Euro. Jnl of Applied Mathematics (2021), *vol.* 32, *pp.* 865–887

doi:10.1017/S0956792520000492

Identifiability of stochastically modelled reaction networks

GERMAN ENCISO[1], RADEK ERBAN[2] and JINSU KIM[3]

[1]*Department of Mathematics, University of California, Irvine, CA 92697, USA*
e-mail: enciso@uci.edu
[2]*Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford, OX2 6GG, UK*
e-mail: erban@maths.ox.ac.uk
[3]*NSF-Simons Center for Multiscale Cell Fate Research, University of California, Irvine, CA 92697, USA*
e-mail: jinsu.kim@uci.edu

(*Received 2 June 2020*; *revised 23 December 2020*; *accepted 30 December 2020*; *first published online 15 February 2021*)

Chemical reaction networks describe interactions between biochemical species. Once an underlying reaction network is given for a biochemical system, the system dynamics can be modelled with various mathematical frameworks such as continuous-time Markov processes. In this manuscript, the identifiability of the underlying network structure with a given stochastic system dynamics is studied. It is shown that some data types related to the associated stochastic dynamics can uniquely identify the underlying network structure as well as the system parameters. The accuracy of the presented network inference is investigated when given dynamical data are obtained via stochastic simulations.

Key words: Chemical reaction networks, identifiability, stochastic simulation, network inference, Gillespie algorithm

2020 Mathematics Subject Classification: 60J27 (Primary); 62M10, 92C42, 92E20, 93B30 (Secondary)

1 Introduction

To study the properties and dynamics of a system of reacting biochemical species, a network representation is often used to describe the interactions between the chemical species involved. A reaction network represents the system behaviour with reactions (directed edges) between complexes (nodes) [7, 14]. Each reaction in a reaction network indicates loss or gain of the amount of the corresponding chemical species. Systems of ordinary differential equations (ODEs) are traditionally used for modelling the time evolution of concentrations of chemical species in reaction network theory [13, 3]. Since biochemical systems may contain chemical species with low copy numbers, stochastic approaches are often used for modelling their behaviour [12]. Stochastic models of homogeneous (space independent) chemical reaction networks are written as continuous-time discrete space Markov chains [1, 2].

In some applications, the underlying network structure may be unknown but information on the associated dynamics is given [9, 19]. The main focus of this paper is to identify the unknown

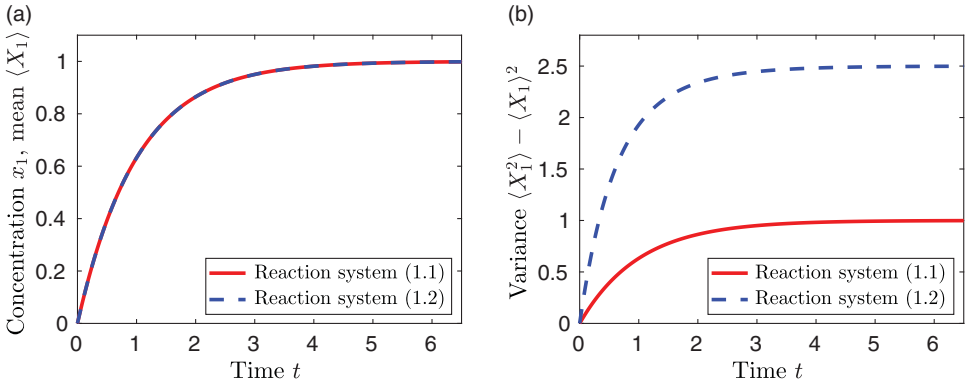

FIGURE 1. (a) The solution of ODE (1.3) with initial condition $x_1(0) = 0$. (b) Variance of the number of molecules of chemical species X_1 for the chemical system (1.1) (red solid line) and the chemical system (1.2) (blue dashed line).

network structure of a stochastic reaction system using dynamical information. Identifiability of reaction systems has been studied under deterministic ODE-based modelling by Craciun and Pantea [9] and Szederkényi et al. [29]. They present examples of reaction systems that admit the same deterministic dynamical system but have different network structure and parameters. In Figure 1, we illustrate this lack of identifiability using two simple reaction systems. They both include one chemical species X_1, which is subject to two chemical reactions:

$$\text{first reaction system:} \qquad \emptyset \xrightarrow{1} X_1, \qquad X_1 \xrightarrow{1} \emptyset, \tag{1.1}$$

$$\text{second reaction system:} \qquad \emptyset \xrightarrow{1/4} 4X_1, \qquad X_1 \xrightarrow{1} \emptyset. \tag{1.2}$$

Denoting $x_1(t)$ the concentration of the chemical species X_1 and using mass-action deterministic description, the time evolution of both reaction systems (1.1) and (1.2) is described by the same ODE:

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = 1 - x_1. \tag{1.3}$$

Solving the ODE (1.3) with the initial condition $x_1(0) = 0$, we obtain $x_1(t) = 1 - \exp[-t]$, which is plotted in Figure 1(a). Since both reaction systems (1.1) and (1.2) contain only reactions of zero and first order, we can analytically solve the chemical master equation corresponding to the stochastic model [15, 18]. We obtain that the mean number of molecules, $\langle X_1 \rangle$, is for both systems given as a solution of the ODE system (1.3). In the case of the first reaction system (1.1), X_1 is Poisson distributed at every time t [12, 18]. Therefore, the variance $\langle X_1^2 \rangle - \langle X_1 \rangle^2$ is equal to the mean $\langle X_1 \rangle = 1 - \exp[-t]$. In Figure 1(b), we show that it differs from the variance obtained using the second reaction system (1.2), which is given as $\langle X_1^2 \rangle - \langle X_1 \rangle^2 = (5 - 2\exp[-t] - 3\exp[-2t])/2$.

Our example illustrates that the dynamics obtained by the ODE model (1.3) cannot be used to distinguish between reaction systems (1.1) and (1.2) and the reaction network is therefore not identifiable in the deterministic context. However, since their stochastic models do differ (as shown in Figure 1(b)), we have potential to use the stochastic data to distinguish between the reaction systems (1.1) and (1.2). This peculiar behaviour is not restricted to our illustrative example. Plesa et al. [28] showed that any reaction network can be redesigned in such a way

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

that the deterministic dynamics are preserved, while the controllable state-dependent noise is introduced into the stochastic dynamics. In this way, one can systematically obtain a family of reaction networks, which have qualitatively different stochastic dynamics, but they are described by the same deterministic model [28]. In applications, the long-term dynamics of some gene regulatory networks (involving multiple timescales) can consist of a unique attractor at the deterministic level (unistability), while the long-term probability distribution at the stochastic level may display multiple maxima (multimodality) [10, 25].

In this paper, we explore how the discrete nature of the associated mass-action stochastic system can help uncover the underlying reaction network. For a given continuous-time Markov chain (CTMC), we quantify the amount of transition rate information needed to uniquely identify the underlying network and the system parameters. For practical implementation of network inference, the presented approach can be used to infer the underlying reaction network with transition data obtained from stochastic simulations. The accuracy of this network inference idea is also investigated.

For each reaction, the reaction intensity determines the likelihood of the reaction taking place. The reaction intensity is proportional to a positive constant; the so-called rate constant. The numbers $1/4$ and 1 in our illustrative reaction systems (1.2) are examples of rate constants. The rate constants can alter the system behaviour significantly and correspond to qualitative differences between deterministic and stochastic descriptions, for example, for systems close to bifurcations of deterministic ODEs [11, 27]. When the reaction network topology is given, the rate constants often need to be estimated as missing parameters. Numerous different statistical and mathematical techniques have been employed in the literature for parameter estimation using dynamical data, such as information theory [19], Bayesian statistics [8, 16, 5, 33], system identification theory [31], machine learning [4], and tensor-structured parametric analysis [22].

In addition to parameter estimation, the underlying network topology is also often unknown or only partially known. There have been a number of methods developed in the literature to infer network information [6, 21, 32]. For instance, Wang et al. [32] study deterministic network inference using multiplex flow cytometry experimental data and toric systems theory. Chattopadhyay et al. [6] proposed a novel inference method for stochastic reaction systems with convex polytopes, which are formed by combinations of reaction vectors captured within a short time window. Other papers focus on statistical information and Bayesian analysis to infer networks of correlations among species [21, 17, 23, 24, 30], but, to our knowledge, there is no previous work that characterises when the transition data of a stochastic system can be used to completely identify the underlying reaction network.

The underlying network structure of a dynamical system may not be uniquely identified if a CTMC is restricted to a subset of the state space because of a conservation law. In this case, the stochastic system can be associated with two different reaction networks, as illustrated in Example 3.2. In Section 3, we prove that the network topology and the system parameters can be uniquely identified provided that we have full dynamic information in a sufficiently large finite region of the state space.

To formulate our results, we begin by introducing our notation in Section 2. In Section 3, we present the main algorithm that uses the transition rates of a given CTMC to infer the underlying network structure and parameters. In Section 4, we show that a general CTMC with polynomial transition rates can be identifiable as a mass-action reaction system. In Section 5, we investigate how accurately the underlying network structure and system parameters can be identified using given stochastic dynamical information about the transition rates.

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

2 Notation and terminology

In this section, we introduce our notation and basic definitions that are used throughout the rest of our manuscript.

2.1 Reaction networks

A reaction network $(\mathcal{S}, \mathcal{C}, \mathcal{R})$ consists of a collection of *species* $\mathcal{S}$, *complexes* $\mathcal{C}$, and *reactions* $\mathcal{R}$. Each reaction is of the form:

$$\sum_{i=1}^{d} y_i X_i \longrightarrow \sum_{i=1}^{d} y_i' X_i, \tag{2.1}$$

where X_i, $i = 1, 2, \ldots, d$, are species, and linear combinations $\sum_{i=1}^{d} y_i X_i$ and $\sum_{i=1}^{d} y_i' X_i$ of species with non-negative integers y_i are complexes. We interchangeably denote by $\mathbf{y} = (y_1, y_2, \ldots, y_d)$ a complex $\sum_{i=1}^{d} y_i X_i$. In the same way, we denote by $\mathbf{y} \to \mathbf{y}'$ the reaction (2.1).

Example 2.1 The typical enzyme-substrate system can be described with a reaction network:

$$X_1 + X_2 \underset{\kappa_2}{\overset{\kappa_1}{\rightleftarrows}} X_3 \xrightarrow{\kappa_3} X_1 + X_4,$$

where the species X_1, X_2, X_3 and X_4 represent the enzyme, substrate, enzyme-substrate complex and product, respectively. For this system, we have $\mathcal{S} = \{X_1, X_2, X_3, X_4\}$, $\mathcal{C} = \{X_1 + X_2, X_3, X_1 + X_4\}$ and $\mathcal{R} = \{X_1 + X_2 \to X_3, X_3 \to X_1 + X_2, X_3 \to X_1 + X_4\}$. Each reaction in $\mathcal{R}$ is associated with the corresponding rate constant κ_1, κ_2 and κ_3.

The time evolution of the concentration of species $X_i \in \mathcal{S}$ is described by a system of ODEs as:

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t}(t) = \sum_{\mathbf{y} \to \mathbf{y}' \in \mathcal{R}} f_{\mathbf{y} \to \mathbf{y}'}(\mathbf{x}(t))(\mathbf{y}' - \mathbf{y}),$$

where $f_{\mathbf{y} \to \mathbf{y}'}$ are positive functions representing the 'weight' of the reaction $\mathbf{y} \to \mathbf{y}'$ at each state. Considering *mass-action kinetics*, we have

$$f_{\mathbf{y} \to \mathbf{y}'}(\mathbf{x}) = \kappa_{\mathbf{y} \to \mathbf{y}'} \mathbf{x}^{\mathbf{y}},$$

where $\mathbf{u}^{\mathbf{v}} = \prod_{i=1}^{d} u_i^{v_i}$ for vectors $\mathbf{u}$ and $\mathbf{v}$ with non-negative entries. The positive constant $\kappa_{\mathbf{y} \to \mathbf{y}'}$ forms the reaction rate for the reaction, and it constitutes one of the parameters of the reaction network. We include this reaction rate by placing it above the arrow of the associated reaction $\mathbf{y} \to \mathbf{y}'$ as in Example 2.1.

2.2 Stochastic description of reaction networks

We model the number of molecules of each chemical species in a reaction network by a CTMC defined on the d-dimensional integer lattice:

$$\mathbb{Z}^d_{\geqslant 0} = \left\{ \mathbf{x} \in \mathbb{Z}^d \,\middle|\, x_i \geqslant 0 \text{ for } i = 1, 2, \ldots, d \right\}. \tag{2.2}$$

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

Using $\mathbf{X}(t) = [X_1(t), X_2(t), \ldots, X_d(t)]$ to denote the number of molecules in a reaction network $(\mathcal{S}, \mathcal{C}, \mathcal{R})$, the corresponding transition rates are defined as:

$$P\big(\mathbf{X}(t+\Delta t) = \mathbf{x} + \mathbf{z} \,\big|\, \mathbf{X}(t) = \mathbf{x}\big) \;=\; \sum_{\substack{\mathbf{y}\to\mathbf{y}'\in\mathcal{R}\\ \mathbf{y}'-\mathbf{y}=\mathbf{z}}} \lambda_{\mathbf{y}\to\mathbf{y}'}(\mathbf{x})\,\Delta t + o(\Delta t),$$

where $o(\Delta t) \to 0$, as $\Delta t \to 0$. We denote by $\mathcal{Z} = \{\mathbf{z} = \mathbf{y}' - \mathbf{y} : \mathbf{y} \to \mathbf{y}' \in \mathcal{R}\}$ the set of the transition vectors of the CTMC $\mathbf{X}(t)$. The function $\lambda_{\mathbf{y}\to\mathbf{y}'} \geqslant 0$ is called the *intensity* of the reaction $\mathbf{y} \to \mathbf{y}'$ and it satisfies

$$\lambda_{\mathbf{y}\to\mathbf{y}'}(\mathbf{x}) > 0 \quad \text{if and only if } x_i \geqslant y_i \text{ for each } i = 1,\ 2,\ \ldots,\ d. \tag{2.3}$$

We say that a reaction $\mathbf{y} \to \mathbf{y}' \in \mathcal{R}$ is *turned off* at $\mathbf{x}$ if $\lambda_{\mathbf{y}\to\mathbf{y}'}(\mathbf{x}) = 0$. Otherwise, we call a reaction $\mathbf{y} \to \mathbf{y}' \in \mathcal{R}$ is *charged* at $\mathbf{x}$. Using (stochastic) mass-action kinetics, we define, for each $\mathbf{y} \to \mathbf{y}' \in \mathcal{R}$:

$$\lambda_{\mathbf{y}\to\mathbf{y}'}(\mathbf{x}) = \kappa_{\mathbf{y}\to\mathbf{y}'}\mathbf{x}^{(\mathbf{y})}, \quad \text{where } \mathbf{u}^{(\mathbf{v})} = \prod_{i=1}^{d} u_i(u_i-1)\cdots(u_i-v_i+1) \tag{2.4}$$

for vectors $\mathbf{u}, \mathbf{v} \in \mathbb{Z}^d_{\geqslant 0}$.

Let $\mathcal{K} = \{\lambda_{\mathbf{y}\to\mathbf{y}'} : \mathbf{y} \to \mathbf{y}' \in \mathcal{R}\}$ be the collection of given intensities for a reaction network $(\mathcal{S}, \mathcal{C}, \mathcal{R})$. Then the associated CTMC is fully characterised by the four tuple $(\mathcal{S}, \mathcal{C}, \mathcal{R}, \mathcal{K})$. Furthermore, since $\mathcal{S}$ and $\mathcal{C}$ can be fully determined using $\mathcal{R}$, the reaction system is fully characterised with $\mathcal{R}$ and $\mathcal{K}$. So in the rest of the paper, we let $(\mathcal{R}, \mathcal{K})$ represent both a reaction network and the associated CTMC, and we call $(\mathcal{R}, \mathcal{K})$ a (stochastic) reaction system.

A reaction network $(\mathcal{R}, \mathcal{K})$ is a subnetwork of another reaction network $(\mathcal{R}', \mathcal{K}')$ if $\mathcal{R} \subset \mathcal{R}'$ and $\lambda_{\mathbf{y}\to\mathbf{y}'} \equiv \lambda'_{\mathbf{y}\to\mathbf{y}'} \in \mathcal{K}$ for each $\mathbf{y} \to \mathbf{y}' \in \mathcal{R}$, where $\lambda_{\mathbf{y}\to\mathbf{y}'}$ and $\lambda'_{\mathbf{y}\to\mathbf{y}'}$ are the reaction intensities of $(\mathcal{R}, \mathcal{K})$ and $(\mathcal{R}', \mathcal{K}')$, respectively. We denote this relation as $(\mathcal{R}, \mathcal{K}) \subset (\mathcal{R}', \mathcal{K}')$. If two systems $(\mathcal{R}, \mathcal{K})$ and $(\mathcal{R}', \mathcal{K}')$ are identical, then $(\mathcal{R}, \mathcal{K}) \subseteq (\mathcal{R}', \mathcal{K}')$ and $(\mathcal{R}, \mathcal{K}) \subseteq (\mathcal{R}', \mathcal{K}')$, which we shortly denote by $(\mathcal{R}, \mathcal{K}) = (\mathcal{R}', \mathcal{K}')$.

2.3 Reaction order and ordering for $\mathbb{Z}^d_{\geqslant 0}$

As indicated in Section 2.1, we use vectors to represent complexes. Hence, for $\mathbf{y} \in \mathbb{Z}^d_{\geqslant 0}$ and $\mathbf{z} \in \mathbb{Z}^d$ such that $\mathbf{y} + \mathbf{z} \in \mathbb{Z}^d_{\geqslant 0}$, we denote by $\mathbf{y} \to \mathbf{y} + \mathbf{z}$ a reaction whose source complex is $\mathbf{y} = \sum_{i=1}^d y_i X_i$ and the product complex is $\mathbf{y} + \mathbf{z} = \sum_{i=1}^d (y_i + z_i)X_i$. For example, for $\mathbf{y} = (1, 2)$ and $\mathbf{z} = (-1, 1)$, the reaction $\mathbf{y} \to \mathbf{y} + \mathbf{z}$ represents $X_1 + 2X_2 \to 3X_2$. For $\mathbf{v} \in \mathbb{Z}^d_{\geqslant 0}$ and an integer N, we define

$$\mathbb{S}_N = \big\{\mathbf{x} \in \mathbb{Z}^d_{\geqslant 0} \,\big|\, \mathbf{x} \text{ satisfies } \|\mathbf{x}\|_1 \leqslant N\big\}, \tag{2.5}$$

$$\mathbb{S}_{\mathbf{v},N} = \big\{\mathbf{x} \in \mathbb{Z}^d_{\geqslant 0} \,\big|\, \mathbf{x} \text{ satisfies } \mathbf{v}\cdot\mathbf{x} = N\big\}, \tag{2.6}$$

where $\cdot$ is the canonical inner product in the Euclidean space. Transition rates of a given CTMC on those sets will play a critical role in the main algorithm of this paper for inferring an underlying network structure. Given two vectors $\mathbf{u} \in \mathbb{Z}^d_{\geqslant 0}$ and $\mathbf{v} \in \mathbb{Z}^d_{\geqslant 0}$, we define the *lexicographical* ordering for $\mathbb{Z}^d_{\geqslant 0}$ by

$$\mathbf{u} \prec \mathbf{v} \text{ if and only if there is } k \text{ such that } u_k < v_k \text{ and } u_i = v_i \text{ for all } i < k. \tag{2.7}$$

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

In particular, the d-dimensional simplex $\mathbb{S}_N$ has n elements which we enumerate in the lexicographical order, that is,

$$\mathbb{S}_N = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}, \quad \text{where} \quad \mathbf{x}^i \prec \mathbf{x}^j \text{ if } i < j \quad \text{and} \quad n = \binom{N+d}{d}. \tag{2.8}$$

A reaction $\mathbf{y} \to \mathbf{y}'$ is of *order* N if $\|\mathbf{y}\|_1 = N$. A reaction system $(\mathcal{R}, \mathcal{K})$ is of order N if the order of all reactions in $\mathcal{R}$ is at most N. A reaction $\mathbf{y} \to \mathbf{y}'$ is of $\mathbf{v}$-*order* N if $\mathbf{v} \cdot \mathbf{y} = N$. A reaction system $(\mathcal{R}, \mathcal{K})$ is of $\mathbf{v}$-order N if the $\mathbf{v}$-order of all reactions in $\mathcal{R}$ is at most N. For example, the reaction system in Example 2.1 is of order 2. However, if we use $\mathbf{v} = (0, 1, 1, 1)$, then the reaction system in Example 2.1 is of $\mathbf{v}$-order 1. In general, if $\mathbf{v} = (1, 1, \ldots, 1)$, the order and the $\mathbf{v}-$order of a reaction are the same.

3 Inference and identifiability of stochastic reaction systems

Main results of this section are stated as Theorems 3.1, 3.3 and 3.5.

3.1 Network inference using the transition rates

Our goal is to construct a reaction system $(\mathcal{R}, \mathcal{K})$ for given transition rates of a CTMC. First, we show that the knowledge of transition rates on a sufficiently large part of the state space uniquely determines the underlying reaction system.

Lemma 3.1 *Let $(\mathcal{R}, \mathcal{K})$ and $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ be two reaction systems of order N_1 and N_2, respectively. Suppose that there exists $N \geqslant \max\{N_1, N_2\}$ such that the two mass-action stochastic models associated with $(\mathcal{R}, \mathcal{K})$ and $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ have the same transition rates on $\mathbb{S}_N$. Then, $(\mathcal{R}, \mathcal{K}) = (\overline{\mathcal{R}}, \overline{\mathcal{K}})$.*

Proof Let $\mathbf{X}(t)$ and $\overline{\mathbf{X}}(t)$ be the CTMCs obtained using stochastic mass-action description of $(\mathcal{R}, \mathcal{K})$ and $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$, respectively. We denote by $\lambda_{\mathbf{y}\to\mathbf{y}'}$ and $\overline{\lambda}_{\mathbf{y}\to\mathbf{y}'}$ the transition rates of reactions $\mathbf{y} \to \mathbf{y}'$ associated with $(\mathcal{R}, \mathcal{K})$ and $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$, respectively. We denote states in $\mathbb{S}_N$ by (2.8). To prove the lemma by contradiction, we suppose that $(\mathcal{R}, \mathcal{K}) \neq (\overline{\mathcal{R}}, \overline{\mathcal{K}})$. Since the order of each reaction $\mathbf{y} \to \mathbf{y}'$ in $\mathcal{R} \cup \overline{\mathcal{R}}$ is less than or equal to N, it can be represented as $\mathbf{x}^k \to \mathbf{x}^k + \mathbf{z}$ for some transition vector $\mathbf{z}$ and for some $\mathbf{x}^k \in \mathbb{S}_N$. Since $(\mathcal{R}, \mathcal{K}) \neq (\overline{\mathcal{R}}, \overline{\mathcal{K}})$, there exists a transition vector $\mathbf{z}$ such that reaction $\mathbf{x}^j \to \mathbf{x}^j + \mathbf{z}$ is the first reaction (in the lexicographical ordering) which is formulated differently in reaction systems $(\mathcal{R}, \mathcal{K})$ and $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$. In other words, we have $\mathbf{x}^i \to \mathbf{x}^i + \mathbf{z} \in \mathcal{R} \cap \overline{\mathcal{R}}$ and $\lambda_{\mathbf{x}^i \to \mathbf{x}^i + \mathbf{z}} \equiv \overline{\lambda}_{\mathbf{x}^i \to \mathbf{x}^i + \mathbf{z}}$ for each $i < j$. Then at $\mathbf{x}^j \in \mathbb{S}_N$, the transition rate for $\mathbf{z}$ of the two systems are different, which is a contradiction to the assumption that both stochastic systems share the same transition rates on $\mathbb{S}_N$. □

Our main result is formulated as Theorem 3.1 below, but before we state this theorem, we begin with a simple example illustrated in Figure 2.

Example 3.1 Consider $d = 2$ and assume that the CTMC has a single transition vector $\mathbf{z} = (1, 1)$. Suppose that we are given data on transition rates $\lambda^*(\mathbf{x})$ of a CTMC defined on $\mathbb{Z}^2_{\geqslant 0}$ as the red numbers indicated in Figure 2. To construct the reaction network, we use $\mathbb{S}_2 = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^6\}$ defined by (2.5) and (2.8), i.e.

$$\mathbf{x}^1 = (0, 0), \quad \mathbf{x}^2 = (0, 1), \quad \mathbf{x}^3 = (0, 2), \quad \mathbf{x}^4 = (1, 0), \quad \mathbf{x}^5 = (1, 1), \quad \mathbf{x}^6 = (2, 0).$$

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

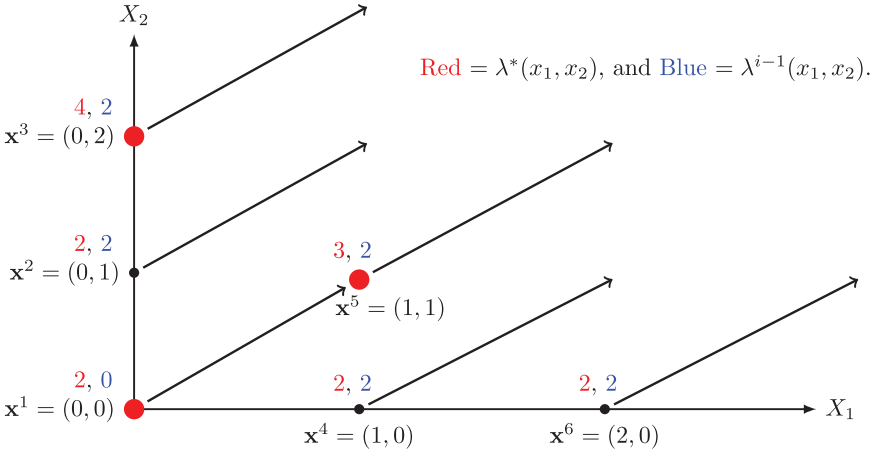

FIGURE 2. The procedure of inferring the underlying reaction system in Example 3.1. The red value at each state is the given transition rate associated with the transition vector $\mathbf{z} = (1, 1)$, indicated by black arrows. The blue is the value of λ^{i-1} updated at the previous state. Red dots indicate the states where $\lambda^*(x_1, x_2) - \lambda^{i-1}(x_1, x_2) > 0$.

Let $\lambda^0 \equiv 0$, $\mathcal{R}^0 = \emptyset$ and $\mathcal{K}^0 = \emptyset$. We iteratively define λ^i, $\mathcal{R}^i$ and $\mathcal{K}^i$ using given information at $\mathbf{x}^i$, for $i = 1, 2, \ldots, 6$. The outcome of this procedure is the transition rate function $\lambda = \lambda^6$, a set of reactions $\mathcal{R} = \mathcal{R}^6$, and a kinetic set $\mathcal{K} = \mathcal{K}^6$ such that

$$\sum_{\mathbf{y} \to \mathbf{y}' \in \mathcal{R}} \lambda_{\mathbf{y} \to \mathbf{y}'}(\mathbf{x}) = \lambda(\mathbf{x}) = \lambda^*(\mathbf{x}) \quad \text{for each } \mathbf{x} \in \mathbb{S}_2. \tag{3.1}$$

Since $\lambda^*(\mathbf{x}^1) - \lambda^0(\mathbf{x}^1) = 2 > 0$ at $\mathbf{x}^1 = (0, 0)$, the reaction $\emptyset \to X_1 + X_2$ must be included in $\mathcal{R}$ with the reaction intensity $\lambda_{\emptyset \to X_1 + X_2}(\mathbf{x}) = 2$. So we let

$$\lambda^1(\mathbf{x}) \equiv 2, \qquad \mathcal{R}^1 = \{\emptyset \to X_1 + X_2\} \qquad \text{and} \qquad \mathcal{K}^1 = \left\{\lambda_{\emptyset \to X_1 + X_2}(\mathbf{x}) = 2\right\}.$$

At the next state $\mathbf{x}^2 = (0, 1)$, we have $\lambda^*(\mathbf{x}^2) - \lambda^1(\mathbf{x}^2) = 0$, hence no additional reaction needs to be included in $\mathcal{R}$. Hence, we put $\lambda^2 = \lambda^1$, $\mathcal{R}^2 = \mathcal{R}^1$ and $\mathcal{K}^2 = \mathcal{K}^1$. Since $\lambda^*(\mathbf{x}^3) - \lambda^2(\mathbf{x}^3) = 2 > 0$ at $\mathbf{x}^3 = (0, 2)$, the reaction $2X_2 \to X_1 + 3X_2$ must be included in $\mathcal{R}$ with the reaction intensity $\lambda_{2X_2 \to X_1 + 3X_2}(\mathbf{x}) = x_2(x_2 - 1)$. So we let

$$\lambda^3(\mathbf{x}) = 2 + x_2(x_2 - 1), \qquad \mathcal{R}^3 = \{\emptyset \to X_1 + X_2,\ 2X_2 \to X_1 + 3X_2\}$$

and $\mathcal{K}^3 = \left\{\lambda_{\emptyset \to X_1 + X_2}(\mathbf{x}) = 2,\ \lambda_{2X_2 \to X_1 + 3X_2}(\mathbf{x}) = x_2(x_2 - 1)\right\}$. We iterate this procedure until the last state $\mathbf{x}^6 = (2, 0) \in \mathbb{S}_2$ as shown in Figure 2. Then, the outcome $(\mathcal{R}, \mathcal{K})$ is the following reaction system:

$$\emptyset \xrightarrow{2} X_1 + X_2, \qquad 2X_2 \xrightarrow{1} X_1 + 3X_2, \qquad X_1 + X_2 \xrightarrow{1} 2X_1 + 2X_2,$$

and the transition rate in the direction $\mathbf{z} = (1, 1)$ is

$$\begin{aligned} \lambda(\mathbf{x}) &= \lambda_{\emptyset \to X_1 + X_2}(\mathbf{x}) + \lambda_{2X_2 \to X_1 + 3X_2}(\mathbf{x}) + \lambda_{X_1 + X_2 \to 2X_1 + 2X_2}(\mathbf{x}) \\ &= 2 + x_2(x_2 - 1) + x_1 x_2. \end{aligned}$$

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

We have observed in Example 3.1 and Lemma 3.1 that a mass-action system of order $N=2$ can be characterised with the transition rates on $\mathbb{S}_N$. Next, we generalise this observation with a simple algorithm. Using the lexicographical order (2.8) of $\mathbb{S}_N$, for a given transition vector $\mathbf{z}$ and the associated transition rate $\lambda^*_{\mathbf{z}}$, we iteratively define $\lambda^0_{\mathbf{z}} \equiv 0$ and

$$\lambda^i_{\mathbf{z}}(\mathbf{x}) = \lambda^{i-1}_{\mathbf{z}}(\mathbf{x}) + c^i_{\mathbf{z}}\,\mathbf{x}^{(\mathbf{x}^i)}, \quad \text{for } i=1,2,\dots,n, \text{ where } \quad c^i_{\mathbf{z}} = \frac{\lambda^*_{\mathbf{z}}(\mathbf{x}^i) - \lambda^{i-1}_{\mathbf{z}}(\mathbf{x}^i)}{\mathbf{x}^{i(\mathbf{x}^i)}}. \tag{3.2}$$

Note that $\lambda^n_{\mathbf{z}}(\mathbf{x}) = \sum_{i\leqslant n} c^i_{\mathbf{z}}\,\mathbf{x}^{(\mathbf{x}^i)}$, and the term $c^i_{\mathbf{z}}\,\mathbf{x}^{(\mathbf{x}^i)}$ can be associated with the mass-action intensity of a reaction $\mathbf{x}^i \to \mathbf{x}^i + \mathbf{z}$ as long as $c^i_{\mathbf{z}} \geqslant 0$. Hence if $c^i_{\mathbf{z}} \geqslant 0$ for each i, we can find a mass-action system that has the same transition rates as $\lambda^*_{\mathbf{z}}$.

Theorem 3.1 *Let* $\mathbf{X}(t)$ *be a* CTMC *defined on the state space* $\mathbb{Z}^d_{\geqslant 0}$ *with the transition rate* $\lambda^*_{\mathbf{z}} : \mathbb{Z}^d_{\geqslant 0} \to [0,\infty)$ *for each transition vector* $\mathbf{z} \in \mathcal{Z} \subset \mathbb{Z}^d$, *where* $|\mathcal{Z}| < \infty$. *Suppose that the constant* $c^i_{\mathbf{z}}$ *in* (3.2) *is non-negative for each* $\mathbf{z} \in \mathcal{Z}$ *and* $\mathbf{x}^i \in \mathbb{S}_N$, *for* $i=1,2,\dots,n$, *where we use notation* (2.8). *Then for each integer* $N>0$, *there exists a unique mass-action reaction system* $(\mathcal{R},\mathcal{K})$ *such that*

(i) *the order of the reaction system* $(\mathcal{R},\mathcal{K})$ *is less than or equal to* N, *and*

(ii) *for each transition vector* $\mathbf{z} \in \mathcal{Z}$, *if* $\lambda^*_{\mathbf{z}}(\mathbf{x}') > 0$ *for some* $\mathbf{x}' \in \mathbb{S}_N$, *then*

$$\lambda^*_{\mathbf{z}}(\mathbf{x}) = \sum_{\substack{\mathbf{y}\to\mathbf{y}'\in\mathcal{R}\\ \mathbf{y}'-\mathbf{y}=\mathbf{z}}} \lambda_{\mathbf{y}\to\mathbf{y}'}(\mathbf{x}) \quad \textit{for all } \mathbf{x} \in \mathbb{S}_N,$$

where $\lambda_{\mathbf{y}\to\mathbf{y}'}$ *is the reaction intensity of* $\mathbf{y}\to\mathbf{y}' \in \mathcal{R}$.

Proof The uniqueness of $(\mathcal{R},\mathcal{K})$ follows from Lemma 3.1. To prove existence, we denote states in $\mathbb{S}_N$ by (2.8). We fix $\mathbf{z} \in \mathcal{Z}$, and let $\lambda^*_{\mathbf{z}}$ be the associated transition rate function of $\mathbf{X}(t)$. Then, let

$$\begin{aligned} \mathcal{R}^{\mathbf{z}} &= \left\{ \mathbf{x}^i \to \mathbf{x}^i + \mathbf{z} \;\middle|\; \text{where } i \text{ satisfies } c^i_{\mathbf{z}} > 0 \right\}, \text{ and} \\ \mathcal{K}^{\mathbf{z}} &= \left\{ \lambda_{\mathbf{x}^i\to\mathbf{x}^i+\mathbf{z}}(\mathbf{x}) = c^i_{\mathbf{z}}\,\mathbf{x}^{(\mathbf{x}^i)} \;\middle|\; \mathbf{x}^i \to \mathbf{x}^i + \mathbf{z} \in \mathcal{R}^{\mathbf{z}} \right\}. \end{aligned} \tag{3.3}$$

Then, we prove that $\lambda^*_{\mathbf{z}}(\mathbf{x}^k) = \lambda^n_{\mathbf{z}}(\mathbf{x}^k)$ for each $\mathbf{x}^k \in \mathbb{S}_N$, where $\lambda^n_{\mathbf{z}}(\mathbf{x}^k)$ is given by (3.2) and n is given by (2.8). Note that for any $k<j$, there is an i such that $x^k_i \leqslant x^j_i$ so that $\mathbf{x}^{k(\mathbf{x}^j)} = 0$. Hence,

$$\lambda^n_{\mathbf{z}}(\mathbf{x}^k) = \lambda^k_{\mathbf{z}}(\mathbf{x}^k) + \sum_{j=k+1}^{n} c^j_{\mathbf{z}}\,\mathbf{x}^{k(\mathbf{x}^j)} = \lambda^k_{\mathbf{z}}(\mathbf{x}^k). \tag{3.4}$$

Therefore, for each k

$$\lambda^*_{\mathbf{z}}(\mathbf{x}^k) = c^k_{\mathbf{z}}\,\mathbf{x}^{k(\mathbf{x}^k)} + \lambda^{k-1}_{\mathbf{z}}(\mathbf{x}^k) = \lambda^k_{\mathbf{z}}(\mathbf{x}^k) = \lambda^n_{\mathbf{z}}(\mathbf{x}^k),$$

where the last equality follows by (3.4). Repeating construction (3.3) for each transition vector $\mathbf{z} \in \mathcal{Z}$, we put

$$\mathcal{R} = \bigcup_{\mathbf{z}\in\mathcal{Z}} \mathcal{R}^{\mathbf{z}}, \qquad \text{and} \qquad \mathcal{K} = \bigcup_{\mathbf{z}\in\mathcal{Z}} \mathcal{K}^{\mathbf{z}},$$

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

By the construction, for each transition vector $\mathbf{z} \in \mathcal{Z}$, we have

$$\lambda^*_{\mathbf{z}}(\mathbf{x}) = \lambda^n_{\mathbf{z}}(\mathbf{x}) = \sum_{\substack{\mathbf{y}\to\mathbf{y}'\in\mathcal{R} \\ \mathbf{y}\to\mathbf{y}'=\mathbf{z}}} \lambda_{\mathbf{y}\to\mathbf{y}'}(\mathbf{x}),$$

for each $\mathbf{x} \in \mathbb{S}_N$, where $\lambda_{\mathbf{y}\to\mathbf{y}'}$ is the intensity of a reaction $\mathbf{y} \to \mathbf{y}'$ in $(\mathcal{R}, \mathcal{K})$. The order of $(\mathcal{R}, \mathcal{K})$ is less than or equal to N since the order of each reaction in $\mathcal{R}$ is less than or equal to N. □

Remark 3.1 *The advantage of Theorem 3.1 is that we do not require any algebraic structure on $\mathbb{S}_N$. Since the mass-action intensity of a reaction is a polynomial, transition rates on an arbitrary set A can be used to infer the underlying reaction network and parameters using a canonical polynomial fitting approach. To do that, however, certain algebraic structure on A is required. More details about network inference with polynomial fitting are provided in Section 4.*

Remark 3.2 *If the transition rates $\lambda^*_{\mathbf{z}}$ of a given CTMC $\mathbf{X}(t)$ are given by an order N mass-action system, then $c^i_{\mathbf{z}} \geqslant 0$ for each transition vector $\mathbf{z}$ and $i = 1, 2, \ldots, n$, and we can uncover the underlying reaction network uniquely by the algorithm illustrated in Figure 2.*

3.2 Identifiability of CTMCs

For a CTMC associated with a given reaction system, one of the main questions is identifiability of the underlying reaction system using the information on the CTMC. We formalise this idea more rigorously.

Definition 3.2 For a CTMC $\mathbf{X}(t)$ with the state space $\mathbb{S}$, the CTMC $\mathbf{X}(t)$ is *identifiable* if there is a unique reaction system $(\mathcal{R}, \mathcal{K})$ such that

1. each $\mathbf{y} \to \mathbf{y}' \in \mathcal{R}$ is charged in at least one state $\mathbf{x} \in \mathbb{S}$,
2. the state space of the CTMC associated with $(\mathcal{R}, \mathcal{K})$ contains $\mathbb{S}$, and
3. the associated mass-action CTMC with $(\mathcal{R}, \mathcal{K})$ admits the same transition rates on $\mathbb{S}$ as $\mathbf{X}(t)$ admits.

Otherwise, $\mathbf{X}(t)$ is not identifiable with a reaction system.

For a CTMC $\mathbf{X}(t)$ associated with an order N reaction system, the uniqueness of Theorem 3.1 implies that $\mathbf{X}(t)$ is identifiable as long as enough information on the transition rates of $\mathbf{X}(t)$ is ensured. We begin with a lemma for identifiability of reaction systems.

Lemma 3.2 *Let $\mathbf{X}_1(t)$ and $\mathbf{X}_2(t)$ be two d-dimensional* CTMC*s associated with mass-action systems $(\mathcal{R}_1, \mathcal{K}_1)$ and $(\mathcal{R}_2, \mathcal{K}_2)$ of order N_1 and N_2, respectively. Suppose that $N_1 > N_2$. Suppose further that $\mathbf{X}_1(t)$ and $\mathbf{X}_2(t)$ have the same transition rates at each state $\mathbf{x} \in \mathbb{S}_{N_2}$. Then, $(\mathcal{R}_2, \mathcal{K}_2) \subset (\mathcal{R}_1, \mathcal{K}_1)$.*

Proof We apply Theorem 3.1 to the transition rates of $\mathbf{X}_1(t)$ on $\mathbb{S}_{N_2}$ to identify a unique order N' reaction system $(\mathcal{R}', \mathcal{K}')$ such that $N' \leqslant N_2$ and the associated CTMC under mass-action kinetics

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

has the same transition rates on $\mathbb{S}_{N_2}$ as the transition rates of $\mathbf{X}_1(t)$. Then by the construction in the proof of Theorem 3.1, reaction $\mathbf{y} \to \mathbf{y}' \in \mathcal{R}'$ if and only if $\mathbf{y} \to \mathbf{y}' \in \mathcal{R}_1$ is of order K for some $K \leqslant N_2$. That is, $\mathcal{R}'$ only contains those reactions in $\mathcal{R}_1$ whose order is less than or equal to N_2. Furthermore, the reaction intensity of each reaction $\mathbf{y} \to \mathbf{y}' \in \mathcal{R}'$ is equal to the reaction intensity of $\mathbf{y} \to \mathbf{y}' \in \mathcal{R}_1$. Therefore, $(\mathcal{R}', \mathcal{K}') \subset (\mathcal{R}_1, \mathcal{K}_1)$. Note also that since $\mathbf{X}_1(t)$ and $\mathbf{X}_2(t)$ have the same transition rates on $\mathbb{S}_{N_2}$, by uniqueness shown in Lemma 3.1, we have $(\mathcal{R}', \mathcal{K}') = (\mathcal{R}_2, \mathcal{K}_2)$ because the order of both reaction systems are less than or equal to N_2, and the associated CTMCs have the same transition rates on $\mathbb{S}_{N_2}$. □

Lemma 3.2 ensures that if two reaction systems have the same transition rates, then the one with lower order is a subsystem of the other. Using this fact, we obtain identifiability of a reaction system.

Theorem 3.3 *Let* $\mathbf{X}(t)$ *be a* CTMC *associated with an order* N *reaction system* $(\mathcal{R}, \mathcal{K})$ *with the state space* $\mathbb{S}$. *If* $\mathbb{S}_N \subseteq \mathbb{S}$, *then* $\mathbf{X}(t)$ *is identifiable.*

Proof First of all, suppose that there exists a reaction system $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ of order $\overline{N}$ where $\overline{N} < N$ such that the associated mass-action system satisfies the conditions (1)–(3) in Definition 3.2. Then Lemma 3.2 implies that $(\overline{\mathcal{R}}, \overline{\mathcal{K}}) \subset (\mathcal{R}, \mathcal{K})$. Since $\overline{N} < N$, there exists a reaction $\tilde{\mathbf{y}} \to \tilde{\mathbf{y}}'$ of order N that belongs to $\mathcal{R} \setminus \overline{\mathcal{R}}$. Let $\mathbf{z} = \tilde{\mathbf{y}}' - \tilde{\mathbf{y}}$. Then at state $\tilde{\mathbf{y}} \in \mathbb{S}_N$,

$$\sum_{\substack{\mathbf{y} \to \mathbf{y}' \in \mathcal{R} \\ \mathbf{y}' - \mathbf{y} = \mathbf{z}}} \lambda_{\mathbf{y} \to \mathbf{y}'}(\tilde{\mathbf{y}}) - \sum_{\substack{\mathbf{y} \to \mathbf{y}' \in \overline{\mathcal{R}} \\ \mathbf{y}' - \mathbf{y} = \mathbf{z}}} \overline{\lambda}_{\mathbf{y} \to \mathbf{y}'}(\tilde{\mathbf{y}}) \geqslant \lambda_{\tilde{\mathbf{y}} \to \tilde{\mathbf{y}}}(\tilde{\mathbf{y}}) > 0,$$

where $\lambda_{\mathbf{y} \to \mathbf{y}}$ and $\overline{\lambda}_{\mathbf{y} \to \mathbf{y}}$ are the reaction intensity associated with a reaction $\mathbf{y} \to \mathbf{y}'$ of $(\mathcal{R}, \mathcal{K})$ and $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$, respectively. Therefore it contradicts to the fact that $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ has the same transition rates on each state $\mathbf{x} \in \mathbb{S}_N$ as $\mathbf{X}(t)$. For the same reason, there does not exists a reaction system, which has higher order than N, satisfies the conditions (1)–(3) in Definition 3.2.

In conclusion, the only reaction network satisfying the conditions (1)–(3) in Definition 3.2 is $(\mathcal{R}, \mathcal{K})$ because uniqueness among reaction systems of order N is guaranteed by Lemma 3.1 and $(\mathcal{R}, \mathcal{K})$ satisfies the condition (1)–(3) in Definition 3.2. □

In practical situations, it is often the case that an associated mass-action CTMC $\mathbf{X}(t)$ is given, but the underlying reaction system $(\mathcal{R}, \mathcal{K})$ is unknown. However, it is reasonable to assume that the order of $(\mathcal{R}, \mathcal{K})$ does not exceed a relatively small number $\overline{N}$ for general biochemical system (for example, many biochemical systems are at most bimolecular, hence we could set $\overline{N} = 2$). Under this assumption, $\mathbf{X}(t)$ is identifiable as long as enough information about the transition rates is given. The case of the unknown order is a consequence of Theorems 3.1 and 3.3 and is formulated as the following corollary.

Corollary 3.1 *Let a CTMC* $\mathbf{X}(t)$ *be a mass-action stochastic system associated with an unknown order* N *reaction system* $(\mathcal{R}, \mathcal{K})$ *with the state space* $\mathbb{S}$. *Suppose that* $N \leqslant \overline{N}$ *for some positive integer* $\overline{N}$. *Suppose further that* $\mathbb{S}_{\overline{N}} \subseteq \mathbb{S}$. *Then* $\mathbf{X}(t)$ *is identifiable. Moreover, using the transition rates of* $\mathbf{X}(t)$, *the true network* $(\mathcal{R}, \mathcal{K})$ *can be explicitly inferred.*

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

3.3 Identifiability of reaction systems with conservation laws

If the transition rate of a Markov process is given over a proper subset $\mathbb{A} \subset \mathbb{S}_N$ for given $N > 0$, then two distinct reaction systems of order N may be constructed having the same transition rates over $\mathbb{A}$. Since $\mathbb{A}$ is the proper subset of $\mathbb{S}_N$, we have $\mathbb{A} \subset \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^m\} \subset \mathbb{S}_N$ where $m < n$. Given the transition rates on $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^m\}$ and considering $\mathbf{x}^\ell \succ \mathbf{x}^m$ such that $\mathbf{x}^\ell \in \mathbb{S}_N$, the mass-action reaction intensity associated with a reaction $\mathbf{x}_\ell \to \mathbf{x}_\ell + \mathbf{z}$ is zero at each state in $\{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^m\}$. Hence by adding or removing $\mathbf{x}_\ell \to \mathbf{x}_\ell + \eta$, we obtain different reaction systems that have the same transition rates on $\mathbb{A}$.

Next, we consider other situations where the underlying reaction system of a CTMC is not uniquely determined. Suppose a given CTMC $\mathbf{X}(t)$ associated with a stochastic reaction network of order N admits a conservation law, i.e. there exists $\mathbf{v} \in \mathbb{Z}^d_{\geqslant 0}$ such that $\mathbf{v} \cdot \mathbf{X}(t) = \mathbf{v} \cdot \mathbf{X}(0)$ for any time $t \geqslant 0$. In this section, we simplify our discussion by considering that the vector $\mathbf{v}$ has all non-zero components, that is, $\mathbf{v} \in \mathbb{Z}^d_{>0}$. Then the state space of $\mathbf{X}(t)$ is confined to a finite hyperplane $\mathbb{S}_{\mathbf{v},N}$ of $\mathbb{Z}^d_{>0}$. In this case, one of the main questions is whether the information about the transition rates over a single hyperplane is sufficient to uniquely infer the underlying reaction system.

In this section, we show how to construct a reaction network of order N with given transition rates over a single hyperplane $\mathbb{S}_{\mathbf{v},N}$, see the definition (2.6). We further show that when a given reaction system $(\mathcal{R}, \mathcal{K})$ is of order N, then the underlying reaction network is not uniquely identified with given transition rates on a single hyperplane $\mathbb{S}_{\mathbf{v},N'}$ such that $N < N'$.

Theorem 3.4 *Let $\mathbf{z} \in \mathbb{Z}^d$, $\mathbf{v} \in \mathbb{Z}^d_{>0}$ and $N > 0$. Let $\lambda(\mathbf{x})$ be a given non-negative function defined on $\mathbb{S}_{\mathbf{v},N}$ such that $\lambda(\mathbf{x}) > 0$ for at least one $\mathbf{x} \in \mathbb{S}_{\mathbf{v},N}$. Then there exists a mass-action reaction system $(\mathcal{R}^{\mathbf{z}}, \mathcal{K}^{\mathbf{z}})$ of $\mathbf{v}$-order N such that the transition rates at each $\mathbf{x} \in \mathbb{S}_{\mathbf{v},N}$ are equal to $\lambda(\mathbf{x})$. That is*

$$\sum_{\mathbf{y} \to \mathbf{y}' \in \mathcal{R}^{\mathbf{z}}} \lambda_{\mathbf{y} \to \mathbf{y}'}(\mathbf{x}) = \lambda(\mathbf{x}) \quad \textit{for each} \quad \mathbf{x} \in \mathbb{S}_{\mathbf{v},N}. \tag{3.5}$$

Proof The key idea of the proof is that (under the mass-action kinetics) every reaction of $\mathbf{v}$-order N is charged at a single state $\mathbf{x} \in \mathbb{S}_{\mathbf{v},N}$ and turned off elsewhere in $\mathbb{S}_{\mathbf{v},N}$. So we will collect all reactions $\mathbf{x} \to \mathbf{x} + \mathbf{z}$ for each $\mathbf{x} \in \mathbb{S}_{\mathbf{v},N}$ as long as $\lambda(\mathbf{x}) > 0$. We define

$$\mathcal{R}^{\mathbf{z}} = \left\{ \mathbf{x} \to \mathbf{x} + \mathbf{z} \;\middle|\; \lambda(\mathbf{x}) > 0,\ \mathbf{x} \in \mathbb{S}_{\mathbf{v},N} \right\}, \text{ and}$$

$$\mathcal{K}^{\mathbf{z}} = \left\{ \lambda_{\mathbf{x} \to \mathbf{x}+\mathbf{z}}(\mathbf{w}) = \frac{\lambda(\mathbf{x})}{\mathbf{x}^{(\mathbf{x})}}\, \mathbf{w}^{(\mathbf{x})} \text{ for any } \mathbf{w} \in \mathbb{Z}^d_{\geqslant 0} \;\middle|\; \lambda(\mathbf{x}) > 0,\ \mathbf{x} \in \mathbb{S}_{\mathbf{v},N} \right\}.$$

Since $\mathbf{v} \in \mathbb{Z}^d_{>0}$, for any two distinct states $\mathbf{x}$ and $\mathbf{x}'$ in $\mathbb{S}_{\mathbf{v},N}$, there is an index k such that $x_k > x'_k$. Therefore, the reaction $\mathbf{x} \to \mathbf{x} + \mathbf{z} \in \mathcal{R}^{\mathbf{x}}$ is turned off at $\mathbf{x}'$ if and only if $\mathbf{x} \neq \mathbf{x}'$. This implies that for any $\mathbf{x} \in \mathbb{S}_{\mathbf{v},N}$ such that $\lambda(\mathbf{x}) > 0$:

$$\lambda(\mathbf{x}) = \frac{\lambda(\mathbf{x})}{\mathbf{x}^{(\mathbf{x})}}\, \mathbf{x}^{(\mathbf{x})} = \lambda_{\mathbf{x} \to \mathbf{x}+\mathbf{z}}(\mathbf{x}) = \sum_{\mathbf{y} \to \mathbf{y}' \in \mathcal{R}^{\mathbf{z}}} \lambda_{\mathbf{y} \to \mathbf{y}'}(\mathbf{x}).$$

Equation (3.5) is also valid for any $\mathbf{x} \in \mathbb{S}_{\mathbf{v},N}$ satisfying $\lambda(\mathbf{x}) = 0$, because we have $(\mathbf{x} \to \mathbf{x} + \mathbf{z}) \notin \mathcal{R}^{\mathbf{z}}$ and each $\mathbf{x}' \to \mathbf{x}' + \mathbf{z} \in \mathcal{R}_{\mathbf{z}}$ is turned off at $\mathbf{x}$. □

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

Theorem 3.4 implies that for a given CTMC defined on a hyperplane $\mathbb{S}_{\mathbf{v},N}$, we can construct a reaction network of $\mathbf{v}$-order N such that the associated mass-action CTMC admits the same transition rates on $\mathbb{S}_{\mathbf{v},N}$. Using this, we prove that a CTMC associated with a conservative reaction system of $\mathbf{v}$-order N is not identifiable if the transition data of the CTMC are only given on $\mathbb{S}_{\mathbf{v},N'}$ for some $N' > N$.

Theorem 3.5 *Let $(\mathcal{R}, \mathcal{K})$ be a mass-action reaction system that admits a conservation law with $\mathbf{v} \in \mathbb{Z}^d_{>0}$ such that $\mathbf{v} \cdot (\mathbf{y}' - \mathbf{y}) = 0$ for each $\mathbf{y} \to \mathbf{y}' \in \mathcal{R}$. Suppose that the $\mathbf{v}$-order of $(\mathcal{R}, \mathcal{K})$ is N. Let $\mathbf{X}(t)$ be the CTMC associated with $(\mathcal{R}, \mathcal{K})$ such that $\mathbf{v} \cdot \mathbf{X}(0) = N'$ and $N' > N$. Then, the CTMC $\mathbf{X}(t)$ is not identifiable.*

Proof Because of the conservation law, the state space of $\mathbf{X}(t)$ is $\mathbb{S}_{\mathbf{v},N'}$, defined by (2.6), because $\mathbf{v} \cdot \mathbf{X}(t) = \mathbf{v} \cdot \mathbf{X}(0)$ for any time $t \geqslant 0$. For a fixed transition vector $\mathbf{z}$ in the set of transition vectors $\mathcal{Z}$ of $\mathbf{X}(t)$, we denote by $\lambda_{\mathbf{z}}(\mathbf{x})$ the transition rate of $\mathbf{X}(t)$ at $\mathbf{x} \in \mathbb{S}_{\mathbf{v},N'}$. Then for each $\mathbf{x} \in \mathbb{S}_{\mathbf{v},N'}$, we have

$$\lambda_{\mathbf{z}}(\mathbf{x}) = \sum_{\substack{\mathbf{y} \to \mathbf{y}' \in \mathcal{R} \\ \mathbf{y}' - \mathbf{y} = \mathbf{z}}} \lambda_{\mathbf{y} \to \mathbf{y}'}(\mathbf{x}),$$

where $\lambda_{\mathbf{y} \to \mathbf{y}'}$ is the intensity of reaction $(\mathbf{y} \to \mathbf{y}') \in \mathcal{R}$. Since $\lambda_{\mathbf{z}}$ is the transition rate of a mass-action reaction system of $\mathbf{v}$-order equal to N, there exists $\mathbf{x}^* \in \mathbb{S}_{\mathbf{v},N}$ such that $(\mathbf{x}^* \to \mathbf{x}^* + \mathbf{z}) \in \mathcal{R}$. Therefore, for $\mathbf{x}' = \mathbf{x}^* + ((N' - N)/v_1, 0, 0, \ldots, 0) \in \mathbb{S}_{\mathbf{v},N'}$, we have $\lambda_{\mathbf{z}}(\mathbf{x}') \geqslant \lambda_{\mathbf{x}^* \to \mathbf{x}^* + \mathbf{z}}(\mathbf{x}') > 0$. This means that there exist at least one $\mathbf{x}' \in \mathbb{S}_{\mathbf{v},N'}$ such that $\lambda_{\mathbf{z}}(\mathbf{x}') > 0$. Hence using Theorem 3.4 with $\lambda_{\mathbf{z}}$ and $\mathbf{z}$, we can construct a reaction system $(\mathcal{R}^{\mathbf{z}}, \mathcal{K}^{\mathbf{z}})$ of $\mathbf{v}$-order N'. Then, we have

$$\lambda_{\mathbf{z}}(\mathbf{x}) = \sum_{\mathbf{y} \to \mathbf{y}' \in \mathcal{R}^{\mathbf{z}}} \overline{\lambda}_{\mathbf{y} \to \mathbf{y}'}(\mathbf{x}),$$

where $\overline{\lambda}_{\mathbf{y} \to \mathbf{y}'}$ is the intensity of reaction $\mathbf{y} \to \mathbf{y}'$ in $(\mathcal{R}^{\mathbf{z}}, \mathcal{K}^{\mathbf{z}})$. Applying Theorem 3.4 in the same way for all transition vectors $\mathbf{z} \in \mathcal{Z}$, we define

$$\overline{\mathcal{R}} = \bigcup_{\mathbf{z} \in \mathcal{Z}} \mathcal{R}^{\mathbf{z}}, \quad \text{and} \quad \overline{\mathcal{K}} = \bigcup_{\mathbf{z} \in \mathcal{Z}} \mathcal{K}^{\mathbf{z}}.$$

Then, we have

$$\lambda_{\mathbf{z}}(\mathbf{x}) = \sum_{\mathbf{y} \to \mathbf{y}' \in \mathcal{R}^{\mathbf{z}}} \overline{\lambda}_{\mathbf{y} \to \mathbf{y}'}(\mathbf{x}) = \sum_{\substack{\mathbf{y} \to \mathbf{y}' \in \overline{\mathcal{R}} \\ \mathbf{y}' - \mathbf{y} = \mathbf{z}}} \overline{\lambda}_{\mathbf{y} \to \mathbf{y}'}(\mathbf{x}),$$

for each $\mathbf{x} \in \mathbb{S}_{\mathbf{v},N'}$ and for each transition vector $\mathbf{z}$ of $\mathbf{X}(t)$. This implies that the CTMC associated with $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ has the same transition rates on $\mathbb{S}_{\mathbf{v},N'}$, which is the state space of $\mathbf{X}(t)$. Since $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ is of $\mathbf{v}$-order N', the two reaction systems $(\mathcal{R}, \mathcal{K})$ and $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ are distinct. Hence, $\mathbf{X}(t)$ is not identifiable. □

We illustrate Theorem 3.5 using the following example.

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

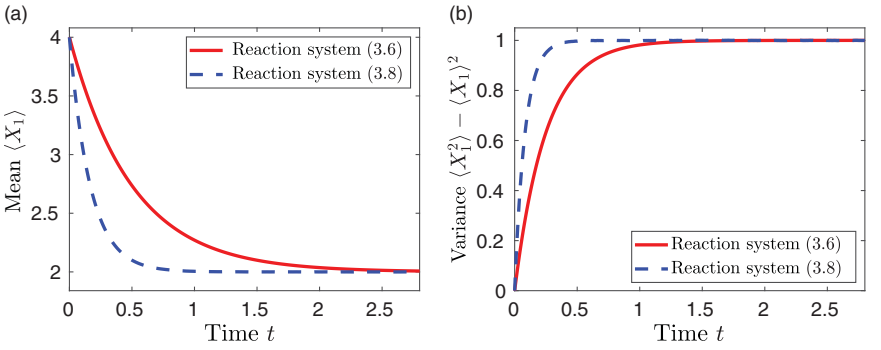

FIGURE 3. (a) The mean number of molecules of the chemical species X_1 for the chemical system (3.6) (red solid line) compared with the result for the chemical system (3.8) (blue dashed line). (b) Time evolution of the variance of the number of molecules of the chemical species X_1. We use the same initial condition $\mathbf{X}(0) = (N, 0)$, where $N = 4$, for both systems. The results for the reaction network (3.6) are calculated by equation (3.9), while the results for the reaction network (3.8) are estimated as averages over 10^7 realisations of the Gillespie SSA.

Example 3.2 Let $\mathbf{X}(t)$ be the CTMC associated with the mass-action reaction system:

$$X_1 \underset{1}{\overset{1}{\rightleftarrows}} X_2 \,. \tag{3.6}$$

Note that this system admits a conservation law such that $\mathbf{v} \cdot \mathbf{X} = X_1(0) + X_2(0)$ where $\mathbf{v} = (1, 1) \in \mathbb{Z}^2_{>0}$. With $\mathbf{X}(0) = (2, 0)$, the transition rates of $\mathbf{X}(t)$ at its state space $\mathbb{S}_{\mathbf{v},2}$ are

$$\lambda_{(-1,1)}(2, 0) = \lambda_{(1,-1)}(0, 2) = 2, \qquad \text{and} \qquad \lambda_{(-1,1)}(1, 1) = \lambda_{(1,-1)}(1, 1) = 1. \tag{3.7}$$

Note that the $\mathbf{v}$-order of the reaction system (3.6) is 1. Using Theorem 3.4, we construct the following reaction system of $\mathbf{v}$-order 2 with the same transition rates (3.7) on $\mathbb{S}_{\mathbf{v},2}$:

$$2\,X_1 \underset{1}{\overset{1}{\rightleftarrows}} X_1 + X_2 \underset{1}{\overset{1}{\rightleftarrows}} 2\,X_2 \,. \tag{3.8}$$

The CTMC associated with the reaction system (3.8) admits the same transition rates on $\mathbb{S}_{\mathbf{v},2}$ as $\mathbf{X}(t)$ does. However, these two reaction systems exhibit different dynamical behaviours if we consider them on a different hyperplane $\mathbb{S}_{\mathbf{v},N}$ as we show in Figure 3 for $N = 4$. Considering the initial condition $\mathbf{X}(0) = (N, 0)$, the mean and variance of the number of molecules of X_1 of the reaction system (3.6) are given by [12]:

$$\langle X_1 \rangle = \frac{N}{2}\left(1 + \exp[-2t]\right), \qquad \langle X_1^2 \rangle - \langle X_1 \rangle^2 = \frac{N}{4}\left(1 - \exp[-4t]\right). \tag{3.9}$$

Using $N = 4$, we plot (3.9) as the red solid lines in Figure 3, where we compare them with the results calculated for the reaction system (3.8) by averaging over 10^7 realisations of the Gillespie stochastic simulation algorithm (SSA).

Remark 3.3 *If we consider the same hyperplane, $\mathbb{S}_{\mathbf{v},2}$, as in Example 3.2, we can also construct an identifiable network if the conditions of Theorem 3.5 are not satisfied. For example, replacing the reaction system (3.6) with the reaction system:*

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

$$2X_1 \underset{\kappa_2}{\overset{\kappa_1}{\rightleftarrows}} 2X_2 .$$

and letting $X(0) = (2, 0)$*, the state space is* $\{(2, 0), (0, 2)\} \subset \mathbb{S}_{\mathbf{v},2}$ *with* $\mathbf{v} = (1, 1)$*. Then the CTMC* $\mathbf{X}(t)$ *is the only reaction network of the* $\mathbf{v}$*-order* 2 *with the same transition rates on* $\mathbb{S}_{\mathbf{v},2}$*, that is, the CTMC* $\mathbf{X}(t)$ *is identifiable.*

4 Reaction networks for Markov processes with polynomial rates

In Section 3.1, we showed that if the transition rates of a CTMC are given at each state in $\mathbb{S}_N$ for some N, then we can uniquely identify an order N stochastic reaction system that has the same transition rates on $\mathbb{S}_N$. In this section, we explore the case where the transition rates of a CTMC are known on arbitrary states, which are not necessarily belonging to $\mathbb{S}_N$. For a d-dimensional CTMC, we will use the transition rates at (compare with (2.5) and (2.8)):

$$n = |\mathbb{S}_N| = \binom{N+d}{d}$$

different states to uniquely identify an order N stochastic reaction system that has the same transition rates at the given states.

Lemma 4.1 *Let* $\mathbf{X}(t)$ *be a CTMC defined on* $\mathbb{Z}^d_{\geqslant 0}$ *with the finite set of transition vectors* $\mathcal{Z}$*. Suppose for each transition vector* $\mathbf{z} \in \mathcal{Z}$*, the transition rates of* $\mathbf{X}(t)$ *are given in finite set* $A_{\mathbf{z}} \subset \mathbb{Z}^d$*. Then there exists a CTMC* $\overline{\mathbf{X}}(t)$ *with polynomial transition rates such that for each* $\mathbf{z} \in \mathcal{Z}$*:*

$$\overline{\lambda}_{\mathbf{z}}(\mathbf{x}) = \lambda_{\mathbf{z}}(\mathbf{x}) \quad \textit{for each } \mathbf{x} \in A_{\mathbf{z}}, \tag{4.1}$$

where $\lambda_{\mathbf{z}}$ *is the given transition rate of* $\mathbf{X}(t)$*, and* $\overline{\lambda}_{\mathbf{z}}$ *is a polynomial transition rate of* $\overline{\mathbf{X}}(t)$*. Moreover, assume that we have* $|A_{\mathbf{z}}| = n = |\mathbb{S}_N|$ *for some positive integer* N*, and denote the elements of* $A_{\mathbf{z}}$ *as* $\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^n$ *and elements of* $\mathbb{S}_N$ *by* (2.8)*. Define a matrix* $M \in \mathbb{Z}^{n\times n}$ *with entries:*

$$M_{ij} = \mathbf{a}^{i(\mathbf{x}^j)} \qquad \textit{for} \quad i = 1, 2, \ldots, n,\ j = 1, 2, \ldots, n.$$

If the matrix M *is invertible, then* $\lambda_{\mathbf{z}}$ *is a unique degree* N *polynomial.*

Proof We can find a polynomial $\overline{\lambda}_{\mathbf{z}}$ such that (4.1) is satisfied because the set $A_{\mathbf{z}}$ is finite for each transition vector $\mathbf{z} \in \mathcal{Z}$ and $|\mathcal{Z}| < \infty$. Suppose that matrix M is invertible. Then, for each $\mathbf{z} \in \mathcal{Z}$, we define $\mathbf{c} \in \mathbb{R}^n$ by

$$\mathbf{c} = M^{-1}\mathbf{b}, \quad \text{where} \quad \mathbf{b} = \left(\lambda_{\mathbf{z}}(\mathbf{a}^1), \lambda_{\mathbf{z}}(\mathbf{a}^2), \ldots, \lambda_{\mathbf{z}}(\mathbf{a}^n)\right)^{\top}. \tag{4.2}$$

Then the degree N polynomial $\overline{\lambda}_{\mathbf{z}}(\mathbf{x})$ is uniquely written as:

$$\overline{\lambda}_{\mathbf{z}}(\mathbf{x}) = \sum_{j=1}^{n} c_j\, \mathbf{x}^{(\mathbf{x}^j)}.$$

□

For a given CTMC, our final goal of this section is to identify a unique mass-action stochastic system that has the same transition rates as the given CTMC admits. By applying Lemma 4.1, we can construct a CMTC whose transition rates are polynomials and have the same values

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

as the given transition rates. However, not every CTMC with polynomial rates is associated with a mass-action reaction network. Negative coefficients cause problems as it is the case of polynomial ODE models which cannot be written as chemical reaction systems [26]. In the case of CTMC, the situation is even more restrictive. To formulate the theorem characterising which CTMC with polynomial transition rates can be identified as a mass-action reaction system, we define $D_i(\lambda) = \min\{y_i^1, y_i^2, \dots, y_i^n\}$ for a polynomial $\lambda(\mathbf{x})$ such that

$$\lambda(\mathbf{x}) = \sum_{j=1}^{n} c_j \, \mathbf{x}^{(\mathbf{y}^j)} \quad \text{for some elements } \mathbf{y}^j \in \mathbb{Z}^d_{\geqslant 0} \text{ and for some constants } c_j > 0. \tag{4.3}$$

Example 4.1 Let us consider polynomial:

$$\lambda(x_1, x_2) = (x_1 + 1)x_2(x_2 - 1) = x_2(x_2 - 1) + x_1x_2(x_2 - 1).$$

Then, we can write $\lambda(\mathbf{x}) = \sum_{j=1}^{2} c_j \, \mathbf{x}^{(\mathbf{y}^j)}$, where $\mathbf{y}^1 = (0, 2)$, $\mathbf{y}^2 = (1, 2)$ and $c_j = 1$ for $j = 1, 2$. Hence, $D_1(\lambda) = \min\{y_1^1, y_1^2\} = 0$ and $D_2(\lambda) = \min\{y_2^1, y_2^2\} = 2$.

Theorem 4.1 *Let $\mathbf{X}(t)$ be a CTMC defined on $\mathbb{Z}^d_{\geqslant 0}$ with the set of transition vectors $\mathcal{Z}$. Suppose that each transition rate $\lambda_{\mathbf{z}}$ of $\mathbf{X}(t)$ associated with $\mathbf{z} \in \mathcal{Z}$ is a polynomial of degree N such that*

$$\lambda_{\mathbf{z}}(\mathbf{x}) = \sum_{j=1}^{n} c_j \, \mathbf{x}^{(\mathbf{x}^j)} \qquad \textit{for some constants} \quad c_j \geqslant 0, \tag{4.4}$$

where $n = |\mathbb{S}_N|$ and elements of $\mathbb{S}_N$ are denoted by (2.8). Suppose further that

$$|z_i| \leqslant D_i(\lambda_{\mathbf{z}}) \quad \textit{if } z_i < 0. \tag{4.5}$$

Then there exists a unique mass-action reaction system such that the associated mass-action stochastic model is equal to the CTMC $\mathbf{X}(t)$.

Proof Let $\mathbf{z} \in \mathcal{Z}$ be fixed. Then, the associated transition rate $\lambda_{\mathbf{z}}$ is given by (4.4). Let $\mathbf{x}^j \in \mathbb{S}_N$ be such that $c_j > 0$ on the right-hand side of equation (4.4). We want to find a chemical reaction with the associated mass-action intensity equal to $c_j \, \mathbf{x}^{(\mathbf{x}^j)}$. Such a reaction is given by $\mathbf{x}^j \to \mathbf{x}^j + \mathbf{z}$ provided that every component of $\mathbf{x}^j + \mathbf{z}$ is non-negative. However, this follows from our assumption (4.5), which implies that

$$x_i^j + z_i \geqslant D_i(\lambda_{\mathbf{z}}) + z_i \geqslant 0, \qquad \text{for all} \quad i = 1, 2, \dots, d, \quad j = 1, 2, \dots, n.$$

Therefore, we define

$$\mathcal{R}^{\mathbf{z}} = \left\{ \mathbf{x}^j \to \mathbf{x}^j + \mathbf{z} \,\middle|\, c_j > 0 \right\} \qquad \text{and} \qquad \mathcal{K}^{\mathbf{z}} = \left\{ \lambda_{\mathbf{x}^j \to \mathbf{x}^j + \mathbf{z}}(\mathbf{x}) = c_j \, \mathbf{x}^{(\mathbf{x}^j)} \,\middle|\, c_j > 0 \right\}.$$

Then,

$$\lambda_{\mathbf{z}}(\mathbf{x}) = \sum_{\mathbf{x}^j \to \mathbf{x}^j + \mathbf{z} \in \mathcal{R}_{\mathbf{z}}} \lambda_{\mathbf{x}^j \to \mathbf{x}^j + \mathbf{z}}(\mathbf{x}).$$

Considering $\mathcal{R}^{\mathbf{z}}$ and $\mathcal{K}^{\mathbf{z}}$ obtained for each $\mathbf{z} \in \mathcal{Z}$, we define $\mathcal{R} = \bigcup_{\mathbf{z} \in \mathcal{Z}} \mathcal{R}^{\mathbf{z}}$ and $\mathcal{K} = \bigcup_{\mathbf{z} \in \mathcal{Z}} \mathcal{K}^{\mathbf{z}}$. The associated CTMC for $(\mathcal{R}, \mathcal{K})$ has the same transition rates as X has. Uniqueness follows since the decomposition (4.4) is unique. □

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

Suppose a given CTMC satisfies the conditions in Lemma 4.1 and that the transition rates of the CTMC satisfy the conditions (4.4) and (4.5) in Theorem 4.1. Then we can infer a reaction network whose associated CTMC has the same transition vectors and the same transition rates at each state in $A_{\mathbf{z}}$ for each transition vector $\mathbf{z}$. We demonstrate this using the following example.

Example 4.2 Let $\mathbf{X}(t)$ be a CTMC defined on $\mathbb{Z}^2_{\geqslant 0}$. Suppose that it is known that $\mathbf{X}(t)$ admits three transition vectors $\mathbf{z}^1 = (1, 0)$, $\mathbf{z}^2 = (-1, 1)$ and $\mathbf{z}^3 = (0, -1)$. We are also given information on the transition rates of $\mathbf{X}(t)$ as:

$$\begin{aligned} &\lambda_{\mathbf{z}^1}(10, 10) = 1, \\ &\lambda_{\mathbf{z}^2}(10, 10) = 20, \qquad \lambda_{\mathbf{z}^2}(9, 11) = 18, \qquad \lambda_{\mathbf{z}^2}(9, 10) = 18, \\ &\lambda_{\mathbf{z}^3}(8, 11) = 33, \qquad \lambda_{\mathbf{z}^3}(8, 10) = 30, \qquad \lambda_{\mathbf{z}^3}(7, 11) = 33. \end{aligned} \tag{4.6}$$

Using Lemma 4.1, we first find a CTMC $\overline{\mathbf{X}}(t)$ with polynomial transition rates. Using the notation of Lemma 4.1 for the first transition vector $\mathbf{z}^1$, we have $A_{\mathbf{z}^1} = \{(10, 10)\}$ such that $n = 1 = |\mathbb{S}_0|$, matrix M is scalar $M = 1$ and 'vector' $\mathbf{b}$ is a scalar as well, $\mathbf{b} = \lambda_{\mathbf{z}^1}(10, 10) = 1$. Thus, the polynomial transition rate $\lambda_{\mathbf{z}^1}$ is a constant given by (4.2) as $\lambda_{\mathbf{z}^1} = M^{-1}\mathbf{b} = 1$. Considering transition vectors $\mathbf{z}^2$ and $\mathbf{z}^3$, we have

$$A_{\mathbf{z}^2} = \{(10, 10),\ (9, 11),\ (9, 10)\}, \quad \text{and} \quad A_{\mathbf{z}^3} = \{(8, 11),\ (8, 10),\ (7, 11)\}.$$

Since $|\mathbb{S}_1| = 3$, we find a linear transition rate $\lambda_{\mathbf{z}^2}$ (resp. $\lambda_{\mathbf{z}^3}$) of $\overline{\mathbf{X}}(t)$ that has the values (4.6) at $A_{\mathbf{z}^2}$ (resp. $A_{\mathbf{z}^3}$). The 3×3 matrix M is given as:

$$M = \begin{bmatrix} 1 & 10 & 10 \\ 1 & 11 & 9 \\ 1 & 10 & 9 \end{bmatrix}, \qquad \text{respectively,} \qquad M = \begin{bmatrix} 1 & 11 & 8 \\ 1 & 10 & 8 \\ 1 & 11 & 7 \end{bmatrix}.$$

Since both matrices are invertible, we can calculate $\mathbf{c}$ by (4.2), where $\mathbf{b} = (20, 18, 18)^\top$, respectively $\mathbf{b} = (33, 30, 33)^\top$. We obtain $\mathbf{c} = M^{-1}\mathbf{b} = (0, 0, 2)^\top$ for the transition vector $\mathbf{z}^2$ and $\mathbf{c} = M^{-1}\mathbf{b} = (0, 3, 0)^\top$ for the transition vector $\mathbf{z}^3$. Therefore, we obtain

$$\lambda_{\mathbf{z}^1} = 1, \qquad \lambda_{\mathbf{z}^2}(\mathbf{x}) = 2x_1, \qquad \lambda_{\mathbf{z}^3}(\mathbf{x}) = 3x_2.$$

Next, we find a reaction network whose associated mass-action dynamics is equal to the CTMC $\overline{\mathbf{X}}(t)$. The conditions (4.4) and (4.5) of Theorem 4.1 are satisfied for all three transition vectors $\mathbf{z}^1$, $\mathbf{z}^2$ and $\mathbf{z}^3$. Thus, the unique reaction system is

$$\emptyset \xrightarrow{1} X_1 \xrightarrow{2} X_2 \xrightarrow{3} \emptyset\,.$$

Example 4.3 Consider the reaction system (3.6) introduced in Example 3.2. Let $\mathbf{z} = (1, -1)$ be one of the two transition vectors of the CTMC $\mathbf{X}(t)$. Given the transition rates (3.7) on $\mathbb{S}_{\mathbf{v},2}$, the first-order reaction $X_2 \to X_1$ is not identified using Theorem 4.1, because matrix M associated with states $\mathbb{S}_{\mathbf{v},2}$ is the singular matrix:

$$M = \begin{pmatrix} 1 & 2 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 2 \end{pmatrix}.$$

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

5 Inference of reaction networks using temporal data

Theorem 3.1 states that we can use transition rates and transition vectors of a mass-action stochastic reaction system to uncover the underlying network structure. However, in applications, we are not given directly the transition rates, but instead only temporal data consisting of states and transition times between them. For example, for an (a priori unknown) underlying network:

$$X_1 \xrightarrow{1} 2X_1, \qquad X_1 + X_2 \xrightarrow{1} 2X_2,$$

we are given transition data of the associated CTMC $\mathbf{X}(t)$ such as:

$$\mathbf{X}(0) = (1,1),\ \mathbf{X}(\tau_1) = (2,1),\ \mathbf{X}(\tau_2) = (1,2), \ldots, \quad \text{and} \quad \tau_1 = 0.2,\ \tau_2 = 1.1, \ldots,$$

where τ_i is the i-th transition time. Thus, to apply results of the previous section, we need to use such time series to estimate the transition vectors $(1,0)$ and $(-1,1)$ and the corresponding transition rates $\lambda_{(1,0)}(\mathbf{x}) = x_1$ and $\lambda_{(-1,1)}(\mathbf{x}) = x_1 x_2$.

Suppose that we are given Q sample trajectories of the CTMC $\mathbf{X}(t)$ consisting of the states of the system $\mathbf{X}^i(\tau_k^i)$, for $i = 1, 2, \ldots, Q$, recorded at times τ_k^i, where $k = 1, 2, \ldots, q(i)$, and $q(i)$ denotes the number of time points in the i-th time series. Assuming that the given time series includes all reaction events, the time of the k-th transition of the CTMC $\mathbf{X}^i(t)$ is equal to τ_k^i. Then, all possible transition vectors $\mathbf{z}$ of the system can be uncovered (as long as they are present in the recorded time series) by collecting the transitions $\mathbf{X}^i(\tau_{k+1}^i) - \mathbf{X}^i(\tau_k^i)$ for all $k = 1, 2, \ldots, q(i)$ and $i = 1, 2, \ldots, Q$.

Next, we estimate the transition rates at each state $\mathbf{x}$ using the sample trajectories. Let CTMC $\mathbf{X}(t)$ be associated with reaction system $(\mathcal{R}, \mathcal{K})$ and let $\mathcal{Z}$ be the finite set of transition vectors. Then, using the random time representation [20, 2], we have

$$\mathbf{X}(t) = \mathbf{X}(0) + \sum_{\mathbf{z} \in \mathcal{Z}} Y_{\mathbf{z}} \left(\int_0^t \lambda_{\mathbf{z}}(\mathbf{X}(s)) ds \right) \mathbf{z},$$

where $Y_{\mathbf{z}}$ are independent unit Poisson processes. Therefore,

$$\mathbb{E}(\tau_{k+1} \mid \mathbf{X}(\tau_k) = \mathbf{x}) = \frac{1}{\lambda(\mathbf{x})}, \quad \text{where} \quad \lambda(\mathbf{x}) = \sum_{\mathbf{z} \in \mathcal{Z}} \lambda_{\mathbf{z}}(\mathbf{x}),$$

and

$$P\left(\mathbf{X}(\tau_{k+1}) = \mathbf{x} + \mathbf{z} \,\middle|\, \mathbf{X}(\tau_k) = \mathbf{x}\right) = \frac{\lambda_{\mathbf{z}}(\mathbf{x})}{\lambda(\mathbf{x})}. \tag{5.1}$$

To estimate $\lambda_{\mathbf{z}}(\mathbf{x})$ at each state $\mathbf{x}$, we identify the data points when this state was reached by defining $G_{\mathbf{x}} = \{(i,k) \mid \mathbf{X}^i(\tau_k^i) = \mathbf{x} \text{ and } k < q(i)\}$. Then, for each state $\mathbf{x}$ and for each transition vector $\mathbf{z}$, we use

$$\begin{aligned} \lambda_{\mathbf{z}}(\mathbf{x}) = \frac{\lambda_{\mathbf{z}}(\mathbf{x})}{\lambda(\mathbf{x})} \lambda(\mathbf{x}) &= \frac{P\left(\mathbf{X}(\tau_{k+1}) = \mathbf{x} + \mathbf{z} \,\middle|\, \mathbf{X}(\tau_k) = \mathbf{x}\right)}{\mathbb{E}(\tau_{k+1} \mid \mathbf{X}(\tau_k) = \mathbf{x})} \\ &\approx \frac{\sum_{(i,k) \in G_{\mathbf{x}}} \mathbb{1}_{\{\mathbf{X}^i(\tau_{k+1}^i) - \mathbf{X}^i(\tau_k^i) = \mathbf{z}\}}}{\sum_{(i,k) \in G_{\mathbf{x}}} \tau_{k+1}^i}, \end{aligned} \tag{5.2}$$

where we assume that $|G_{\mathbf{x}}|$ is sufficiently large to get a good approximation.

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

Example 5.1 Let $(\mathcal{R}, \mathcal{K})$ be the following one-species mass-action reaction system:

$$3X_1 \xrightarrow{3} 4X_1 \xrightarrow{10} \emptyset \xrightarrow{1} X_1 \xrightarrow{2} 2X_1.$$

For the transition 'vector' $\mathbf{z} = 1$, the transition rate of the associated CTMC $\mathbf{X}(t)$ is

$$\lambda_{\mathbf{z}}(x_1) = 1 + 2x_1 + 3x_1(x_1 - 1)(x_1 - 2).$$

Using the Gillespie SSA, we generate $Q = 10^2$ independent sample time trajectories of this system each of which contains $q(i) = 10^3$ transition times τ_k^i and the corresponding states $X_1^i(\tau_k^i)$, for $k = 1, 2, \ldots, 10^3$ and $i = 1, 2, \ldots, 10^2$. Applying (5.2), we obtain for the state $x_1 = 4$ the estimated transition rate $\lambda_{\mathbf{z}}(4) = 80.871$, which compares well with the true transition rate $\lambda_{\mathbf{z}}(4) = 81$.

5.1 Distance between two reaction systems

For a given (unknown) mass-action reaction system $(\mathcal{R}, \mathcal{K})$, suppose we know the number of species and the order of the network. Suppose further that we use transition data associated with $(\mathcal{R}, \mathcal{K})$ to estimate the transition rates of $(\mathcal{R}, \mathcal{K})$ by equation (5.2). Then, we can use the estimated transition rates to infer a reaction system $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ by applying Theorem 3.1. In this section, we discuss how we can measure the accuracy of the inferred reaction system $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ by comparing to the original system $(\mathcal{R}, \mathcal{K})$.

Definition 5.1 For two reaction systems $(\mathcal{R}, \mathcal{K})$ and $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ defined on $\mathbb{Z}^d_{\geqslant 0}$, their distance at time t is defined as the total variance distance as $\|p(\cdot, t) - \overline{p}(\cdot, t)\|_{TV}$, where $p(\mathbf{x}, t) = P(\mathbf{X}(t) = \mathbf{x})$ and $\overline{p}(\mathbf{x}, t) = P(\overline{\mathbf{X}}(t) = \mathbf{x})$ are the probability distributions of the stochastic systems $\mathbf{X}(t)$ and $\overline{\mathbf{X}}$ associated with $(\mathcal{R}, \mathcal{K})$ and $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$, respectively. In particular, we measure the similarity of the two reaction systems on a finite set U with their distance at time t with respect to a finite set U, which we define as:

$$\delta_U = \frac{1}{2} \sum_{\mathbf{x} \in U} \left| p(\mathbf{x}, t) - \overline{p}(\mathbf{x}, t) \right|.$$

An alternative distance can also be defined by measuring the difference between the reaction intensities of $(\mathcal{R}, \mathcal{K})$ and $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ over a fixed finite set.

Definition 5.2 For two reaction systems $(\mathcal{R}, \mathcal{K})$ and $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ defined on $\mathbb{Z}^d_{\geqslant 0}$, let $\mathbf{X}(t)$ and $\mathbf{X}(t)$ be the associated CTMCs with the set of transition vectors $\mathcal{Z}$ and $\overline{\mathcal{Z}}$, respectively. Let further that $\lambda_{\mathbf{z}}$ and $\overline{\lambda}_{\bar{\mathbf{z}}}$ be the transition rates associated with transition vectors $\mathbf{z} \in \mathcal{Z}$ and $\bar{\mathbf{z}} \in \overline{\mathcal{Z}}$, respectively. Then for a fixed finite set U, we define

$$\delta_U^I = \max_{\mathbf{x} \in U} \left\{ \max_{\mathbf{z} \in \mathcal{Z} \cap \overline{\mathcal{Z}}} \left| \lambda_{\mathbf{z}}(\mathbf{x}) - \overline{\lambda}_{\mathbf{z}}(\mathbf{x}) \right|, \max_{\mathbf{z} \in \mathcal{Z}} \lambda_{\mathbf{z}}(\mathbf{x}), \max_{\bar{\mathbf{z}} \in \overline{\mathcal{Z}}} \overline{\lambda}_{\bar{\mathbf{z}}}(\mathbf{x}) \right\}.$$

Both the distances δ_U and δ_U^I measure the similarity of two reaction systems confined to a finite set U. For a given (unknown) reaction system of order N, we can apply Theorem 3.1 to infer a network system using the transition data over $U = \mathbb{S}_N$. Then we can test with either δ_U or δ_U^I how close the inferred network is to the original reaction system. The following example demonstrates this process.

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

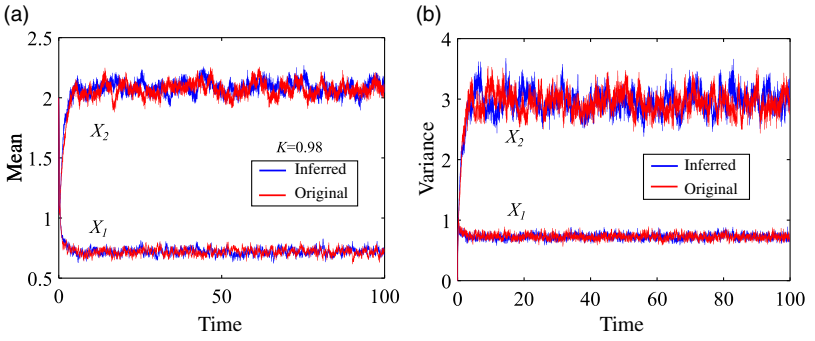

FIGURE 4. (a) Mean values of $X_1(t)$ and $X_2(t)$ of the reaction network (5.3) (red lines) and the inferred reaction network (5.4)–(5.5) (blue lines) obtained by averaging over 10^4 realisations of the Gillespie SSA with initial condition $X_1(0) = 1$ and $X_2(0) = 1$. The average number of transitions by the reactions (5.5) in $\overline{\mathcal{R}} \setminus \mathcal{R}$ is denoted by K. (b) The variance of $X_1(t)$ and $X_2(t)$ estimated from the same time series.

Example 5.2 Consider the following mass-action reaction system of order 3:

$$\begin{gathered} X_1 \underset{1}{\overset{1}{\rightleftarrows}} \emptyset, \qquad X_2 \underset{1}{\overset{1}{\rightleftarrows}} \emptyset, \\ 2X_1 + X_2 \xrightarrow{1} \emptyset \xrightarrow{1} X_1 + X_2 \xrightarrow{1} 2X_1 + 2X_2 \end{gathered} \tag{5.3}$$

We use the Gillespie SSA to simulate the reaction system (5.3) until we collect 15.625×10^5 sample transition times τ_k^i for each state $\mathbf{x}^j \in \mathbb{S}_3$, where $j = 1, 2, \ldots, 10$. Then we estimate the transition rates by (5.2) and apply Theorem 3.1 with the estimated transition rates over $\mathbb{S}_3$. We obtain the mass-action reaction system which contain both original reactions (with modified rate constants):

$$\begin{gathered} X_1 \underset{1.0008}{\overset{0.9999}{\rightleftarrows}} \emptyset, \qquad X_2 \underset{0.9996}{\overset{1.0025}{\rightleftarrows}} \emptyset, \\ 2X_1 + X_2 \xrightarrow{0.9994} \emptyset \xrightarrow{1.0002} X_1 + X_2 \xrightarrow{1.0027} 2X_1 + 2X_2, \end{gathered} \tag{5.4}$$

and additional reactions (with relatively small rate constants):

$$\begin{gathered} 2X_1 + X_2 \xrightarrow{0.0022} X_1 + X_2 \xrightarrow{0.0013} X_2 \xrightarrow{1.3\times10^{-4}} 2X_2 \xrightarrow{5.3\times10^{-4}} 3X_2 \\ 2X_1 \xrightarrow{5.2\times10^{-4}} X_1, \qquad 2X_1 \xrightarrow{4.9\times10^{-4}} 2X_1 + X_2, \end{gathered} \tag{5.5}$$

where the reactions in (5.5) are the reactions in $\overline{\mathcal{R}} \setminus \mathcal{R}$. To compare the original reaction system (5.3) with the inferred reaction system (5.4)–(5.5), we first estimate the distance δ_U by computing the empirical measures with 10^4 realisations of the Gillespie SSA. We obtain $\delta_U = 0.0083$ (for a larger set $U' = \mathbb{S}_{100}$, we get $\delta_{U'} = 0.0244$). The alternative distance δ_U^I can also be computed using the mass-action intensities of the reaction systems as $\delta_U^I = 0.0090$ (for the larger set $U' = \mathbb{S}_{100}$, we get $\delta_{U'}^I = 331.9268$). Mean trajectories of species X_1 and X_2 in the original reaction system (5.3) and the inferred reaction network (5.4)–(5.5) are shown in Figure 4.

Remark 5.1 *As shown in Example 5.2, the distance δ_U is robust to the size of U because this distance is defined using the probability densities. However, the distance δ_U^I is sensitive*

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

to the choice of the set U since the transition rates $\lambda_{\mathbf{z}}(\mathbf{x})$ and $\overline{\lambda}_{\overline{\mathbf{z}}}(\mathbf{x})$ rapidly increase as $\|\mathbf{x}\|_1$ is increased.

5.2 Error analysis

For a given CTMC, the true underlying network structure and the true parameter values are often unknown. Thus, the distance between the true network and the estimated network cannot be calculated. Using the central limit theorem, however, we can find confidence intervals for given stochastic simulation data to ensure that the alternative distance δ_U^I is less than some bound. Let $(\mathcal{R}, \mathcal{K})$ be a given reaction system and let $\lambda_{\mathbf{z}}(\mathbf{x})$ be the transition rate of the associated CTMC. Note that

$$\lambda_{\mathbf{z}}(\mathbf{x}) = P\left(\mathbf{X}(\tau_{k+1}) = \mathbf{x} + \mathbf{z} \,\middle|\, \mathbf{X}(\tau_k) = \mathbf{x}\right)\ \lambda(\mathbf{x})$$

as shown in (5.1), where $\lambda(\mathbf{x}) = \sum_{\mathbf{z}\in\mathcal{Z}} \lambda_{\mathbf{z}}(\mathbf{x})$ is the total intensity of the CTMC $\mathbf{X}(t)$. Thus, letting $\overline{\lambda}(\mathbf{x}) = |G_{\mathbf{x}}| / \left(\sum_{(i,k)\in G_{\mathbf{x}}} \tau_{k+1}^i\right)$ be the sample mean of the total intensity, we define the sample transition rate for a transition vector $\mathbf{z}$ as:

$$\overline{\lambda}_{\mathbf{z}}^{(i,k)}(\mathbf{x}) = \mathbb{1}_{\{\mathbf{X}^i(\tau_{k+1}^i)-\mathbf{X}^i(\tau_k^i)=\mathbf{z}\}}\ \overline{\lambda}(\mathbf{x}).$$

Then, the sample mean of the transition rate $\overline{\lambda}_{\mathbf{z}}(\mathbf{x})$ can be computed as:

$$\overline{\lambda}_{\mathbf{z}}(\mathbf{x}) = \frac{1}{|G_{\mathbf{x}}|} \sum_{(i,k)\in G_{\mathbf{x}}} \mathbb{1}_{\{\mathbf{X}^i(\tau_{k+1}^i)-\mathbf{X}^i(\tau_k^i)=\mathbf{z}\}}\ \lambda(\mathbf{x}) \approx \frac{1}{|G_{\mathbf{x}}|} \sum_{(i,k)\in G_{\mathbf{x}}} \overline{\lambda}_{\mathbf{z}}^{(i,k)}(\mathbf{x}).$$

Then by the central limit theorem, for $\varepsilon > 0$:

$$P\left(\lambda_{\mathbf{z}}(\mathbf{x}) - \varepsilon \leqslant \overline{\lambda}_{\mathbf{z}}(\mathbf{x}) \leqslant \lambda_{\mathbf{z}}(\mathbf{x}) + \varepsilon\right) \approx P\left(-\frac{\varepsilon\sqrt{|G_{\mathbf{x}}|}}{\sigma_{\mathbf{z}}(\mathbf{x})} \leqslant Z \leqslant \frac{\varepsilon\sqrt{|G_{\mathbf{x}}|}}{\sigma_{\mathbf{z}}(\mathbf{x})}\right),$$

where

$$\sigma_{\mathbf{z}}^2(\mathbf{x}) = \frac{1}{|G_{\mathbf{x}}| - 1} \sum_{(i,k)\in G_{\mathbf{x}}} \left(\overline{\lambda}_{\mathbf{z}}^{(i,k)}(\mathbf{x}) - \overline{\lambda}_{\mathbf{z}}(\mathbf{x})\right)^2$$

is the sample variance, and Z is an independent standard normal random variable. Thus, we can formulate the following proposition on confidence intervals.

Proposition 5.1 *Let $(\mathcal{R}, \mathcal{K})$ be a reaction system. For a finite subset $U \subseteq \mathbb{Z}^d_{\geqslant 0}$, let $\overline{\lambda}_{\mathbf{z}}(\mathbf{x})$ and $\sigma_{\mathbf{z}}^2(\mathbf{x})$ be the sample mean and the sample variance for each transition vector $\mathbf{z} \in \mathcal{Z}$ and $\mathbf{x} \in A$, respectively. For some $0 < \alpha < 1$, suppose that $\varepsilon > 0$ satisfies*

$$\varepsilon \geqslant \frac{z_\alpha \sigma_{\mathbf{z}}(\mathbf{x})}{\sqrt{|G_{\mathbf{x}}|}} \quad \text{for each } \mathbf{z} \in \mathcal{Z} \text{ and for each } \mathbf{x} \in U, \tag{5.6}$$

where $[-z_\alpha, z_\alpha]$ is the $(1-\alpha)$-confidence interval of a standard normal random variable, i.e. $P(-z_\alpha \leqslant Z \leqslant z_\alpha) = 1 - \alpha$, where Z is the standard normal random variable. Then for the inferred reaction system $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ obtained by Theorem 3.1 with the sample transition rates $\overline{\lambda}_{\mathbf{z}}(\mathbf{x})$, the distance δ_U^I between $(\mathcal{R}, \mathcal{K})$ and $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$ is less than ε with $(1-\alpha)^{|\overline{\mathcal{Z}}|} \times 100\%$ accuracy, where $\overline{\mathcal{Z}}$ is the set of transition vectors of the CTMC associated with $(\overline{\mathcal{R}}, \overline{\mathcal{K}})$.

https://doi.org/10.1017/S0956792520000492 Published online by Cambridge University Press

Example 5.3 Consider again the inferred reaction system (5.4)–(5.5) in Example 5.2. Note that we have the sample transition rates $\overline{\lambda}_{\mathbf{z}}(\mathbf{x})$ at each state $\mathbf{x} \in \mathbb{S}_3$ and for each transition vector $\mathbf{z}$. Hence, we can calculate the sample variance. We obtain

$\mathbf{x}$	$(0,0)$	$(1,0)$	$(2,1)$	$\ldots$	$(3,0)$
$\mathbf{z}$	$(1,1)$	$(1,1)$	$(-1,0)$	$\ldots$	$(-1,0)$
$\overline{\lambda}_{\mathbf{z}}(\mathbf{x})$	1.002	0.9984	2.0073	$\ldots$	2.9975
$\sigma_{\mathbf{z}}(\mathbf{x})$	2.007	2.9966	8.0045	$\ldots$	2.9972

Since six transition vectors are inferred in Example 5.2 (i.e. $|\overline{\mathcal{Z}}| = 6$) and $(1-\alpha)^6 = 0.95$ with $\alpha = 0.0085$, we choose $z_\alpha \approx 2.4$ such that $P(-z_\alpha \leqslant Z \leqslant z_\alpha) = 1-\alpha$ with the standard normal random variable Z. Hence, if we let

$$\varepsilon = 0.0172 = \max_{\mathbf{x}\in\mathbb{S}_3, \mathbf{z}\in\mathcal{Z}} z_\alpha \sigma_{\mathbf{z}}(\mathbf{x}) |G_{\mathbf{x}}|^{-1/2},$$

then the distance $\delta^I_{\mathbb{S}_3}$ between the given system and the estimated reaction system is less than 0.0172 with 95% accuracy.

6 Discussion

In this paper, we have explored identifiability of reaction systems. Identifiability of a stochastic reaction system $(\mathcal{R}, \mathcal{K})$ holds if this is the only set of reactions that produces its transition rates on the corresponding state space. Therefore, identifiability of a reaction system must be verified *prior to* inference of a network structure and parameter estimation. Using the fact that a mass-action system is fully characterised with the transition rates on a certain finite region, we have proved that any stochastic mass-action system of order at most N is identifiable as long as the associated state space contains $\mathbb{S}_N$.

Using the mass-action property, we have also proposed an algorithm that enables us to infer the underlying reaction network and the associated parameters with the transition data of a given CTMC. In the case that the transition data are given by stochastic simulations, we have investigated how to approximate the true transition data, and in turn, how to infer an estimated underlying network. Then by using the confidence intervals, we can measure the accuracy of the estimated underlying network comparing to the true network.

The presented network inference method relies on the exact transition data consisting of the transition vectors and the transition times. Hence, this method is not directly applicable to data that consist of partial information of the system at discrete time points. However, we have shown that as the transition information and confidence on transition rate estimates increases, the distance between the actual and approximated networks tends to decrease. Given that increasingly precise measurements are being made for specific reaction networks in experimental studies, we expect that the presented method can be used in the future to infer underlying networks and kinetic parameters for realistic biological systems.

Acknowledgements

Radek Erban and German Enciso would like to thank the organisers of the ‘Recent Developments in Mathematical and Computational Biomedicine’ (19w5085) workshop at the Casa Matemática

Oaxaca (CMO) in Oaxaca, in November 2019, where this research project was initiated. German Enciso and Jinsu Kim are partially supported by NSF grant DMS1763272, Simons Foundation grant 594598 (Qing Nie) and by NSF grant DMS1616233.

Conflict of interest

None.

References

[1] Anderson, D. & Kurtz, T. (2011) Continuous time Markov chain models for chemical reaction networks. In: H. Koeppl (editor), *Design and Analysis of Biomolecular Circuits: Engineering Approaches to Systems and Synthetic Biology*, Springer, New York, pp. 3–42.
[2] Anderson, D. & Kurtz, T. (2015) *Stochastic Analysis of Biochemical Systems*, Springer, Cham.
[3] Angeli, D. (2009) A tutorial on chemical reaction network dynamics. *Eur. J. Control* **15**, 398–406.
[4] Baldi, P. & Brunak, S. (2001) *Bioinformatics: The Machine Learning Approach*, MIT Press.
[5] Catanach, T., Vo, H. & Munsky, B. (2020) Bayesian inference of stochastic reaction networks using multifidelity sequential tempered Markov chain Monte Carlo. arXiv preprint arXiv:2001.01373.
[6] Chattopadhyay, I., Kuchina, A., Süel, G. & Lipson, H. (2013) Inverse Gillespie for inferring stochastic reaction mechanisms from intermittent samples. *Proc. Natl. Acad. Sci.* **110**(32), 12990–12995.
[7] Craciun, G. & Feinberg, M. (2006) Multiple equilibria in complex chemical reaction networks: Ii. the species-reactions graph. *SIAM J. Appl. Math.* **66**(4), 1321–1338.
[8] Craciun, G., Kim, J., Pantea, C. & Rempala, G. (2013) Statistical model for biochemical network inference. *Commun. Stat. Simul. Comput.* **42**(1), 121–137.
[9] Craciun, G. & Pantea, C. (2008) Identifiability of chemical reaction networks. *J. Math. Chem.* **44**, 244–259.
[10] Duncan, A., Liao, S., Vejchodský, T., Erban, R. & Grima, R. (2015) Noise-induced multistability in chemical systems: discrete versus continuum modeling. *Phys. Rev. E* **91**, 042111.
[11] Erban, R., Chapman, S. J., Kevrekidis, I. & Vejchodský, T. (2009) Analysis of a stochastic chemical system close to a SNIPER bifurcation of its mean-field model. *SIAM J. Appl. Math.* **70**(3), 984–1016.
[12] Erban, R. & Chapman, S. J. (2020) *Stochastic Modelling of Reaction-Diffusion Processes*. Cambridge Texts in Applied Mathematics, Cambridge University Press, 308 p.
[13] Feinberg, M. (1989) Necessary and sufficient conditions for detailed balancing in mass action systems of arbitrary complexity. *Chem. Eng. Sci.* **44**(9), 1819–1827.
[14] Feinberg, M. (2019) *Foundations of Chemical Reaction Network Theory*, Springer, Cham, Switzerland.
[15] Gadgil, C., Lee, C. & Othmer, H. (2005) A stochastic analysis of first-order reaction networks. *Bull. Math. Biol.* **67**, 901–946.
[16] Golightly, A. & Wilkinson, D. (2006) Bayesian sequential inference for stochastic kinetic biochemical network models. *J. Comput. Biol.* **13**(3), 838–851.
[17] Gupta, A. & Rawlings, J. (2014) Comparison of parameter estimation methods in stochastic chemical kinetic models: examples in systems biology. *AIChE J.* **60**(4), 1253–1268.
[18] Jahnke, T. & Huisinga, W. (2007) Solving the chemical master equation for monomolecular reaction systems analytically. *J. Math. Biol.* **54**(1), 1–26.
[19] Komorowski, M., Costa, M., Rand, D. & Stumpf, M. (2011) Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proc. Natl. Acad. Sci.* **108**(21), 8645–8650.
[20] Kurtz, T. (1972) The relationship between stochastic and deterministic models for chemical reactions. *J. Chem. Phys.* **57**(7), 2976–2978.

[21] LANGARY, D. & NIKOLOSKI, Z. (2019) Inference of chemical reaction networks based on concentration profiles using an optimization framework. *Chaos Interdiscip. J. Nonlinear Sci.* **29**(11), 113121.

[22] LIAO, S., VEJCHODSKÝ, T. & ERBAN, R. (2015) Tensor methods for parameter estimation and bifurcation analysis of stochastic reaction networks. *J. R. Soc. Interface* **12**(108), 20150233.

[23] LOSKOT, P., ATITEY, K. & MIHAYLOVA, L. (2019) Comprehensive review of models and methods for inferences in bio-chemical reaction networks. *Front. Genet.* **10**, 549.

[24] MARKOWETZ, F. & SPANG, R. (2007) Inferring cellular networks–a review. *BMC Bioinformatics* **8**(6), S5.

[25] PLESA, T., ERBAN, R. & OTHMER, H. (2019) Noice-induced mixing and multimodality in reaction networks. *Eur. J. Appl. Math.* **30**, 887–911.

[26] PLESA, T., VEJCHODSKÝ, T. & ERBAN, R. (2016) Chemical reaction systems with a homoclinic bifurcation: an inverse problem. *J. Math. Chem.* **54**(10), 1884–1915.

[27] PLESA, T., VEJCHODSKÝ, T. & ERBAN, R. (2017) Test models for statistical inference: two-dimensional reaction systems displaying limit cycle bifurcations and bistability. In: *Stochastic Processes, Multiscale Modeling, and Numerical Methods for Computational Cellular Biology*, Springer International Publishing, Cham, Switzerland, pp. 3–27.

[28] PLESA, T., ZYGALAKIS, K., ANDERSON, D. & ERBAN, R. (2018) Noise control for molecular computing. *J. R. Soc. Interface* **15**(144), 20180199.

[29] SZEDERKÉNYI, G., BANGA, J. & ALONSO, A. (2011) Inference of complex biological networks: distinguishability issues and optimization-based solutions. *BMC Syst. Biol.* **5**(1), 177.

[30] VILLAVERDE, A. & BANGA, J. (2014) Reverse engineering and identification in systems biology: strategies, perspectives and challenges. *J. R. Soc. Interface* **11**(91), 20130505.

[31] WALTER, E. & PRONZATO, L. (1997) Identification of parametric models from experimental data. *Commun. Control Eng.*, Springer, London, **8**.

[32] WANG, S., LIN, J., SONTAG, E. & SORGER, P. (2019) Inferring reaction network structure from single-cell, multiplex data, using toric systems theory. *PLOS Comput. Biol.* **15**, 1–25.

[33] WARNE, D., BAKER, R. & SIMPSON, M. (2019) Simulation and inference algorithms for stochastic biochemical reaction networks: from basic concepts to state-of-the-art. *J. R. Soc. Interface* **16**(151), 20180943.