

A State-space Realization Approach to Set Identification of Biochemical Kinetic Parameters

Yutaka Hori and Richard M. Murray

Abstract—This paper proposes a set-based parameter identification method for biochemical systems. The developed method identifies not a single parameter but a set of parameters that all explain time-series experimental data, enabling the systematic characterization of the uncertainty of identified parameters. Our key idea is to use a state-space realization that has the same input-output behavior as experimental data instead of the experimental data itself for the identification. This allows us to relax the originally nonlinear identification problem to an LMI feasibility problem validating the norm bound of an error system. We show that regions of parameters can be efficiently classified into consistent and inconsistent parameter sets by combining the LMI feasibility problems and a generalized bisection algorithm.

I. INTRODUCTION

The success of synthetic biology in the last decade has proved that mathematical models can be powerful prediction and diagnostic tools in the design of biological circuits. Although our ability to model biochemical systems had long been limited by the existence of many uncontrollable and unseen black boxes in a cell, it has recently become possible to test small biological circuit modules in a relatively well-controlled environment using in vitro cell-free expression systems [1]–[3]. These emerging technologies have motivated development of a grey-box, rather than a black-box, parameter identification tool that helps the modeling of small circuit parts for given reaction stoichiometries. To date, a number of identification methods were developed for biochemical systems (see [4], [5] for examples).

The major challenges of today’s biochemical parameter identification come from the non-convexity of the problem and the limitation of the measurable molecular species. In fact, direct application of classical deterministic identification methods such as least square fitting often suffers from local minima, and leaves the uncertainty to the identified parameters. Although it is often overlooked, the estimation of such uncertainty is crucial especially in the design of large biological circuits, since the combinations of uncertain models could limit the ability of prediction and lead to a wrong conclusion. It is thus desirable to develop an identification method that can systematically assess the uncertainty as a set of the parameters instead of finding a single point in the parameter space.

This work was supported in part by the Defense Advanced Research Projects Agency (DARPA) under Living Foundries program (HR0011-12-C-0065). Y. Hori is supported by JSPS Fellowship for Research Abroad.

Y. Hori and R. M. Murray are with Department of Computing and Mathematical Sciences, California Institute of Technology, USA. yhori@caltech.edu, murray@cds.caltech.edu

One of the existing approaches to the parameter set identification is Bayesian inference, which can provide posterior probability distributions of the estimated parameters as a measure of uncertainty. In Toni *et al.* [6], a Monte Carlo based approximate Bayesian computation (ABC) method was applied to infer the parameters of deterministic models of biochemical systems, and credible intervals of the identified parameters were successfully obtained from posterior distributions. Theoretical guarantees of the algorithm’s performance and the reduction of computational cost are, however, still open research topics.

In another line of research, two algebraic approaches were proposed to identify parameter sets with theoretical rigor. El-Samad *et al.* [7] and Anderson and Papachristodoulou [8] formulated parameter set validation problems using barrier certificates [9] and SOSTOOLS [10]. In contrast to the Bayesian approach, the algebraic approach can strictly identify a set of parameters by bounding the identification error using a Lyapunov-like approach, though the search for a barrier certificate involves substantial computational effort. A computationally less demanding approach was proposed by Kuepfer *et al.* [11] based on semidefinite relaxation of steady state equations. It was later shown that this approach is also useful for kinetic measurements by introducing an approximation to the model [12], [13], but the effect of the approximation has yet to be studied thoroughly [14].

In this paper, we propose a novel approach to parameter set identification that overcomes the issues in existing identification methods in that it has (i) theoretical guarantees of the identification results, (ii) the ability to handle time-series data, and (iii) competitive computational performance with the SDP relaxation approach [11]. The proposed method utilizes a time-series measurement of a perturbation experiment and identifies a parameter set that explains the experimental data, using a linearized ODE model. A key idea of our approach is to use a state-space realization that has the same input-output behavior as experimental data for the identification instead of the experimental data itself. This allows recasting the set identification problem as a norm minimization problem of the error system. The proposed algorithm solves linear matrix inequality (LMI) feasibility problems iteratively and efficiently identifies the parameter sets using a binary space partitioning method.

The rest of this paper is organized as follows. In Section II, we introduce the model of biochemical systems and mathematically formulate the set identification problem. The proposed method and its technical proof are then given in

Section III. Section IV is devoted to the demonstration of the proposed method on a simple enzymatic reaction model. Finally, Section V concludes the paper.

II. MODEL OF BIOCHEMICAL SYSTEMS AND PROBLEM STATEMENT

In this section, we first introduce a class of biochemical systems considered in this paper. Then, we formulate the parameter set identification problem mathematically.

A. Model of biochemical systems

We consider a set of biochemical reactions that involves n molecular species, $\mathcal{M}_1, \mathcal{M}_2, \dots, \mathcal{M}_n$, and r reaction channels, $\mathcal{R}_1, \mathcal{R}_2, \dots, \mathcal{R}_r$. For example, the molecules \mathcal{M}_i ($i = 1, 2, \dots, n$) stand for different mRNA and protein species, and \mathcal{R}_i ($i = 1, 2, \dots, r$) represent transcription, translation, degradation and so on. We denote the concentration of \mathcal{M}_i at time t by $\xi_i(t)$ and define a vector $\boldsymbol{\xi}(t) := [\xi_1(t), \xi_2(t), \dots, \xi_n(t)]^T$. The dynamics of the molecular concentrations can then be modeled by the law of mass action [15] as

$$\dot{\boldsymbol{\xi}} = M\boldsymbol{\varphi}(\boldsymbol{\xi}; \boldsymbol{\theta}), \quad (1)$$

where $M \in \mathbb{Z}^{n \times r}$ is the stoichiometry matrix of the reactions, and $\boldsymbol{\varphi}(\boldsymbol{\xi}; \boldsymbol{\theta}) := [\varphi_1(\boldsymbol{\xi}; \theta_1), \varphi_2(\boldsymbol{\xi}; \theta_2), \dots, \varphi_r(\boldsymbol{\xi}; \theta_r)]^T$ is the vector of propensity functions, where each entry $\varphi_i(\boldsymbol{\xi}; \theta_i)$ represents the propensity function of the reaction \mathcal{R}_i with a rate constant $\theta_i \in \mathbb{R}_+$. The vector of the rate constants is defined by $\boldsymbol{\theta} := [\theta_1, \theta_2, \dots, \theta_r] \in \mathbb{R}_+^r$. For example, the propensity function of a dimerization reaction of \mathcal{M}_{i_1} and \mathcal{M}_{i_2} can be written as $\varphi_i(\boldsymbol{\xi}; \theta_i) = \theta_i \xi_{i_1} \xi_{i_2}$. It should be noted that propensity functions are linear in terms of reaction rates due to the law of mass action.

When the reactions are near equilibrium, the dynamics of the concentrations around the equilibrium can be represented by a linearized model of equation (1). Let $\boldsymbol{\xi}^* \in \mathbb{R}_+^n$ denote an equilibrium concentration of the biochemical system (1) and $\boldsymbol{x} \in \mathbb{R}^n$ denote the difference of $\boldsymbol{\xi}$ from $\boldsymbol{\xi}^*$, i.e., $\boldsymbol{x} := \boldsymbol{\xi} - \boldsymbol{\xi}^*$. Using these notations, we introduce the following linear time-invariant system describing the local dynamics of the biochemical system (1).

$$\begin{aligned} \dot{\boldsymbol{x}} &= \sum_{i=1}^r \theta_i A_i \boldsymbol{x} = A(\boldsymbol{\theta}) \boldsymbol{x}, \boldsymbol{x}(0) = \boldsymbol{x}_0 \\ \boldsymbol{y} &= C \boldsymbol{x}, \end{aligned} \quad (2)$$

where the first equation is obtained by the Jacobian linearization of equation (1). The matrices A_i ($i = 1, 2, \dots, r$) and $A(\boldsymbol{\theta})$ are defined by

$$A_i := M \frac{\partial \varphi_i}{\partial \boldsymbol{\xi}} \Big|_{\boldsymbol{\xi}=\boldsymbol{\xi}^*} \quad \text{and} \quad A(\boldsymbol{\theta}) := \sum_{i=1}^r \theta_i A_i, \quad (3)$$

and $\boldsymbol{x}_0 \in \mathbb{R}^n$ is the initial value, which is an external input at $t = 0$ in a perturbation experiment. The second equation in the model (2) is a measurement equation with $C \in \mathbb{R}^{q \times n}$ and $\boldsymbol{y} \in \mathbb{R}^q$. In many cases, only a subset of molecular species is measurable, and the row entries of C are the standard bases

$\{\boldsymbol{e}_{i_k}\}_{k=1}^q$ corresponding to the measurable molecular species $\{\mathcal{M}_{i_k}\}_{k=1}^q$. In what follows, the notation $\boldsymbol{y}(t; \boldsymbol{\theta})$ is also used instead of \boldsymbol{y} to explicitly show the dependence of \boldsymbol{y} on the parameter $\boldsymbol{\theta}$.

Remark 1. In this paper, we assume that experiments are conducted in a relatively well-controlled environment such as in vitro systems, and the stoichiometry M , the propensity $\boldsymbol{\varphi}(\boldsymbol{\xi}, \boldsymbol{\theta})$ and the equilibrium $\boldsymbol{\xi}^*$ are known *a priori*. When the equilibrium is unknown, we can still linearize the system at an arbitrary non-equilibrium point and obtain the model $\dot{\boldsymbol{x}} = A(\boldsymbol{\theta})\boldsymbol{x} + M\boldsymbol{\varphi}(\boldsymbol{\xi}; \boldsymbol{\theta})$ with an additional constant bias term $M\boldsymbol{\varphi}(\boldsymbol{\xi}; \boldsymbol{\theta})$. Taking a derivative in time, we can obtain a similar form to the model (2) and proceed with a similar approach shown below. A thorough study of such extension is, however, left for our future work. \square

B. Problem statement

In this paper, we consider an identification problem of the parameters θ_i ($i = 1, 2, \dots, r$), given a time-series measurement of \boldsymbol{y} and the grey box model (2). In particular, we here propose a method to identify a set of parameters rather than a single point in the parameter space.

Suppose a time-series of molecular concentrations is obtained at $N + 1$ time points with sampling period T_s by a perturbation experiment around an equilibrium point. We denote the measured output of the biochemical system at $t = kT_s$ by $\hat{\boldsymbol{y}}[k] \in \mathbb{R}^q$ ($k = 0, 1, 2, \dots, N$) and define a set of the measurements $\hat{\mathcal{Y}}$ and the output of the model $\mathcal{Y}(\boldsymbol{\theta})$ by

$$\begin{aligned} \hat{\mathcal{Y}} &:= \{\hat{\boldsymbol{y}}[0], \hat{\boldsymbol{y}}[1], \hat{\boldsymbol{y}}[2], \dots, \hat{\boldsymbol{y}}[N]\}, \\ \mathcal{Y}(\boldsymbol{\theta}) &:= \{\boldsymbol{y}(t; \boldsymbol{\theta}) \mid \dot{\boldsymbol{x}} = A(\boldsymbol{\theta})\boldsymbol{x}, \boldsymbol{y} = C\boldsymbol{x}, \boldsymbol{x}(0) = \boldsymbol{x}_0\}. \end{aligned}$$

The goal of the parameter set identification is to find regions in the parameter space such that the difference between $\hat{\mathcal{Y}}$ and $\mathcal{Y}(\boldsymbol{\theta})$ is within a given tolerance and/or noise level γ . We define consistent and inconsistent parameter sets to refer to the regions that satisfy and dissatisfy the distance constraint.

Definition. Let $\mathcal{D}(\hat{\mathcal{Y}}, \mathcal{Y}(\boldsymbol{\theta}))$ denote a given metric that quantifies the difference between the measurement and the output of the model. We define a consistent parameter set \mathcal{C} and an inconsistent parameter set \mathcal{I} as follows.

$$\mathcal{C} := \left\{ \boldsymbol{\theta} \in \bar{\Theta} \mid \mathcal{D}(\hat{\mathcal{Y}}, \mathcal{Y}(\boldsymbol{\theta})) \leq \gamma \right\}, \quad (4)$$

$$\mathcal{I} := \left\{ \boldsymbol{\theta} \in \bar{\Theta} \mid \mathcal{D}(\hat{\mathcal{Y}}, \mathcal{Y}(\boldsymbol{\theta})) > \gamma \right\}, \quad (5)$$

where $\bar{\Theta}$ is a given parameter search space.

The concrete form of the metric $\mathcal{D}(\hat{\mathcal{Y}}, \mathcal{Y}(\boldsymbol{\theta}))$ is introduced in the next section.

The exact identification of the possibly non-convex continuous regions \mathcal{C} and \mathcal{I} is a difficult open problem. On the other hand, knowing the approximate shapes of \mathcal{C} and \mathcal{I} is helpful and more tractable from an engineering viewpoint. In

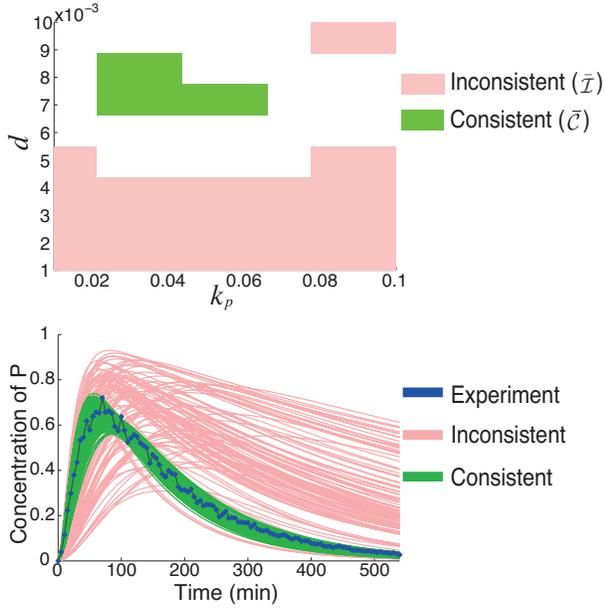


Fig. 1. (Top) Inner approximations of consistent parameters $\bar{\mathcal{C}}$ and inconsistent parameters $\bar{\mathcal{I}}$. (Bottom) Comparison of simulated experimental data and model simulation results with parameters in $\bar{\mathcal{C}}$ and $\bar{\mathcal{I}}$.

what follows, we consider the following inner approximation problem.

Problem. *Given a measurement of a perturbation experiment $\hat{\mathcal{Y}}$ and the grey box model (2), find inner approximations of the consistent and the inconsistent parameter sets $\bar{\mathcal{C}}(\subseteq \mathcal{C})$ and $\bar{\mathcal{I}}(\subseteq \mathcal{I})$.*

It should be noted that $\bar{\mathcal{I}}$ is equivalent to an outer approximation of the consistent parameter set \mathcal{C} , i.e., $\bar{\mathcal{C}} \subseteq \mathcal{C} \subseteq \bar{\mathcal{I}}$.

Figure 1 illustrates an example of the sets $\bar{\mathcal{C}}$ and $\bar{\mathcal{I}}$, and corresponding simulation results, whose details are presented in Section IV. The set $\bar{\mathcal{C}}$ can be interpreted as the uncertainty of the parameters that comes from the noise in the measurement, the limitation of measurable molecules and so on. In the design and analysis of biological circuits, the set identification result helps us plan another experiment that narrows down the uncertainty of specific parameters. Moreover, parameter sensitivity of the system can also be discussed from the shape of the regions $\bar{\mathcal{C}}$ and $\bar{\mathcal{I}}$.

III. PARAMETER SET IDENTIFICATION METHOD

A. Outline of the proposed algorithm

We propose an identification method that consists of two stages: (i) construction of a state-space realization and (ii) iterative search of consistent and inconsistent parameter sets.

In the first stage, we solve a state-space realization problem for given $\hat{\mathcal{Y}}$ and generate a linear time-invariant system whose output matches $\hat{\mathcal{Y}}$. Specifically, we define the ν -th order realization

$$\begin{aligned} \dot{\hat{\mathbf{x}}} &= \hat{A}\hat{\mathbf{x}}, \quad \hat{\mathbf{x}}(0) = \hat{\mathbf{x}}_0, \\ \mathbf{z} &= \hat{C}\hat{\mathbf{x}}, \end{aligned} \quad (6)$$

Algorithm 1: Realization algorithm

Input: Measurement $\hat{\mathcal{Y}}$

Output: Realization $(\hat{A}, \hat{\mathbf{x}}_0, \hat{C})$

- 1) Compute a singular value decomposition (SVD) of $H_{\ell_1, \ell_2}(0)$, where ℓ_1 and ℓ_2 are some constant satisfying $\ell_1 + \ell_2 = N$.

$$H_{\ell_1, \ell_2}(0) = [U, \bar{U}] \begin{bmatrix} \Sigma & O \\ O & O \end{bmatrix} \begin{bmatrix} V^T \\ \bar{V}^T \end{bmatrix},$$

where $\Sigma := \text{diag}(\sigma_1, \sigma_2, \dots, \sigma_\nu) \in \mathbb{R}^{\nu \times \nu}$ with $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_\nu$. The matrices U, \bar{U}, V and \bar{V} are defined accordingly.

- 2) Obtain a discrete-time realization $(\hat{F}, \hat{\mathbf{p}}_0, \hat{H})$ by

$$\begin{aligned} \hat{F} &= \Sigma^{-\frac{1}{2}} U^T (H_{\ell_1, \ell_2}(1)) V \Sigma^{-\frac{1}{2}} \\ \hat{\mathbf{p}}_0 &= \Sigma^{-\frac{1}{2}} V^T(:, 1), \quad \hat{H} = U(1, :) \Sigma^{\frac{1}{2}}, \end{aligned}$$

where $V^T(:, 1)$ and $U(1, :)$ stand for the first column and row of V^T and U , respectively.

- 3) Obtain a continuous-time realization $(\hat{A}, \hat{\mathbf{x}}_0, \hat{C})$ by

$$\hat{A} = \frac{\log(\hat{F})}{T_s}, \quad \hat{\mathbf{x}}_0 = \hat{\mathbf{p}}_0, \quad \hat{C} = \hat{H}.$$

where $\hat{A} \in \mathbb{R}^{\nu \times \nu}$, $\hat{\mathbf{x}}_0 \in \mathbb{R}^\nu$ and $\hat{C} \in \mathbb{R}^{q \times \nu}$. As shown later, ideally we can construct a realization such that $\mathbf{z}(kT_s) = \hat{\mathbf{y}}[k]$. This implies that we can use the realization (6) as a surrogate of the measurement $\hat{\mathcal{Y}}$ in the parameter set identification.

In the second stage, the parameter sets $\bar{\mathcal{C}}$ and $\bar{\mathcal{I}}$ are identified using the realization (6). We define the distance between the model and the measurement by

$$\mathcal{D}(\hat{\mathcal{Y}}, \mathcal{Y}(\theta)) := \int_0^\infty \|\mathbf{y}(t; \theta) - \mathbf{z}(t)\|_2^2 dt. \quad (7)$$

We then search parameter sets $\bar{\mathcal{C}}$ and $\bar{\mathcal{I}}$ over the given search space $\bar{\Theta}$. In Section III-C, we propose an algorithm that approximates the regions of consistent and inconsistent parameters using multiple convex polytopes.

B. Construction of a realization

We construct the realization (6) using the celebrated Ho-Kalman method [16], or equivalently eigensystem realization algorithm (ERA) [17]. Let $H_{\ell_1, \ell_2}(m) \in \mathbb{R}^{\ell_1 q \times \ell_2}$ denote a block Hankel matrix

$$H_{\ell_1, \ell_2}(m) := \begin{bmatrix} \hat{\mathbf{y}}[m] & \hat{\mathbf{y}}[m+1] & \cdots & \hat{\mathbf{y}}[m+\ell_1] \\ \hat{\mathbf{y}}[m+1] & \hat{\mathbf{y}}[m+2] & \cdots & \hat{\mathbf{y}}[m+\ell_1+1] \\ \vdots & \vdots & \ddots & \vdots \\ \hat{\mathbf{y}}[m+\ell_2] & \hat{\mathbf{y}}[m+\ell_2+1] & \cdots & \hat{\mathbf{y}}[m+\ell_1+\ell_2] \end{bmatrix}.$$

The realization (6) is then obtained by Algorithm 1. It is worth noting that Step 3) of Algorithm 1 can be replaced with other discrete-to-continuous time conversion methods

such as Tustin transform. The presented version is the inverse zero-order-hold transform ¹.

It should be noted that the order of the realization ν is determined based on the number of non-zero singular values. It is known that the realization given by Algorithm 1 is minimum among those LTI systems satisfying $z(kT_s) = \hat{\mathbf{y}}[k]$ for $k = 0, 1, 2, \dots$, when the underlying dynamics of the measurement $\hat{\mathcal{Y}}$ is given by a LTI system [17], [19].

In practice, however, the singular value does not completely decay to zero due to the underlying nonlinearity, measurement noise and numerical errors. In such cases, the order of the system ν can be determined so that the truncated singular values $\sigma_{\nu+1}, \sigma_{\nu+2}, \dots$ are sufficiently small, since each Hankel singular value represents the contribution of the corresponding subspace to the dynamics. We show in Section IV that this truncation helps avoid overfitting and extract the true underlying dynamics.

C. Inner approximation of consistent and inconsistent parameter regions

Once the realization is constructed, the next step is to compute the parameter sets $\bar{\mathcal{C}}$ and $\bar{\mathcal{I}}$. We here introduce LMI feasibility problems that determine whether a given set of parameters lies inside either of \mathcal{C} and \mathcal{I} based on relaxation. The parameter sets are then identified by iteratively solving the feasibility problems for different sets of parameters.

Let $\Theta = \{\theta^{(i)}\}_{i=1}^{\zeta}$ denote a set of ζ distinct points in the parameter search space $\bar{\Theta}$. The following proposition states that the convex hull of Θ lies inside the consistent (or inconsistent) parameter region, *i.e.*, $\text{co}(\Theta) \subseteq \mathcal{C}$ (or \mathcal{I} , respectively), if the LMIs are feasible.

Proposition 1. *Consider the linearized model of biochemical systems (2) and the realization (6) generated by Algorithm 1. Given a set of parameter points $\Theta = \{\theta^{(i)}\}_{i=1}^{\zeta}$, suppose $(A(\theta), C)$ is observable for all $\theta \in \text{co}(\Theta)$. Then, the distance $\mathcal{D}(\hat{\mathcal{Y}}, \mathcal{Y}(\theta))$ satisfies $\mathcal{D}(\hat{\mathcal{Y}}, \mathcal{Y}(\theta)) > \gamma$ for all $\theta \in \text{co}(\Theta)$ and a given γ , if there exists a symmetric matrix Q such that*

$$\begin{aligned} (\star) &= Q \begin{bmatrix} A(\theta^{(i)}) & O \\ O & \hat{A} \end{bmatrix} + \begin{bmatrix} A(\theta^{(i)}) & O \\ O & \hat{A} \end{bmatrix}^T Q + \begin{bmatrix} C^T C & -C^T \hat{C} \\ -\hat{C}^T C & \hat{C}^T \hat{C} \end{bmatrix} \succ O, \\ (*) &= [\mathbf{x}_0^T, \hat{\mathbf{x}}_0^T] Q \begin{bmatrix} \mathbf{x}_0 \\ \hat{\mathbf{x}}_0 \end{bmatrix} > \gamma \end{aligned}$$

with $i = 1, 2, \dots, \zeta$. Accordingly, $\mathcal{D}(\hat{\mathcal{Y}}, \mathcal{Y}(\theta)) \leq \gamma$ for all $\theta \in \text{co}(\Theta)$ and a given γ , if there exists a symmetric matrix Q such that $(\star) \preceq O$ and $(*) \preceq \gamma$.

The proof of this proposition is omitted due to the limitation of the space but it can be found in [20]. The idea of Proposition 1 is that we consider an error system whose output is given by $e := \mathbf{y} - \mathbf{z}$ and calculate the upper

¹When \hat{F} has a negative real eigenvalue, careful treatment is necessary to avoid the logarithm of a negative real number [18]. The command 'd2c' in MATLAB implements the conversion of a negative real number to complex conjugate pairs and is useful for practical implementation.

Algorithm 2: Iterative method to compute $\bar{\mathcal{C}}$ and $\bar{\mathcal{I}}$

Input: Realization $(\hat{A}, \hat{\mathbf{x}}_0, \hat{C})$, minimum block size ϵ
Output: Approximated consistent and inconsistent parameter sets, $\bar{\mathcal{C}}$ and $\bar{\mathcal{I}}$.
Initialization: Let $\bar{\mathcal{C}} = \emptyset$ and $\bar{\mathcal{I}} = \emptyset$. Define a family of parameter sets ϑ and let $\vartheta = \{\Theta_0\}$ with Θ_0 satisfying $\bar{\Theta} \subseteq \text{co}(\Theta_0)$
while ϑ is not empty **do**
 Pick a parameter set Θ from ϑ .
 Remove Θ from ϑ .
 if $\text{LMI}_{\succ}(\Theta)$ is feasible **then**
 Add $\text{co}\{\Theta\}$ to the inconsistent parameter set $\bar{\mathcal{I}}$.
 else
 if $\text{LMI}_{\preceq}(\Theta)$ is feasible **then**
 Add $\text{co}\{\Theta\}$ to the consistent parameter set $\bar{\mathcal{C}}$.
 else
 Divide the region Θ into Θ_1 and Θ_2 .
 Add Θ_1 to ϑ if $\text{vol}(\Theta_1) \geq \epsilon$.
 Add Θ_2 to ϑ if $\text{vol}(\Theta_2) \geq \epsilon$.
 end
 end
end

and lower bounds of impulse-to-energy gain $\int_0^\infty \|e\|^2 dt = \mathcal{D}(\hat{\mathcal{Y}}, \mathcal{Y}(\theta))$. Note that the LMIs for the upper bound $\mathcal{D}(\hat{\mathcal{Y}}, \mathcal{Y}(\theta)) \leq \gamma$ is a version of the previous result (see Chapter 6 of [21]) in that the positive semi-definiteness of Q is replaced with the observability assumption.

Proposition 1 allows us to verify that a given continuous parameter region is contained in \mathcal{C} or \mathcal{I} by solving the LMI feasibility problems. Thus, inner approximations of \mathcal{C} and \mathcal{I} can be obtained by iteratively solving the feasibility problems for different sets of Θ .

Remark 2. The gap between the necessity and the sufficiency can be arbitrarily small in Proposition 1 as the size of the region $\text{co}(\Theta)$ becomes smaller. In fact, Proposition 1 is necessary and sufficient when Θ is a single point in the parameter space (see Theorem 4.6.1 of [22]). \square

A relatively simple way to screen the parameter sets Θ would be to partition the search space $\bar{\Theta}$ into equally sized blocks and solve the feasibility problems for each block. This, however, requires combinatorial scan over the parameter search space $\bar{\Theta}$. Instead, binary space partitioning can be used to efficiently compute the consistent and inconsistent parameter sets. Detailed steps are shown in Algorithm 2, where $\text{LMI}_{\succ}(\cdot)$ and $\text{LMI}_{\preceq}(\cdot)$ stand for the LMI feasibility problem in Proposition 1 with $(\star) \succ O, (*) > O$ and $(\star) \preceq O, (*) \leq 0$, respectively. We here use the k-d tree [23] for efficiently bisecting n -dimensional space. Note that the LMI feasibility problems in Proposition 1 are independent of the space partitioning method.

Remark 3. In many cases, the consistent parameter set \mathcal{C} is much smaller than the inconsistent parameter set \mathcal{I} . Thus, it

is more efficient to test $\text{LMI}_{>}(\cdot)$ first as shown in Algorithm 2. An alternative of Algorithm 2 would be to identify the inconsistent set $\bar{\mathcal{L}}$, or outer approximation of \mathcal{C} , in the first run, then search for $\bar{\mathcal{C}}$ over the complementary set of $\bar{\mathcal{L}}$, which would be much smaller than the original search space $\bar{\Theta}$. \square

IV. EXAMPLE: A SIMPLE ENZYMATIC REACTION

In this section, we demonstrate the proposed parameter set identification method using a numerically simulated measurement of a simple enzymatic reaction system.

A. Description of the system

We consider the following set of enzymatic reactions



where $S, E, S:E$ and P represent an input substrate, an enzyme, a complex of S and E and a final product, respectively, and the symbol \emptyset stands for the degradation of a molecule. The propensity functions of these reactions are obtained based on the law of mass action as

$$\begin{aligned} \varphi_1(\boldsymbol{\xi}, k_f) &= k_f[S](E_0 - [S:E]), & \varphi_2(\boldsymbol{\xi}, k_r) &= k_r[S:E], \\ \varphi_3(\boldsymbol{\xi}, k_p) &= k_p[S:E], & \varphi_4(\boldsymbol{\xi}, d) &= d[P] \end{aligned}$$

where $\boldsymbol{\xi} := [[S], [S:E], [P]]^T$, and E_0 is a given total amount of enzyme, *i.e.*, $E_0 := [S] + [S:E]$. The corresponding stoichiometry matrix M is also obtained as

$$M = \begin{bmatrix} -1 & 1 & 0 & 0 \\ 1 & -1 & -1 & 0 \\ 0 & 0 & 1 & -1 \end{bmatrix}. \quad (8)$$

The dynamics of the reaction $\dot{\boldsymbol{\xi}} = M\boldsymbol{\varphi}(\boldsymbol{\xi}; \boldsymbol{\theta})$ can then be modeled as

$$\begin{aligned} \frac{d}{dt}[S] &= -k_f[S](E_0 - [S:E]) + k_r[S:E] \\ \frac{d}{dt}[S:E] &= k_f[S](E_0 - [S:E]) - k_r[S:E] - k_p[S:E] \\ \frac{d}{dt}[P] &= k_p[S:E] - d_p[P]. \end{aligned} \quad (9)$$

Our goal here is to obtain a set of the rate parameters $\boldsymbol{\theta} := [k_f, k_r, k_p, d]$ that is consistent or inconsistent with experimental data. Suppose a perturbation experiment was conducted around an equilibrium point $\boldsymbol{\xi} = [0, 0, 0]^T$ by adding the input substrate S by one unit at $t = 0$, *i.e.*, $[S](0) = 1$, and a time-series of the product concentration $[P]$ was measured over $t = 155$ minutes with sampling period $T_s = 5$ minutes as shown in the shaded region of Fig. 2 (left). The time-series data were produced by simulating the model (9) with

$$\boldsymbol{\theta} = [k_f, k_r, k_p, d]^T = [0.07, 0.0035, 0.056, 0.007]^T \quad (10)$$

and $E_0 = 1.12$. Multiplicative Gaussian noise was added to the simulation output by $[P](1 + 0.05\delta)$, where δ is a random variable drawn from a standard Gaussian distribution. In Fig. 2 (left), the time-series after $t = 155$ is also shown for later convenience, but it is not used in the following identification steps.

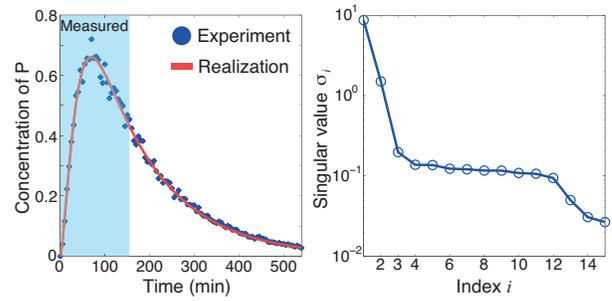


Fig. 2. (Left) Simulated experimental data and the output of the realization (6). (Right) Hankel singular values.

B. Parameter set identification

Given the time-series measurement, we first construct the realization (6). Figure 2 (right) shows the singular values of the Hankel matrix $H_{15,15}(0)$. We can see that the decay rate of the singular values significantly decreases after σ_4 , implying that the first three subspaces of U and V contain essential information on the dynamics. We here choose $\nu = 3$ and produce a realization by Algorithm 1. The realization is specifically obtained as

$$\hat{A} = 10^{-2} \times \begin{bmatrix} 0.283 & -1.67 & 0.306 \\ 1.67 & -2.39 & 3.46 \\ 0.306 & -3.46 & -8.33 \end{bmatrix},$$

$$\hat{\boldsymbol{x}}_0 = [-0.584, 0.630, 0.209]^T, \hat{C} = [-0.584, -0.630, 0.209].$$

The output of the realization \boldsymbol{z} is shown in Fig. 2 (left). We can see that the constructed realization tracks the time course of P not only for the measured interval $0 \leq t \leq 155$ but also for $t > 155$. Moreover, the trajectory of \boldsymbol{z} is smoother than the measurement, since the higher order dynamics were eliminated by the truncation of the Hankel singular values. Note that this allows us to reduce the risk of overfitting.

The consistent and inconsistent parameter sets $\bar{\mathcal{C}}$ and $\bar{\mathcal{L}}$ are then obtained by applying Algorithm 2. Here we set the search region $\bar{\Theta}$ as

$$\bar{\Theta} = \{\boldsymbol{\theta} \mid k_f \in [0.01, 0.1], k_r \in [0.001, 0.01], k_p \in [0.01, 0.1], d \in [0.001, 0.01]\},$$

and the error bound γ as

$$\gamma = 0.05 \times \int_0^\infty z^2 dt = 2.839, \quad (11)$$

so that the threshold is 5% of the relative error between \boldsymbol{z} and \boldsymbol{y} . The matrix $A(\boldsymbol{\theta})$ in Proposition 1 can be easily calculated from the definition (3), and the perturbation at $t = 0$ is $\boldsymbol{x}_0 = [1, 0, 0]$.

Figure 3 illustrates a slice of the identified parameter sets $\bar{\mathcal{C}}$ and $\bar{\mathcal{L}}$ at $k_r = 0.0035$. In the figure, each cuboid corresponds to the parameter region Θ for which the feasibility problem in Proposition 1 is solved. We can see that relatively large parameter regions can be classified into $\bar{\mathcal{C}}$ or $\bar{\mathcal{L}}$ by a single feasibility problem when the regions are far from the actual

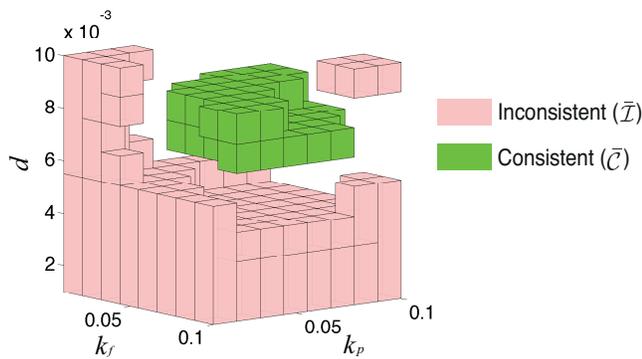


Fig. 3. A three-dimensional slice of the identified sets \bar{C} and \bar{I} .

parameter (10). This reduces unnecessary mesh of the parameter space and saves the computation time. Figure 1 (top) is another slice in two dimensions at $(k_f, k_r) = (0.09, 0.0025)$. These identification results are useful for robust biocircuit design since they provide the landscape of the parameter uncertainty. We used SeDuMi [24] and YALMIP [25] to solve LMIs in Proposition 1.

In order to confirm that the parameters were correctly identified, we simulated the model (9) with 100 different parameters that were randomly selected from the regions \bar{C} and \bar{I} , respectively. The result is shown in Fig. 1 (bottom). The figure illustrates that the identified parameter sets \bar{C} and \bar{I} are consistent and inconsistent with the experimental data, respectively.

V. CONCLUSION

We have proposed a novel parameter set identification method that utilizes time-series data of a perturbation experiment. Given the error bound of a model and a time-series measurement, the proposed method can efficiently classify parameter regions into consistent and inconsistent sets by iteratively solving the LMI feasibility problems and bisecting the parameter regions. We have shown that the utilization of the state space realization as a surrogate of experimental data allows us to overcome the non-convexity of the problem and to handle time-series data based on a solid theoretical foundation.

In our future work, the proposed framework will be extended so that it can also handle time-series data around non-equilibrium state (see Remark 1) and uncertain or unknown initial values. Experimental work is also in progress to demonstrate the proposed method using data obtained from an in vitro cell-free expression system.

Acknowledgements: The authors would like to thank Anandh Swaminathan at California Institute of Technology for helpful comments to improve the manuscript.

REFERENCES

- [1] Y. Shimizu, *et al.*, "Cell-free translation reconstituted with purified components," *Nature Biotechnology*, vol. 19, pp. 751–755, 2001.
- [2] J. Shin and V. Noireaux, "An E. coli cell-free expression toolbox: application to synthetic gene circuits and artificial cells," *ACS Synthetic Biology*, vol. 1, no. 1, pp. 29–41, 2012.
- [3] Z. Z. Sun, *et al.*, "Protocols for implementing an escherichia coli based TX-TL cell-free expression system for synthetic biology," *Journal of Visualized Experiments*, vol. 79, no. e50762, 2013.
- [4] N. van Riel and E. Sontag, "Parameter estimation in models combining signal transduction and metabolic pathways: The dependent input approach," *IET Systems Biology*, vol. 153, 2006.
- [5] A. Dayarian, *et al.*, "Shape, size, and robustness: feasible regions in the parameter space of biochemical networks," *PLoS Computational Biology*, vol. 5, 2009.
- [6] T. Toni, *et al.*, "Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems," *Journal of the Royal Society Interface*, vol. 6, no. 31, pp. 187–202, 2009.
- [7] H. El-Samad, *et al.*, "Model validation and robust stability analysis of the bacterial heat shock response using SOSTOOLS," in *Proceedings of the 42nd Conference on Decision and Control*, 2003, pp. 3766–3771.
- [8] J. Anderson and A. Papachristodoulou, "On validation and invalidation of biological models," *BMC Bioinformatics*, vol. 10, no. 132, 2009.
- [9] S. Prajna and A. Jadbabaie, "Safety verification of hybrid systems using barrier certificates," in *Hybrid Systems: Computation and Control*. Springer Berlin Heidelberg, 2004, vol. 2993, pp. 477–492.
- [10] A. Papachristodoulou, *et al.*, *SOSTOOLS: Sum of squares optimization toolbox for MATLAB*, <http://arxiv.org/abs/1310.4716>, 2013, available from <http://www.eng.ox.ac.uk/control/sostools>.
- [11] L. Kuepfer, U. Sauer, and P. A. Parrilo, "Efficient classification of complete parameter regions based on semidefinite programming," *BMC Bioinformatics*, vol. 8, no. 12, 2007.
- [12] P. Rumschinski, *et al.*, "Set-based dynamical parameter estimation and model invalidation for biochemical reaction networks," *BMC Systems Biology*, vol. 4, no. 69, 2010.
- [13] J. Hasenauer, *et al.*, "Parameter identification, experimental design and model falsification for biological network models using semidefinite programming," *IET Systems Biology*, vol. 4, no. 2, pp. 119–130, 2010.
- [14] P. Rumschinski, *et al.*, "Influence of discretization errors on set-based parameter estimation," in *Proceedings of the 49th IEEE Conference on Decision and Control*, 2010, pp. 296–301.
- [15] P. A. Iglesias and B. P. Ingalls, *Control theory and systems biology*. The MIT Press, 2009.
- [16] B. L. Ho and R. E. Kalman, "Effective construction of linear state-variable models from input-output functions," *Regelungstechnik*, vol. 14, no. 12, pp. 545–548, 1966.
- [17] J. N. Juang and R. S. Pappa, "An eigensystem realization algorithm for modal parameter identification and model reduction," *Journal of Guidance, Control and Dynamics*, vol. 8, no. 5, pp. 620–627, 1985.
- [18] I. Kollár, G. F. Franklin, and R. Pintelon, "On the equivalence of z-domain and s-domain models in system identification," in *Proceedings of the IEEE Instrumentation and Measurement Technology Conference*, 1996, pp. 14–19.
- [19] Y. Hori, M. H. Khammash, and S. Hara, "Efficient parameter identification for stochastic biochemical networks using a reduced-order realization," in *Proceedings of the European Control Conference*, 2013, pp. 4154–4159.
- [20] Y. Hori and R. M. Murray, "A state-space realization approach to set identification of biochemical kinetic parameters," California Institute of Technology, Tech. Rep., 2014, available at http://www.cds.caltech.edu/~murray/papers/2014k_hm15-ecc.html.
- [21] S. Boyd, *et al.*, *Linear matrix inequalities in system and control theory*. SIAM, 1994.
- [22] R. E. Skelton, T. Iwasaki, and K. Grigoriadis, *A unified algebraic approach to linear control design*. CRC Press, 1997.
- [23] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.
- [24] J. F. Sturm, "Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones," *Optimization methods and software*, vol. 11, pp. 625–653, 1999.
- [25] J. Löfberg, "Yalmip : A toolbox for modeling and optimization in MATLAB," in *Proceedings of the CACSD Conference*, 2004. [Online]. Available: <http://users.isy.liu.se/johanl/yalmip/>