant. In multistage scenario MPC, the non-anticipativity constraints ensure that the first control move is equal for all the scenarios. This enables closed-loop implementation. However, if the master problem and subproblems fail to converge within the required sampling time, the non-anticipativity constraints are not satisfied. Consequently, the first control input computed by the different scenarios are different, thus impeding closed-loop implementation.

One way to address this issue is to take a weighted average of the manipulated inputs at the first sample based on the probabilities of the different scenarios [11]. However, this may not be a good approach since the weighted average can lead to an infeasible solution. The authors in [5] proposed to compute an average of the control inputs at the first sample such that the worst-case constraint violation for the local subproblems is minimized, which is given by solving an additional linear programming (LP) problem. In this paper, we instead propose a primal decomposition approach to solve this issue, which always ensures the feasibility of the non-anticipativity constraints

References

[1] P.J. Campo, M. Morari, Robust model predictive control, in: American Control Conference, 1987, IEEE, 1987, pp. 1021–1026.

[2] P. Scokaert, D. Mayne, Min-max feedback model predictive control for constrained linear systems, IEEE Trans. Autom. Control 43 (8) (1998) 1136–1142.

[3] S. Lucia, T. Finkler, S. Engell, Multi-stage nonlinear model predictive control applied to a semi-batch polymerization reactor under uncertainty, J. Process Control 23 (9) (2013) 1306–1319.

[4] S. Lucia, S. Subramanian, S. Engell, Non-conservative robust nonlinear model predictive control via scenario decomposition, in: 2013 IEEE International Conference on Control Applications (CCA), IEEE, 2013, pp. 586–591.

[5] R. Martí, S. Lucia, D. Sarabia, R. Paulen, S. Engell, C. de Prada, Improving scenario decomposition algorithms for robust nonlinear model predictive control, Comput. Chem. Eng. 79 (2015) 30–45.

Received: 4 July 2016 | Revised: 20 July 2017 | Accepted: 16 November 2017

DOI: 10.1002/rnc.4019

RESEARCH ARTICLE

WILEY

A dual Newton strategy for scenario decomposition in robust multistage MPC

D. Kouzoupis[1] | E. Klintberg[2] | M. Diehl[1,3] | S. Gros[4]

[1]Department of Microsystems Engineering (IMTEK), University of Freiburg, Freiburg im Breisgau, Germany

[2]Qamcom Research & Technology, Gothenburg, Sweden

[3]Department of Mathematics, University of Freiburg, Freiburg im Breisgau, Germany

[4]Department of Electrical Engineering, Chalmers University of Technology, Gothenburg, Sweden

Correspondence
D. Kouzoupis, Department of Microsystems Engineering (IMTEK), University of Freiburg, 79110 Freiburg im Breisgau, Germany.
Email: dimitris.kouzoupis@imtek.uni-freiburg.de

Summary

This paper considers the solution of tree-structured quadratic programs as they may arise in multistage model predictive control. In this context, sampling the uncertainty on prescribed decision points gives rise to different scenarios that are linked to each other via the so-called nonanticipativity constraints. Previous work suggests to dualize these constraints and apply Newton's method on the dual problem to achieve a parallelizable scheme. However, it has been observed that the globalization strategy in such an approach can be expensive. To alleviate this problem, we propose to dualize both the nonanticipativity constraints and the dynamics to obtain a computationally cheap globalization. The dual Newton system is then reformulated into small highly structured linear systems that can be solved in parallel to a large extent. The algorithm is complemented by an open-source software implementation that targets embedded optimal control applications.

Computational aspects of robust MPC have been considered, eg, in the work of Hansson[9] where a primal-dual interior point method is proposed for min-max MPC, and in the work of Zeilinger et al[10] where real-time feasibility is achieved via

a tube-based robust MPC formulation together with the early termination of an interior point method. Moreover, several optimization methods have been proposed that are applicable to multistage MPC problems. Other works[11-15] proposed to use tailored interior point methods. Leidereiter et al[16] and Klintberg et al[17] used parallelizable active set methods, whereas other works[18-20] used various decomposition techniques to exploit the intrinsic structure of the problem.

Conference on Decision and Control; 2016; Las Vegas, NV.

18. Birge J. Decomposition and partitioning methods for multistage stochastic linear programs. *Oper Res*. 1985;33:989-1007.
19. Mulvey J, Ruszczyński A. A new scenario decomposition method for large-scale stochastic optimization. *Oper Res*. 1995;43:477-490.
20. Marti R, Lucia S, Sarabia D, Paulen R, Engell S. Improving scenario decomposition algorithms for robust nonlinear model predictive control. *Comput Chem Eng*. 2015;79:30-45.
21. Ferreau HJ, Kozma A, Diehl M. A parallel active-set strategy to solve sparse parametric quadratic programs arising in MPC. *IFAC Nonlin Model Predictive Control Conf*. 2012;45(17):74-79.
22. Frasch JV, Sager S, Diehl M. A parallel quadratic programming method for dynamic optimization problems. *Math Program Comput*. 2015;7:289-329.
23. Liu JWH. The multifrontal method for sparse matrix solution: theory and practice. *SIAM Rev*. 1992;34(1):82-109.

S. Thangavel [75%] – R. Paulen [15%] – S. Engell [10%]: Robust Multi-Stage Nonlinear Model Predictive Control Using Sigma Points. *Processes*, no. 7, vol. 8, pp. 0851, 2020.

4. Soloperto, Raffaele – Muller, Matthias A. – Allgower, Frank: Guaranteed Closed-Loop Learning in Model Predictive Control. *IEEE Transactions on Automatic Control*, no. 2, vol. 68, pp. 991-1006, 2023.
5. Casas, Carlos Andres Elorza – Valipour, Mahshad – Sandoval, Luis A. Ricardez: Multi-scenario and multi-stage robust NMPC with state estimation application on the Tennessee-Eastman process. *Control Engineering Practice*, vol. 139, pp. 105635, 2023.

processes

MDPI

Article

Robust Multi-Stage Nonlinear Model Predictive Control Using Sigma Points

Sakthi Thangavel [1,*], Radoslav Paulen [2] and Sebastian Engell [1]

1 Process Dynamics and Operations Group, Department of Chemical and Biochemical Engineering, Technische Universität Dortmund, Emil-Figge-Strasse 70, 44227 Dortmund, Germany; sebastian.engell@tu-dortmund.de

2 Faculty of Chemical and Food Technology, Slovak University of Technology in Bratislava, 81237 Bratislava, Slovakia; radoslav.paulen@stuba.sk

* Correspondence: sakthi.thangavel@tu-dortmund.de; Tel.: +49-231-755-5341

Received: 2 June 2020; Accepted: 9 July 2020; Published: 16 July 2020

check for updates

Guaranteed Closed-Loop Learning in Model Predictive Control

Raffaele Soloperto, Matthias A. Müller, *Senior Member, IEEE*, and Frank Allgöwer, *Member, IEEE*

analyzed in [5]. During the past few years, several explicit dual adaptive MPC approaches were proposed, see, e.g., [10] and [11]. In [6] an additional term, which increases with the prediction error, was introduced in the cost function. This results in a control law able to indirectly seek for informative

A. Considered System

We consider a nonlinear, time invariant, perturbed discrete-time system

$$x_{t+1} = f_w(x_t, u_t, d_t) \tag{1}$$

[7] A. Mesbah, "Stochastic model predictive control with active uncertainty learning: A survey on dual control," *Annu. Rev. Control*, vol. 45, pp. 107–117, 2018.
[8] A. A. Feldbaum, "Dual control theory. I," *Automat. Remote Control*, vol. 21, no. 9, pp. 1240–1249, 1960.
[9] T. A. N. Heirung, B. E. Ydstie, and B. Foss, "Dual adaptive model predictive control," *Automatica*, vol. 80, pp. 340–348, 2017.
[10] S. Thangavel, S. Lucia, R. Paulen, and S. Engell, "Dual robust nonlinear model predictive control: A multi-stage approach," *J. Process. Control*, vol. 72, pp. 39–51, 2018.
[11] A. Iannelli, M. Khosravi, and R. S. Smith, "Structured exploration in the
[33] R. Soloperto, M. A. Müller, S. Trimpe, and F. Allgöwer, "Learning-based robust model predictive control with state-dependent uncertainty," *IFAC-PapersOnLine*, vol. 51, no. 20, pp. 442–447, 2018.
[34] Y. Lian and C. N. Jones, "Learning feature maps of the Koopman operator: A subspace viewpoint," in *Proc. IEEE 58th Conf. Decis. Control*, 2019, pp. 860–866.
[35] J. Köhler, P. Kötting, R. Soloperto, F. Allgöwer, and M. A. Müller, "A robust adaptive model predictive control framework for nonlinear uncertain systems," *Int. J. Robust Nonlinear Control*, vol. 31, no. 18, pp. 8725–8749, 2021.
[36] J. Köhler, M. A. Müller, and F. Allgöwer, "A novel constraint tightening

Control Engineering Practice 139 (2023) 105635

Contents lists available at ScienceDirect

Control Engineering Practice

journal homepage: www.elsevier.com/locate/conengprac

Multi-scenario and multi-stage robust NMPC with state estimation application on the Tennessee-Eastman process

Carlos Andrés Elorza Casas, Mahshad Valipour, Luis A. Ricardez Sandoval *

Department of Chemical Engineering, University of Waterloo, Waterloo, Canada, N2L 3G1

ARTICLE INFO

Keywords:
Tennessee-Eastman process
Robust NMPC
State estimation
Plant-model mismatch

ABSTRACT

This study presents the implementation of two discrete robust approaches to Non-linear Model Predictive Control (NMPC), multi-scenario NMPC (MSc-NMPC) and multi-stage NMPC (MS-NMPC), to the benchmark Tennessee-Eastman (TE) challenge, with Extended Kalman Filter (EKF) and Moving Horizon Estimation (MHE) as state estimators. The robust NMPC formulation results in closed-loop responses that prevent constraint violation and closely track the process set-point under parameter uncertainty, even in scenarios where traditional NMPC results in an unstable response for this process. Additionally, unconstrained state estimators such as EKF are unsuitable because the parameter uncertainty may cause estimates to fall outside the feasible region of the process, which ultimately destabilizes the process. MHE was able to overcome this challenge because it considers process constraints in its formulation. The additional computational time required to solve the robust NMPC formulations and MHE does not cause significant delays for the sampling time considered, demonstrating their applicability to challenging large-scale industrial chemical processes.

has been performed on small-case studies involving less than 10 states (Holtorf, Mitsos, & Biegler, 2019; Kummer, Nagy, & Varga, 2020; Lucia et al., 2014, 2017; Piceno-Dıaz et al., 2020; Puschke & Mitsos, 2018; Skupin et al., 2022; Subramanian, Lucia, & Engel, 2015; Thangavel, Paulen, & Engell, 2020; Thangavel et al., 2018; Thombre et al., 2021; Tătulea-Codrean et al., 2020). To name a few, Holtorf et al. (2019) implemented an MS-NMPC with online-generated scenario trees for a semi-batch polymerization process with seven states. Kummer et al. (2020) applied an MS-NMPC to a semi-batch Williams-Otto process

a flowsheet of the TE process. There are eight chemical components involved, A to H. Components A to C are non-condensable gases and components D to H may exist in both liquid and gaseous phases. Component B is an inert component. The reactions occur in the gas phase of the reactor (Downs & Vogel, 1993). Products G and H are desirable and component F is a by-product. The reactions, (1)-(4), are exothermic and irreversible, and the rates are temperature sensitive. The process model is highly nonlinear, and it is open-loop unstable, which means control is required to stabilize the system and maintain it

Tatjewski, P., & Ławryńczuk, M. (2020). Algorithms with state estimation in linear and nonlinear model predictive control. *Computers & Chemical Engineering*, *143*, Article 107065. http://dx.doi.org/10.1016/j.compchemeng.2020.107065.

Thangavel, S., Paulen, R., & Engell, S. (2020). Robust multi-stage nonlinear model predictive control using sigma points. *Processes*, *8*(7), 851. http://dx.doi.org/10.3390/pr8070851.

Thangavel, S., et al. (2018). Dual robust nonlinear model predictive control: A multi-stage approach. *Journal of Process Control*, *72*, 39–51. http://dx.doi.org/10.1016/j.jprocont.2018.10.003.

Thombre, M., et al. (2021). Sensitivity-assisted multistage nonlinear model predictive

Valipour, M., Toffolo, K. M., & Ricardez-Sandoval, L. A. (2021). State estimation and sensor location for entrained-flow gasification systems using Kalman filter. *Control Engineering Practice*, *108*, Article 104702. http://dx.doi.org/10.1016/j.conengprac.2020.104702.

Vinoth Upendra, J., & Prakash, J. (2013). Comparison of state estimation algorithms on the Tennessee Eastman process. In R. Malathi, & J. Krishnan (Eds.), *Lecture notes in electrical engineering, Recent advancements in system modelling applications* (pp. 357–368). India: Springer India, http://dx.doi.org/10.1007/978-81-322-1035-1_31.

Welch, G., & Bishop, G. (2006). *An introduction to the kalman filter.*

Yan, M., & Ricker, N. L. (1995). Multi-objective control of the Tennessee Eastman