2013 European Control Conference (ECC)
July 17-19, 2013, Zürich, Switzerland.

Outlier Analysis in Set-based Estimation for Nonlinear Systems Using Convex Relaxations

Stefan Streif, Matthias Karl, and Rolf Findeisen

Abstract— **Set-based estimation for nonlinear systems is a useful tool to handle sparse and uncertain data. The tool provides outer bounds on feasible parameter sets and reachable states, as well as provable inconsistency certificates for entire parameter regions. In case of errors in the data such as outliers or incorrect *a priori* assumptions on variable uncertainties, set-based approaches can, however, lead to poor estimates or even rejection of a consistent model. We present a set-based approach to systematically identify outliers or incorrect variable uncertainty assumptions. The basic idea is to detect outliers by quantifying the influence they have on the inconsistency of an underlying feasibility problem. The results build on a set-based estimation framework that employs convex relaxations. Specifically we derive model consistency measures and sensitivity measures that combine the sensitivity information stored in the Lagrange dual variables. An algorithm is developed that iteratively detects outliers that contribute most to inconsistency. The algorithm terminates once the data and model are no longer proved inconsistent. The approach is illustrated by an example.**

I. Introduction

In many fields of science and engineering, model-based analysis, prediction, and design play an important role. Models are often nonlinear and depend on unknown or uncertain parameters that first have to be estimated from data. It is however difficult to provide conclusive parameter estimates, because available experimental data are often sparse, corrupted by noise and measurement outliers. To deal with sparse and uncertain data and to allow guaranteed tests of model consistency, and to provide outer bounds on consistent parameters or predictions of model behavior, set-based analysis methods are useful tools (e. g. [1]–[8]).

Real measurement data frequently contain outliers, i. e. measurements which deviate notably from the expected system behavior or the remaining data [9]. Reasons for outliers can be, e. g., errors in data transmission, flushing, cleaning, or calibration of sensors, or sensor failures [9], [10]. Thus, outliers are often caused by other mechanisms and not by the system itself. A review of data outlier analysis methods and arising problems can be found e. g. in [10].

Another frequently occurring problem is that *a priori* bounds on uncertain parameters are often poor or not available at all. Often parameter values are then assumed to range over several orders of magnitude to ensure that all possible consistent parameter values are covered. This can result in arbitrary or ambiguous model-based predictions, numerical difficulties, as well as long computing and estimation times [11]. On the contrary, if parameters uncertainties are incorrectly assumed too small, no consistent solution might exist and an actually consistent model is rejected. It is thus crucial to detect and deal with incorrect assumptions on parameters.

In this contribution we present a set-based approach to detect outliers in the data or incorrect assumptions on parameter bounds. The proposed approach to outlier detection builds on a set-based estimation framework that employs a feasibility formulation and checks infeasibility of a convex relaxation thereof [3], [4], all of which are available in the free Matlab toolbox ADMIT [11].

Specifically, we define outliers as uncertainty bounds that render the feasibility problem inconsistent. To detect outliers and inconsistent parameter bounds we derive suitable model consistency and sensitivity measures and exploit the sensitivity information available from the dual variables of the optimization/feasibility problem. This allows the quantification of the influence of uncertainty ranges on the consistency measure. As shown, however, additionally introduced constraints to tighten the solution sets in the relaxation- and set-based estimation require special attention. While these constraints can produce tighter solution sets [12], [13], a much larger number of constraints has to be considered in the outlier analysis, which is often not straightforward.

Related approaches for sensitivity analysis, however not for outlier detection, using dual variables have been presented in the literature. Castillo and coworkers [14], [15] used a similar sensitivity measure and applied it to nonlinear programming problems and various statistical error distributions. Frenklach and coworkers [16]–[19] performed sensitivity analysis and data set consistency analysis using an S-procedure relaxation for a different problem class and apply it to model response prediction. To the best of our knowledge, outlier detection within a set-based estimation or model invalidation framework based on convex linear relaxation has not been presented so far.

This contribution is structured as follows. In Section II and III we state the problem setup and briefly recap the relaxation- and set-based analysis framework used here, which is based on [3], [4]. In Sections IV–V we introduce the consistency and sensitivity measures used for outlier detection. We illustrate and discuss the outlier analysis approach at a small example in Section VI.

Note that this paper mainly addresses outlier analysis and correction for set-based estimation of consistent parameter sets. However, the presented results are directly applicable

All authors are with the Laboratory for Systems Theory and Automatic Control, Institute for Automation Engineering, Otto-von-Guericke-University Magdeburg, Germany. Corresponding author: Stefan Streif.
Emails: stefan.streif@ovgu.de, matthias.karl@st.ovgu.de, rolf.findeisen@ovgu.de.

for outlier analysis e. g. in set-based state estimation.

II. Set-based Estimation and Outliers

Basically we define outliers as measurements or uncertainty bounds which lead to inconsistency of a set-based estimation problem. We therefore first introduce in this section the considered set-based estimation problem and the underlying feasibility problem. Then we introduce the set-based outlier analysis problem.

We consider dynamical, nonlinear, discrete-time systems in implicit form

$$\text{MP}: \begin{cases} F\big(x(k{+}1), x(k), p\big) = 0 \\ Y\big(y(k), x(k), p\big) = 0 \end{cases}$$

Here $k \in \mathcal{T} = \{1, \dots, n_t\}$ denotes the time index. The states are denoted by $x(k) \in \mathbb{R}^{n_x}$, the outputs by $y(k) \in \mathbb{R}^{n_y}$, and the parameters by $p \in \mathbb{R}^{n_p}$. We assume $F : \mathbb{R}^{n_x} \times \mathbb{R}^{n_x} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_x}$ and $Y : \mathbb{R}^{n_x} \times \mathbb{R}^{n_y} \times \mathbb{R}^{n_p} \to \mathbb{R}^{n_y}$ to be polynomial or rational functions.

Within the set-based framework, we do not assume that the measurements are given as a vector of output values. Rather we assume that a time series of uncertainty sets for the outputs and *a priori* bounds on the states are available. In particular, we assume that a lower and upper bound on each uncertain variable is given. The corresponding inequalities are stored in a collection of the following form:

Assumption 1 (Description of uncertainty bounds) *The uncertainties of the parameters p, output measurements $y(k)$ and states $x(k)$ are defined by collections of inequalities*

$$\mathbb{C}_y = \bigcup_{\substack{k\in\mathcal{T} \\ i\in\{1,\dots,n_y\}}} \Big\{\underline{y_i(k)} \le y_i(k)\Big\} \cup \Big\{y_i(k) \le \overline{y_i(k)}\Big\}$$

$$\mathbb{C}_x = \bigcup_{\substack{k\in\mathcal{T} \\ i\in\{1,\dots,n_x\}}} \Big\{\underline{x_i(k)} \le x_i(k)\Big\} \cup \Big\{x_i(k) \le \overline{x_i(k)}\Big\}$$

$$\mathbb{C}_p = \bigcup_{i\in\{1,\dots,n_p\}} \Big\{\underline{p_i} \le p_i\Big\} \cup \Big\{p_i \le \overline{p_i}\Big\}$$

where $\underline{v} \ge 0$ and $\overline{v} < \infty$ denoting a lower and, respectively, upper bound on the variable v.

Note that each element in the collections $\mathbb{C}_y$, $\mathbb{C}_x$, and $\mathbb{C}_p$ represents an inequality. We use this notation later in the definition of outliers. By removing or modifying the inequalities corresponding to outlier from the collections, the resulting set-based estimation problem becomes consistent. Note also that from now on we only analyze the collections $\mathbb{C}_p$ and $\mathbb{C}_y$ for outliers, but not $\mathbb{C}_x$. The latter collection has been added for the sake of completeness. In the following, we denote by $\mathcal{P}$ the set of admissible parameter values, which is given by the intersection of the half-spaces represented by the inequalities in Assumption 1.

A. Set-based Parameter Estimation

The aim of set-based parameter estimation is to find the consistent parameter set $\mathcal{P}^* \subseteq \mathcal{P}$ for which the model MP is consistent $\forall\, p \in \mathcal{P}^*$ with the uncertain output measurements $\mathbb{C}_y$ and uncertain initial state estimates $\mathbb{C}_x$. Note that the same framework can be used for set-based state estimation or reachability analysis [11].

To check consistency of parameters, we combine MP and the collection of inequalities from Assumption 1 in a feasibility problem (FP) [3], [4]:

$$\text{FP}: \begin{cases} \text{find} & \xi_{\text{FP}} & \\ \text{subject to} & F(x(k{+}1), x(k), p) = 0 & \forall\, k \in \mathcal{T}^+ \\ & Y(y(k), x(k), p) = 0 & k \in \mathcal{T} \\ & \mathbb{C}_y, \quad \mathbb{C}_x, \quad \mathbb{C}_p & \end{cases}$$

The vector $\xi_{\text{FP}} \in \mathbb{R}^{n_{\xi_{\text{FP}}}}$ lumps all variables, i. e. $\xi_{\text{FP}} = \big[x(1)^T, \dots, x(n_t)^T, y(1), \dots, y(n_t)^T, p^T\big]^T$, and $\mathcal{T}^+ = \mathcal{T} \setminus n_t$. It is easy to see that FP allows to detect inconsistent parameter sets:

Theorem 1 (Inconsistent parameter set [4]) *If the feasibility problem FP does not admit a solution ξ^*_{FP}, then the model MP is inconsistent $\forall\, p \in \mathcal{P}$.*

To directly check if FP admits a solution or not is in general difficult due to the nonlinearities and nonconvexities. We therefore use convex relaxations (see Section III) to prove inconsistency of $\mathcal{P}$ (i. e. invalidate entire models) or of partitions $\mathcal{P}_i \subset \mathcal{P}$. Then an outer approximations $\hat{\mathcal{P}}$ of the consistent parameter set $\mathcal{P}^*$ is found by the union of partitions $\mathcal{P}_i$ which were not proved inconsistent (cf. [4]).

B. Set-based Outlier Analysis

In [10], different definitions for outliers in different contexts are presented. Similar to [9], but adapted to the set-based setting and feasibility formulation, we consider outliers to be measurements that are inconsistent with the model behavior under all possible uncertainties of the parameters and remaining data. Thus, the presence of the outlying measurement data precludes a consistent solution of FP; but if the outliers are removed, then the model dynamics and output equations in FP are consistent with the remaining data:

Definition 1 (Set-based measurement data outliers) *Given that FP admits no consistent solution for $\mathbb{C}_y$. Then we define measurement outliers $\mathbb{O}_y \subseteq \mathbb{C}_y$ as the collection of inequalities such that FP admits a consistent solution ξ^*_{FP} if the corresponding inequalities are removed from FP, i. e. $\mathbb{C}_y \coloneqq \mathbb{C}_y \setminus \{\mathbb{O}_y\}$.*

$\mathbb{O}_y$ collects the inequalities of $\mathbb{C}_y$ (see Assumption 1) for which the corresponding $y_i(k)^*$ from ξ^*_{FP} violate the associated inequalities. Here we made the implicit assumption that a consistent solution ξ^*_{FP} exists if all outliers are removed.

Remark 1 (Non-uniqueness of outliers) *Note that the outliers depend on the uncertainty bounds (Assumption 1) and the resulting set of consistent solutions to FP. This definition does in general not lead to a unique definition of outliers. In the extreme case, a consistent solution of FP can be obtained if all uncertainty bounds are removed. Also the presence of outlying measurements can determine which other measurements are considered outliers, so called swamping and masking effects [10].*

We also address the problem of correcting incorrect initial parameter assumptions:

Definition 2 (Incorrect set-based parameter assumptions) *Given that FP admits no consistent solution for $\mathbb{C}_p$. Then we define incorrect parameter assumptions $\mathbb{O}_p \subseteq \mathbb{C}_p$ as the collection of inequalities such that FP admits a consistent solution ξ^*_{FP} if the corresponding inequalities are removed from FP, i. e. $\mathbb{C}_p \coloneqq \mathbb{C}_p \setminus \{\mathbb{O}_p\}$.*

Note that $\mathbb{O}_p$ might not be unique (cf. Remark 1) and that we implicitly assumed the existence of a consistent solution if all incorrect parameter assumption were removed.

Definition 1 is a model-based or model-generic [10] definition for outliers, whereas Definition 2 is a data-dependent definition for incorrect model assumptions. Before presenting the set-based approach to find and correct outliers, we recap the mathematical details of the set-based parameter estimation framework that is used to check consistency.

III. Preliminaries: Convex and Linear Relaxations for Set-based Parameter Estimation

Outliers are defined in terms of inconsistency of the FP. Before proceeding with the description of our approach to set-based outlier detection, we outline the key steps and approaches for solving the FP. All necessary steps and algorithms are implemented in the toolbox ADMIT [11].

A. Convex Relaxations

The considered FP is usually nonlinear and non-convex which makes it difficult to directly check consistency and derive consistent solution sets. We therefore apply convex relaxation techniques to determine outer approximations and provable inconsistency certificates. Several relaxation steps are made [3], [4] and detailed below.

The FP can be transformed into a quadratic problem (QP). But the obtained QP is still non-convex and is therefore relaxed into a semi-definite programming problem (SDP). However, solving large SDPs is very challenging. Transforming it into a linear programming problem (LP) allows large problems to be solved much faster using a range of efficient solvers.

In the first step, the FP is rewritten as a QP by expressing all dynamic and output equations in quadratic form $\xi^T A \xi = 0$, where $A \in \mathbb{R}^{n_\xi \times n_\xi}$ is a symmetric matrix. $\xi \in \mathbb{R}^{n_\xi}$ is a minimal basis of monomials for the equations and contains the elements of ξ_{FP}, the constant 1, and additional monomials of degree ≥ 2 that are required to represent all equations.

The inequalities in the collections $\mathbb{C}_p$, $\mathbb{C}_x$, and $\mathbb{C}_y$, which define the bounds on the uncertain variables (cf. Assumption 1) can be formulated as n_{in} linear inequality constraints in matrix form $B\xi \geq 0$, $B \in \mathbb{R}^{n_{\text{in}} \times n_\xi}$. Define the index set $\mathcal{J} = \{1, \dots, n_{\text{eq}}\}$ to address all matrices A_j corresponding to the n_{eq} equality constraints. Then the FP can be rewritten as

$$\text{QP}: \begin{cases} \text{find} & \xi \in \mathbb{R}^{n_\xi} \\ \text{subject to} & \xi^T A_j \xi = 0 \quad j \in \mathcal{J} \\ & \xi_1 = 1 \\ & B\xi \geq 0 \end{cases}$$

Such a quadratic decomposition can always be found, however it is still not convex.

In the second step, we obtain a convex semidefinite program by introducing the symmetric variable matrix $X = \xi\xi^T \in \mathbb{R}^{n_\xi \times n_\xi}$ and relaxing the resulting conditions $\text{rank}\,(X) = 1$ and $\text{trace}\,(X) \geq 1$ with the weaker constraint $X \succeq 0$, see e.g. [20]. The semidefinite program reads:

$$\text{SDP}: \begin{cases} \text{find} & X \\ \text{subject to} & \text{trace}\,(A_j X) = 0 \quad j \in \mathcal{J} \\ & \text{trace}\left(ee^T X\right) = 1 \\ & BXe \geq 0 \\ & X \succeq 0 \end{cases}$$

where $e = (1, 0, ..., 0)^T \in \mathbb{R}^{n_\xi}$.

B. Redundant Constraints

Due to the semidefinite relaxation, the solution set will increase compared to FP, which might lead to the inclusion of false solutions. To tighten the relaxation, a common technique [12], [13] is to add constraints that are redundant in the basis ξ_{FP}, but not necessarily redundant in the higher-dimensional variable basis X of the semidefinite relaxation. The following constraints are suggested:

$$BXB^T \succeq 0 \tag{1a}$$

$$A_j X B^T = 0 \quad \forall\, j \in \mathcal{J}_{\text{leq}} \tag{1b}$$

$$A_{j_1} X A_{j_2}^T = 0 \quad \forall\, j_1, j_2 \in \mathcal{J}_{\text{leq}} \tag{1c}$$

$\mathcal{J}_{\text{leq}} = \{1, \dots, n_{\text{leq}}\} \subseteq \mathcal{J}$ is the index set for all equality constraint which are linear in ξ. Note that Equation (1a) also contains the McCormick relaxation for bilinear monomials.

The constraints (1) are added to the SDP, which increases the problem size by $\frac{n_\xi(n_\xi+1)}{2} + n_{\text{leq}} n_{\text{in}} + \frac{n_{\text{leq}}(n_{\text{leq}}+1)}{2}$ constraints, but tightens the solution set due to additional cuts. The downside to it is that the large number of additional constraints make an outlier detection with standard methods not straightforward [14], [21]. An approach to overcome this problem is presented in Section V.

C. Linear Relaxation

To deal with larger problems with more constraints and more variables, we relax the SDP to a linear program LP. In the third step, therefore, a linear relaxation is derived from the SDP (and constraints (1)) by substituting the semi-positive definiteness constraint $X \succeq 0$ with element-wise non-strict positivity of X ($X \geq 0$):

$$\text{LP}: \begin{cases} \text{find} & X \\ \text{subject to} & \text{trace}\,(A_j X) = 0 \quad \forall\, j \in \mathcal{J} \\ & \text{trace}\left(ee^T X\right) = 1 \\ & BXe \geq 0 \\ & X \geq 0 \\ & BXB^T \geq 0 \\ & A_j X B^T = 0 \quad \forall\, j \in \mathcal{J}_{\text{leq}} \\ & A_{j_1} X A_{j_2}^T = 0 \quad \forall\, j_1, j_2 \in \mathcal{J}_{\text{leq}} \end{cases}$$

Note that $X \geq 0$ is already implied by $BXe \geq 0$ and due to Assumption 1.

A more intuitive representation that simplifies the subsequent notation and analysis is obtained by rewriting LP in an equivalent aggregated and vectorized form. The column

vector $\tilde{\zeta} = \text{vec}(X) \in \mathbb{R}^{n_{\tilde{\zeta}}}$, $n_{\tilde{\zeta}} = \frac{n_\xi(n_\xi+1)}{2}$, then contains all unique elements of X. Straightforward matrix reformulations yields a linear problem in familiar form:

$$\text{vLP}: \begin{cases} \text{find} & \tilde{\zeta} \in \mathbb{R}^{n_{\tilde{\zeta}}} \\ \text{subject to} & \tilde{A}_{\text{eq}}\tilde{\zeta} = b_{\text{eq}} \\ & \tilde{B}_{\text{in}}\tilde{\zeta} \geq 0 \end{cases}$$

with $b_{\text{eq}} = (1, 0, 0, \ldots, 0)^T \in \mathbb{R}^{n_{\tilde{\zeta}}}$, and $\tilde{A}_{\text{eq}}$ and $\tilde{B}_{\text{in}}$ of appropriate dimensions.

D. Inconsistency Certificates Based on the Lagrange Dual

As stated above, we are interested in proving inconsistency of the FP. An efficient approach [3] in this case is to consider the dual formulation dLP of the linear relaxation vLP:

$$\text{dLP}: \begin{cases} \max\limits_{\lambda,\nu} & \inf\limits_{\tilde{\zeta}} \left(\nu^T \left(\tilde{A}_{\text{eq}}\tilde{\zeta} - b_{\text{eq}}\right) - \lambda^T \tilde{B}_{\text{in}}\tilde{\zeta}\right) \\ \text{subject to} & \lambda \geq 0 \end{cases}$$

where the argument of the infimum operator is called the Lagrangian, and ν and λ are the dual variables corresponding to the equality and inequality constraints, respectively [21].

Theorem 2 (Inconsistency certificate [4]) *If the objective of the dual linear program dLP is unbounded, then MP is inconsistent $\forall\, p \in \mathcal{P}$.*

The weak-duality theorem and the relaxation process guarantee that if the objective of the dual program is unbounded, then the FP does not admit a solution [4] hence is inconsistent. Note that to prove consistency, other approaches such as Monte Carlo sampling or nonlinear programming have to be used [11].

For subsequent purposes it is necessary to consider an optimization problem. The vLP can be converted into an optimization problem by adding an objective function $f = c^T\tilde{\zeta}$, $c \in \mathbb{R}^{n_{\tilde{\zeta}}}$ to the Lagrangian in dLP [21]. Minimization then yields a lower bound on the global optimizer of FP. By iteratively minimizing and maximizing over different variables, outer approximations can be determined [4]. In practice, the linear relaxation presented here yields satisfactory results and allows large problems to be considered [11].

The Lagrange dual variables λ and ν contain useful sensitivity information [21]. We use this information for outlier detection in Sections IV and V. First, however, we have to define a suitable optimization problem that quantifies consistency of a model and data. Sensitivity of the consistency measure with respect to parameter and measurement uncertainty bounds will then be determined to identify outliers and incorrect parameter assumptions, and to account for the large number of redundant constraints.

IV. Model Consistency Measure

In this section we derive an optimization-based model consistency measure, which allows the identification of outliers. The basic purpose of the consistency measure is that its value quantifies how inconsistent a model MP and the data $\mathbb{C}_y$ and $\mathbb{C}_p$ are. This can be imagined as follows. Assume inconsistency has been proved with Theorem 2. Then, e. g., the measurement uncertainties $\mathbb{C}_y$ (Assumption 1) will be increased (i. e. the inequality constraints will be weakened) until inconsistency cannot be proved any longer, because the uncertainties may now cover consistent solutions. In Section V, we use sensitivity analysis to determine which uncertainties contributed most to consistency.

First of all, we have to reformulate the vLP such that the widths of the uncertainty intervals (according to Assumption 1 for a set of variables of interest, i. e. either parameters p or measurements y) can be increased as described. To this end, we introduce the consistency variable $\gamma \in \mathbb{R}$ and reformulate the lower and upper bounds of interest appearing in the matrices $\tilde{A}_{\text{eq}}$ and $\tilde{B}_{\text{in}}$. More specifically, we make the following substitutions: $\underline{\xi_{\text{FP}}} := \xi_{\text{FP}}^C - \gamma\xi_{\text{FP}}^W$ and $\overline{\xi_{\text{FP}}} := \xi_{\text{FP}}^C + \gamma\xi_{\text{FP}}^W$. The central value ξ_{FP}^C of each uncertainty interval is given by $\frac{\underline{\xi_{\text{FP}}}+\overline{\xi_{\text{FP}}}}{2}$, and its half-width ξ_{FP}^W by $\frac{\overline{\xi_{\text{FP}}}-\underline{\xi_{\text{FP}}}}{2}$. To exemplify this, one obtains for $v - \underline{v} \geq 0$ the constraint $v - (v^C - \gamma v^W) \geq 0$. Note that we make symbolic substitutions in the equations using the nonlinear bases ξ_{FP} and higher order monomials ξ. This is easily possible if one keeps track of the nonlinear definitions of all relaxation variables in ξ and $\tilde{\zeta} = \text{vec}(X) = \text{vec}\left(\xi\xi^T\right)$. We also keep the lower (resp. upper) bounds as symbolic variables in the equations. Prior to solving the problem, the symbolic variables are substituted by their corresponding numerical values from Assumption 1. The symbolic form is also needed later for sensitivity analysis (Proposition 2) to be able to derive partial derivatives. Finally, the obtained equations are rewritten in matrix form as follows:

Proposition 1 (Model consistency) *The model consistency measure γ^* is defined via the optimal solution to*

$$\gamma LP: \begin{cases} \min\limits_{\zeta} & \gamma & \\ \text{subject to} & A_{eq}\zeta & = b_{eq} \\ & B_{in}\zeta & \geq 0 \\ & \gamma & \geq 0 \end{cases}$$

with linear basis $\zeta^T = \left[\tilde{\zeta}^T, \tilde{\zeta}_\gamma^T, \gamma, \gamma_2\right]$, where $\tilde{\zeta}_\gamma := \gamma\tilde{\zeta}$ and $\gamma_2 := \gamma^2$.

Note that due to the constraint (1a) new products involving γ appeared, which required to extend the linear basis. Also higher order products of γ can appear due to higher-order monomials in ξ, which then have to be accounted for in ζ.

The model consistency measure γ^* can be interpreted as follows. For $\gamma^* \leq 1$, the original model is not found inconsistent and set-based parameter estimation can be addressed immediately. For $\gamma^* > 1$, the uncertainties had to be increased to obtain a non-inconsistent model and an outlier analysis should be performed.

The presented consistency measure (Proposition 1) lays the basis for outlier detection in Section V. Other optimization and relaxation-based consistency measures have been proposed in the literature to allow the comparison of model hypotheses, see e. g. [5] (and references therein) using a Sum-of-Square relaxation. However, our measure is particularly well-suited for the considered relaxations with a large number of redundant constraints as shown next.

V. Sensitivity Measure and Outlier Analysis

To detect outliers and incorrect parameter assumptions, we aim to identify which bounds on the variables in ξ_{FP} (cf. Assumption 1) contribute most to an increase of γ^*. What we need are sensitivity measures that link γ^* and bounds of ξ_{FP}. As it is well-known [21], Lagrange dual variables contain information about sensitivity of the objective function value with respect to perturbations of the constraints. To exemplify this, consider the constraint $c_i := v - \underline{v} \geq 0$ with associated dual variable λ_i. Then $\lambda_i = \frac{\partial \gamma^*}{\partial c_i}$ quantifies how much γ^* changes if the constraint is perturbed. This can be used to quantify the influence of the bound $\underline{v}$ [14], [15], [17]: $\frac{\partial \gamma^*}{\partial \underline{v}} = \frac{\partial \gamma^*}{\partial c_i}\frac{\partial c_i}{\partial \underline{v}} = \lambda_i \frac{\partial c_i}{\partial \underline{v}}$. However, due to the large number of tightening constraints (1) and associated dual variables, it is not straightforward to systematically quantify the sensitivity of the different uncertainty bounds. The following sensitivity measure provides means to overcome this problem.

Proposition 2 (Sensitivity measure) *The sensitivities $S_w \in \mathbb{R}$, $w \in \{\underline{v}, \overline{v}\}$, of the consistency measure γ^* with respect to the lower bound $\underline{v}$ or upper bound $\overline{v}$ of the uncertain variable $v \in \xi_{FP}$ are given by*

$$S_w = \left| \nu^T \frac{\partial}{\partial w}\left(A_{eq}\zeta^*_{FP} - b_{eq}\right) - \lambda^T \frac{\partial}{\partial w}\left(B_{in}\zeta^*_{FP}\right) \right|$$

where λ and ν are the Lagrange dual variables associated with γLP. $\zeta^*_{FP} = vec\left(\xi_{FP}\,\xi^T_{FP}\right)$ is a column vector of appropriate dimension containing the nonlinear monomial vector ξ_{FP} and is evaluated at the solution ζ^* of γLP.

It is not difficult to see that the sensitivities are obtained from the partial derivative of a modified Lagrangian (dLP for γLP) in which symbolic substitutions for the lower and upper bounds were made as described in the previous section.

Remark 2 (Normalization) *The sensitivity measure from Proposition 2 can be normalized in the usual fashion [22], [23] to account for variables of different scales.*

Using Proposition 2 allows a systematic ranking of the variable bounds that contribute most to model inconsistency. The following algorithm can be used to iteratively determine, remove or weaken inconsistent constraints until the relaxed problem becomes feasible. This then allows the identification of outliers or incorrect parameter bounds (cf. Section II-B).

Algorithm 1 (Measurement outlier detection and correction)

```
Input:  inconsistent model MP and data 𝕄_y,
        magnitude of modification κ > 0
Output: non-inconsistent model MP and
        modified inequality collection 𝕄_y
```

(1) determine consistency measure γ^* by solving γLP
(2) stop if $\gamma^* \leq 1$
(3) perform sensitivity analysis for measurement uncertainties $\mathbb{C}_y$
(4) select lower or upper bound j with largest sensitivity and replace:
  - if lower: $\underline{\zeta_j} := \max\left(0, \zeta_j^C - (\gamma^* + \kappa)\,\zeta_j^W\right)$
  - if upper: $\overline{\zeta_j} := \zeta_j^C + (\gamma^* + \kappa)\,\zeta_j^W$

(5) go to step (1)

Remark 3 (Parameter outlier detection and correction) *By replacing $\mathbb{C}_y$ with $\mathbb{C}_p$ in Algorithm 1, one can also analyze incorrect parameter assumptions (cf. Definition 2).*

The algorithm follows a greedy strategy and modifies only one potentially inconsistent bound per iteration, thereby trying to keep the number of modifications of $\mathbb{C}_y$ small. The magnitude of modification is proportional to $(\gamma^*+\kappa)$ and the width of the corresponding uncertainty interval ζ_j^W.

Depending on the value of κ, the bounds are either weakened or completely removed from the FP. For small values, potentially inconsistent bounds will be weakened only slightly. This might then require many iterations of the algorithm until inconsistency cannot be proved any longer. For $\kappa > 0$, it is more likely that the set of consistent solutions is larger once the algorithm terminates. Whereas for $\kappa = 0$, a single consistent solutions might only be found on the boundaries of the uncertainty sets. Choosing κ large, potential outliers are weakened to a large extent. The algorithm might then terminate faster, but set-based predictions with the resulting model might be poor (cf. Introduction). For $\kappa \to \infty$, the corresponding constraints are basically removed from the set of equations, thus: $\mathbb{C}_y \setminus \{\mathbb{O}_y\}$.

Remark 4 (Modification of several bounds) *It may happen that several bounds have the same sensitivity, which then requires to modify step (4): either only one of the sensitive variable bounds is picked at random, or all bounds are chosen and modified.*

Remark 5 (Algorithm run-time) *The run-time of the algorithm is proportional to the average time needed for solving one instance of γLP and proportional to the number of iterations. Because efficient solvers are available, γLP can be solved usually very quickly. See [3], [4], [11] for discussions on further complexity and performance issues related to set-based approaches.*

VI. Example

To illustrate the set-based approach for outlier detection, we analyze a small reaction motif that is often found in systems biology models. More complex systems have been analyzed using set- and relaxation-based methods (see e. g. [3], [4], [11]). The model and the set-based outlier analysis have been implemented in ADMIT [11].

In the considered reaction network, enzyme E converts substrate S into product P via the intermediary complex C. The time-discretized dynamics (using an implicit Euler scheme) reads:

$$S(k) = S(k\text{–}1) + \Delta T\left(-p_1\,S(k)E(k) + p_2\,C(k)\right)$$
$$C(k) = C(k\text{–}1) + \Delta T\left(p_1\,S(k)E(k) - (p_2 + p_3)\,C(k)\right)$$

with $S(k) + P(k) = 1$, $E(k) + C(k) = 1, \forall\, k \in \mathcal{T}$, $\mathcal{T} = \{0,1,\ldots,10\}$, and $\Delta T = 0.1$. The parameters bounds were set to $0.1 \leq p_i \leq 10, i \in \{1,2,3\}$. A nominal and consistent solution was obtained from simulation with $p_1 = p_2 = p_3 = S(0) = 1$, $C(0) = 0$. Measurement values were generated by sampling from a normal distribution (mean values from simulation and standard deviations of 0.01 for S and 0.04 for C) and by adding $\pm 5\%$ relative error to the sampled values.

After the linear relaxation, the vLP contains 741 variables and 3 930 constraints in total. Due to space limitations, we only analyze measurement outliers using the presented method and algorithm. Outliers were generated by adding 1 to the lower and upper bounds on $S(2)$, and by decreasing lower and upper bounds on $S(3)$ by 0.5 and on $C(7)$ by 0.1.

Starting with $\gamma^* = 7.84$, the algorithm ($\kappa = 0.1$) terminated after 16 iterations with $\gamma^* = 0.99$. All outlying bounds were corrected and consistent solutions were obtained. In addition, the lower bounds on the variables $C(4)$ and $C(8)$ were spuriously considered outliers and were also modified. The reason is that wrong solutions can be determined due to the relaxation. For these solutions, different constraints (including non-outliers) can be active with non-zero associated sensitivities, which then are treated like outliers by the algorithm.

VII. Conclusions

While set-based inconsistency certificates for an entire model or parameter regions can be a useful analysis tool [1]–[8]. However, set-based approaches might fail or produce wrong results in case of data outliers or incorrect *a priori* parameter uncertainty bounds.

The presented outlier analysis approach can be used in a first step to systematically identify and correct those uncertainty bounds and constraints that caused inconsistency and is therefore a crucial preliminary step prior to further set-based analyses. Note that the outlier analysis approach can also be used for set-based state estimation or reachability analysis.

The presented approach could also be used for set-based fault detection [24]. It should be possible to find bounds on the uncertainties which do not lead to inconsistency of fault candidate models. These bounds could then be used to quantify sensitivity of fault detection approaches with respect to measurement or parameter uncertainties.

As discussed, it is not straightforward to detect outliers by simple inspection of the constraints due to the large amount of redundant constraints. These redundant constraints are actually important to tighten the linear relaxation-based estimation results. For smaller problems, one could also use the usually tighter SDP relaxation instead of the LP relaxation. But the redundant constraints are also needed to tighten the SDP relaxation [12], [13], [20]. However, the presented results can be applied to the SDP relaxation without any modifications.

Acknowledgment

We thank Philipp Rumschinski and Anton Savchenko for helpful discussions and critical reading of the manuscript.

References

[1] M. Kieffer and E. Walter, "Guaranteed nonlinear state estimation for continuous–time dynamical models from discrete–time measurements," in *In Proc. 5th IFAC Symposium on Robust Control Design*, 2006.

[2] M. Kieffer, E. Walter, and I. Simeonov, "Guaranteed nonlinear parameter estimation for continuous-time dynamical models," in *Proc. 14th IFAC Symposium on System Identification*, 2006, pp. 843–848.

[3] P. Rumschinski, S. Borchers, S. Bosio, R. Weismantel, and R. Findeisen, "Set-base dynamical parameter estimation and model invalidation for biochemical reaction networks." *BMC Syst. Biol.*, vol. 4, p. 69, 2010.

[4] P. Rumschinski, S. Streif, and R. Findeisen, "Combining qualitative information and semi-quantitative data for guaranteed invalidation of biochemical network models," *Int. J. Robust Nonlin. Control*, vol. 22, no. 10, pp. 1157–1173, 2012.

[5] J. Anderson and A. Papachristodoulou, "On validation and invalidation of biological models." *BMC Bioinformatics*, vol. 10, p. 132, 2009.

[6] M. Milanese and A. Vicino, "Optimal estimation theory for dynamic systems with set membership uncertainty: an overview," *Automatica*, vol. 27, no. 6, pp. 997–1009, 1991.

[7] M. Milanese and C. Novara, "Set membership identification of nonlinear systems," *Automatica*, vol. 40, no. 6, pp. 957–975, 2004.

[8] V. Cerone, D. Piga, and D. Regruto, "Set-membership error-in-variables identification through convex relaxation techniques," *IEEE Transactions on Automatic Control*, vol. 57, no. 2, pp. 517 –522, 2012.

[9] R. K. Pearson, "Outliers in process modeling and identification," *IEEE T. Contr. Syst. T.*, vol. 10, no. 1, pp. 55–63, 2002.

[10] I. Ben-Gal, *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*. Kluwer Academic Publishers, 2005, ch. Outlier detection.

[11] S. Streif, A. Savchenko, P. Rumschinski, S. Borchers, and R. Findeisen, "ADMIT: a toolbox for guaranteed model invalidation, estimation and qualitative-quantitative modeling," *Bioinformatics*, vol. 28, no. 9, pp. 1290–1291, 2012. [Online]. Available: http://ifatwww.et.uni-magdeburg.de/syst/ADMIT/

[12] J. B. Lasserre, "Global optimization with polynomials and the problem of moments," *SIAM J. Optimiz.*, vol. 11, no. 3, pp. 796–817, 2001.

[13] M. Anstreicher, "Semidefinite programming versus the reformulation-linearization technique for nonconvex quadratically constrained quadratic programming," *J. Glob. Opt.*, vol. 43, no. 2-3, pp. 471–484, 2009.

[14] E. Castillo, A. J. Conejo, R. Mnguez, and C. Castillo, "A closed formula for local sensitivity analysis in mathematical programming," *Eng. Optimiz.*, vol. 38, no. 1, pp. 93–112, 2006.

[15] E. Castillo, A. Hadi, A. Conejo, and A. Fernández-Canteli, "A general method for local sensitivity analysis with application to regression models and other optimization problems," *Technometrics*, vol. 46, no. 4, pp. 430–444, 2004.

[16] T. M. Russi, A. Packard, R. P. Feeley, and M. Frenklach, "Sensitivity analysis of uncertainty in model prediction," *J. Phys. Chem. A*, vol. 112, no. 12, pp. 2579–2588, 2008.

[17] T. M. Russi, "Uncertainty quantification with experimental data and complex system models," Ph.D. dissertation, Mechanical Engineering, University of California, Berkeley, 2010.

[18] R. P. Feeley, "Fighting the curse of dimensionality: A method for model validation and uncertainty propagation for complex simulation models," Ph.D. dissertation, Mechanical Engineering, University of California, Berkeley, 2008.

[19] R. P. Feeley, M. Frenklach, M. Onsum, T. M. Russi, A. Arkin, and A. Packard, "Model discrimination using data collaboration," *J. Phys. Chem. A*, vol. 110, no. 21, pp. 6803–6813, 2006.

[20] P. A. Parrilo, "Semidefinite programming relaxations for semialgebraic problems," *Math. Program.*, vol. 96, no. 2, pp. 293–320, 2003.

[21] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

[22] A. Saltelli, *Sensitivity Analysis*. Wiley, 2000.

[23] S. Streif, S. Waldherr, F. Allgöwer, and R. Findeisen, *Systems Analysis of Biological Networks*, ser. Methods in Bioengineering. Artech House MIT Press, 2009, ch. Steady state sensitivity analysis of biochemical reaction networks: a brief review and new methods, pp. 129–148.

[24] A. Savchenko, P. Rumschinski, S. Streif, and R. Findeisen, "Complete diagnosability of abrupt faults using set-based sensitivities," in *Proc. 8th IFAC Symposium on Fault Detection, Supervision and Safety of Technical Processes (SafeProcess 2012)*, 2012, pp. 860–865.