QUANTICOL

A Quantitative Approach to Management and Design of Collective and Adaptive Behaviours

quantiCOL

http://www.quanticol.eu

TR-QC-4-2014

Multi-Scale Reductions of Mean-Field and Stochastic Models

Tutorial and a survey of the literature

Revision: 1.0; Feb 5, 2014

Author(s): Rytis Paškauskas (CNR), Luca Bortolussi (Univ. Trieste & CNR)

Publication date: Feb 5, 2014

Funding Scheme: Small or medium scale focused research project (STREP)
Topic: ICT-2011 9.10: FET-Proactive 'Fundamentals of Collective Adaptive Systems' (FOCAS)
Project number: 600708

Coordinator: Jane Hillston (UEDIN)
e-mail: Jane.Hillston@ed.ac.uk
Fax: +44 131 651 1426

Part. no.	Participant organisation name	Acronym	Country
1 (Coord.)	University of Edinburgh	UEDIN	UK
2	Consiglio Nazionale delle Ricerche – Istituto di Scienza e Tecnologie della Informazione "A. Faedo"	CNR	Italy
3	Ludwig-Maximilians-Universität München	LMU	Germany
4	Ecole Polytechnique Fédérale de Lausanne	EPFL	Switzerland
5	IMT Lucca	IMT	Italy
6	University of Southampton	SOTON	UK

COOPERATION

Contents

1 Introduction

Population continuous Markov chain is a framework for describing Markovian population dynamics. At the basis of any PCTMC, is the set of rules that characterize completely all possible states and transitions of populations. There are several representations of PCTMCs, the two common ones being the 'master equation' or 'Kolmogorov equation' on the one hand, and the mean field 'rate equations' on the other hand. There are other, mixed options, but we will not consider them here. An important

feature that is characteristic to most PCTMCs is the presence of multiple temporal scales, either through explicit rate constants or as an intrinsic property.

This report aims at presenting the temporal multi-scale phenomena, the associated potential of model reductions, and lastly, to survey the original chemistry and physics literature on the topic. Our specific interest here is to explore the links between the known mean field reductions, and the appropriate stochastic reductions of which, unfortunately, only a few studies exist. Because there are only two groups of mean field reductions, going by the quasi-steady state (QSS) and quasi-equilibrium (QE) approximation, the discussion of stochastic reductions will be limited to those approximations. This does not exhaust all the possibilities of stochastic reduction. Notably, we do not discuss the most popular topic, the matrix perturbation theory, because it doesn't have a useful mean field equivalent. For matrix methods, we can point to the general literature [Ste90, BO91, GVL12, Ste01, Kat95] and remark that the simplest, singular perturbation theory setting could probably be reformulated using the matrix theory (in terms of generators).

The review is structured with the goal of presenting an easy tutorial first (sections 3, 4, 5) and a collection of hopefully useful references second (sections 6 and 7). For compleness, the basic notions of a PCTMC, the master equation, and the mean field limit are summarized in section 2. To illustrate key ideas and to motivate, the Wireless Sensor Network model is presented in section 3. Is a quintessential, familiar model[1], and yet it can teach a few valuable lessons. The original QSS/QE approximations were invented as mean field reductions; we start there as well, emphasizing their differences in section 4, all the while avoiding the hard subject of differential equations (the only ODE-related prerequisite is the knowledge of how to solve $\dot{x} = -kx$). Then, concrete implementations of the WSN model reductions are presented and nicely illustrated in section 5

2 Population Continuous Time Markov Chain (PCTMC)

Population continuous time Markov chain (PCTMC) is a tuple $\chi_C = (\mathbf{X}, \mathcal{M}, \mathcal{T}, \mathbf{D})$, where

1. $\mathbf{X} = (X_1, \ldots, X_n)$ is a vector of variables describing n species of the model.
2. $\mathcal{M}$ is the domain of $\mathbf{X}$. Usually X_i counts the elements in a population of species i, therefore we can assume that $X_i \subset \mathbb{N}$ and $\mathcal{M} \subset \mathbb{N}^n$ (in order to allow the extinction of populations, we assume $0 \in \mathbb{N}$).
3. $\mathcal{T} = \{\tau_1, \ldots, \tau_r\}$ is the set of r transitions of the form $\tau = (\lambda, \mathbf{b}, \mathbf{a}, w)$, where:
   (a) λ is the label of a transition,
   (b) $\mathbf{a} = (a_1, \ldots, a_n)$ is the *post-vector*; $a_i \in \mathbb{N}$ is the number of agents of species i, created by the transition.
   (c) $\mathbf{b} = (b_1, \ldots, b_n)$ is the *pre-vector*; $b_i \in \mathbb{N}$ is the number of agents of species i, consumed by the transition.
   (d) $w : \mathcal{M} \mapsto \mathbb{R}^+ \cup \{0\}$ is the *rate function*, satisfying the conditions $w(\mathbf{X}) \geq 0$, and $w(\mathbf{X}) = 0$ if $\mathbf{X} + \mathbf{a} - \mathbf{b} \notin \mathcal{M}$.

   The pre-, and post-vectors are combined in a *state-change vector* $\mathbf{s} = \mathbf{a} - \mathbf{b}$. This vector gives the net change on each variable due to the transition.
4. $\mathbf{D} \in \mathcal{M}$ is the initial data (population counts at $t = 0$).

[1] by being inspired by the Michaelis-Menten model[JG11] which is well known in chemistry and biology literature

2.1 Stochastic model

The state of a stochastic model in continuous time is described by $P(\mathbf{X};t)$, which is a joint probability of there being $\mathbf{X}$ agents in the population at time t, conditional to an unspecified initial probability at $t=0$. The probability at later times is found by solving a system of linear ODEs, known as the 'Master equation' (ME) or 'Kolmogorov's equation'[VK92, Gar85]

$$\partial_t P(\mathbf{X};t) = \sum_{j=1}^{r} w_j(\mathbf{X}-\mathbf{s}_j)P(\mathbf{X}-\mathbf{s}_j;t) - \sum_{j=1}^{r} w_j(\mathbf{X})P(\mathbf{X};t)\,. \tag{1}$$

A slightly different formulation of the same equation can be found in the literature

$$\partial_t P(\mathbf{X};t) = \sum_{\mathbf{X}'\in\mathcal{M}} W(\mathbf{X}|\mathbf{X}')P(\mathbf{X}';t) - \sum_{\mathbf{X}'\in\mathcal{M}} W(\mathbf{X}'|\mathbf{X})P(\mathbf{X};t)\,.$$

The relationship between the two is established by $W(\mathbf{X}|\mathbf{X}') = \sum_j w_j(\mathbf{X}')\delta_{\mathbf{X},\mathbf{X}'+\mathbf{s}_j}$.

2.2 Mean field limit

The mean field model is described in terms of the population density vector $\mathbf{x}(t)$ which is the average population size, normalized to the total population N, in the limit when total population goes to infinity, i.e. $\mathbf{x}(t) = \lim_{N\to\infty} N^{-1}\langle\mathbf{X}\rangle(t)$, where the average of some function $f(\mathbf{X})$ is defined by $\langle f(\mathbf{X})\rangle(t) = \sum_{\mathbf{X}} f(\mathbf{X})P(\mathbf{X};t)$. Its temporal evolution is governed by a system of ODEs, called the *rate equations*:

$$\partial_t \mathbf{x} = \sum_{j=1}^{r} \mathbf{s}_j w_j(\mathbf{x}) = \mathsf{S}\cdot\mathbf{w}(\mathbf{x}) \tag{2}$$

Here $\mathsf{S} = (\mathsf{s}_1, \mathsf{s}_2, \ldots, \mathsf{s}_r)$ is the *stoichiometry matrix*, consisting of all the state change vectors arranged as columns ("$\cdot$" denotes a matrix-vector product).

3 Wireless sensor network model (WSN)

Consider a model of a network of sensors and gateways, whose purpose is to collect and transmit data. It could be pictured as two layers, with N_s sensors placed on the top layer, and N_g gateways placed on the bottom layer, as in figure 1. The total number of elements in both layers is $N = N_s + N_g$. In principle, there is the ambient space in which the elements are distributed, but to simplify the discussion, we will assume that the system is 'well mixed', meaning that all the elements are distributed randomly, and that each element of the gateway layer can communicate with any element in the sensor layer.

A sensor is modelled as a automaton with three states, with the following interpretation and notation:

- sensor in a 'collecting data' state (SC) $X_1 = \#\mathrm{SC}$
- sensor in a 'transmitting data' state (ST) $X_2 = \#\mathrm{ST}$
- sensor in a 'done' state (SD) $X_3 = \#\mathrm{SD}$

In our simple model, sensors are single-use devices; when a sensor is dispatched, it becomes inactive. The variables X_1, X_2, X_3 contain the count of sensors in a corresponding state. Colors match figure 1.

A gateway is modelled as an even simpler automaton with two states:

- gateway in an 'idle' state (GI) $X_4 = \#\mathrm{GI}$
- gateway in a 'receiving data' state (GR) $X_5 = \#\mathrm{GR}$

Similarly, X_4 and X_5 contain population counts in a corresponding state.

The variable of the WSN model is therefore five-dimensional: $\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)^\intercal$.

To complete the definition of our PCTMC, We propose the following scenario of transitions. To make a data transfer, an idle gateway (GI) must establish a connection with one collecting sensor (SC)

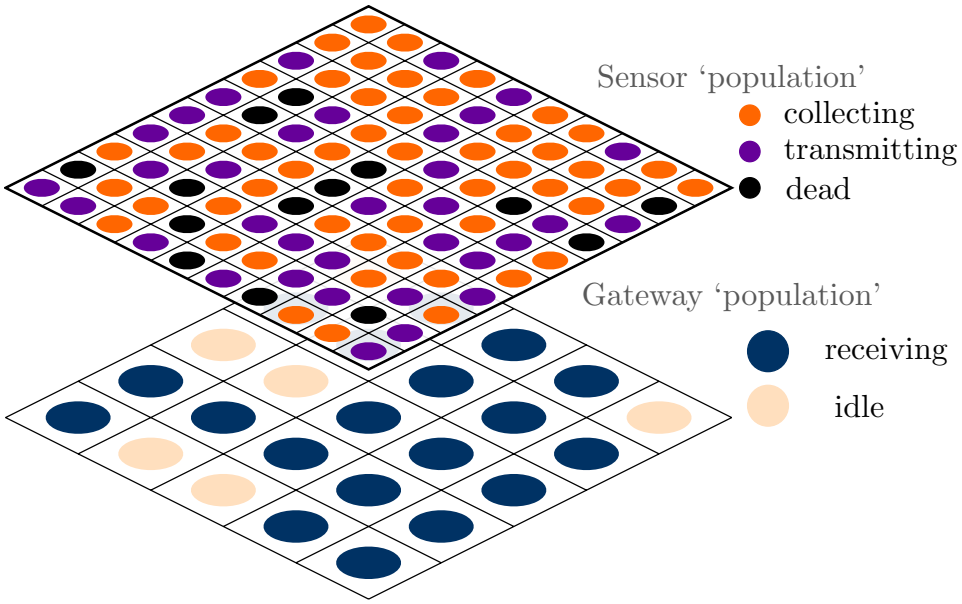


Figure 1: Two layer WSN model, populated by $N_s = 100$ sensors (upper layer) and $N_g = 25$ gateways (lower layer), each of which is assigned a state randomly, with probabilities $p(\mathrm{SC}) = 1/2$, $p(\mathrm{ST}) = 1/3$, $p(\mathrm{SD}) = 1/6$, $p(\mathrm{GR}) = 3/4$, $p(\mathrm{GI}) = 1/4$.

to become an ST-GR pair. It is easy to see that one unit of both X_1 and X_4 is consumed, and one unit of both X_2 and X_5 is created in this transition, so the state change vector is $\mathbf{s}_1 = (-1, +1, 0, -1, +1)^\intercal$.

The fundamental quantity of the model is the probability of a transition to occur within a small time interval Δt. This probability is $w_1(\mathbf{X}) \cdot \Delta t$, where $w_1(\mathbf{X})$ is interpreted as a rate (or frequency) of transition[Gil77]. We suppose that each gateway can decide to initiate this transition independently of others, so in a system consisting of N_g gateways, the correct scaling of the frequency is $w_1 \sim N_g$. We postulate that this transition is proportional to a joint probability that a die, tossed on the upper plane in figure 1, lands in a square marked by ● (labelled SC) *and* that a second die, tossed onto the lower plane lands in a square marked by ● (labeled GI), i.e.

$$w_1(\mathbf{X}) = k_1 X_1 X_4 / N_s \,.$$

Normalization by the total number of elements $N = N_s + N_g$ could also have been used, but N_s will be more convenient in the limit $N \to \infty$. There is no difference between the two normalizations so long as N_s and N approach ∞ at the 'same speed', so that

$$n_g = \frac{N_g}{N_s} \tag{3}$$

is a fixed number. This scaling would not be appropriate in other situations.

Transmission could be interrupted due to something going awry with the transmitting sensor, resulting in a breakup of an existing ST-GR pair. The state change vector of this transition is $-\mathbf{s}_1$. The frequency of this transition is proportional to a chance of landing a die in a square marked with ● and labeled (ST),

$$w_2(\mathbf{X}) = k_2 X_2 \,.$$

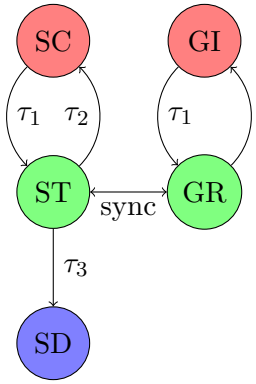
An existing ST-GR pair breaks up upon a successful completion. The state change vector is now $\mathbf{s}_2 = (0, -1, +1, +1, -1)$. We follow the same argument as before for the frequency, obtaining

$$w_3(\mathbf{X}) = k_3 X_2 \,.$$

To complete the discussion we write out all the transitions with their corresponding labels:

$$\begin{aligned}\tau_1 = &\quad (\text{connect}, &\quad \mathbf{s}_1, &\quad w_1)\\ \tau_2 = &\quad (\text{disconnect}, &\quad -\mathbf{s}_1, &\quad w_2)\\ \tau_3 = &\quad (\text{dispatch}, &\quad \mathbf{s}_2, &\quad w_3)\end{aligned}$$

We may draw these transitions as a picture (make it a Petri net?):



3.1 Stoichiometry reduction of the WSN model

The stoichiometry matrix S is formed from all state change vectors, arranged as columns, so it is an $n \times r$ integer valued matrix. The stoichiometry matrix of the WSN model is

$$\mathsf{S} = \left(\mathsf{s}_1, -\mathsf{s}_1, \mathsf{s}_2\right) = \begin{pmatrix} -1 & 1 & 0 \\ 1 & -1 & -1 \\ 0 & 0 & 1 \\ -1 & 1 & 1 \\ 1 & -1 & -1 \end{pmatrix}. \tag{4}$$

Stoichiometry matrix is useful, because it allows to analyze the first basic type of reduction, the *stoichiometry reduction*. Consider the following three row vectors:

$$\begin{aligned}\mathbf{I}_1 &= (1,1,1,0,0)\\ \mathbf{I}_2 &= (0,0,0,1,1)\\ \mathbf{I}_3 &= (0,1,0,0,-1)\end{aligned}$$

As one may easily check, each $\mathbf{I}_i$ nullifies S: $\mathbf{I}_i \cdot \mathsf{S} = 0$. To each $\mathbf{I}$, we associate an *invariant* $\mathcal{I}_i = \mathbf{I}_i \cdot \mathbf{X}$. Invariants remain unchanged with respect to any of the transitions allowed by the model. Each invariant may be assigned a sensible value $\mathcal{I}_i = c_i$ that describes the initial state of the system. Their meaning becomes more clear as we look at each of them individually: $\mathcal{I}_1 = X_1 + X_2 + X_3$, $\mathcal{I}_2 = X_4 + X_5$, $\mathcal{I}_3 = X_2 - X_5$. Now it is easy to guess that the proper constants are $\mathcal{I}_1 = N_s$, which expresses the fact that there is a fixed number N_s of sensors among all three groups SC, ST, SD; and $\mathcal{I}_2$ should be assigned a value N_g which is, similarly, the total number of gateways. It is not evident what number to assign to the last one, I_3. The correct answer is that any valid number could be used, different choices implying something about the initial state of the model. $\mathcal{I}_3 = 0$ implies that each transmitting sensors is coupled to an idle gateway, $\mathcal{I}_3 = 1$ implies that there is one unaccounted for 'transmitting' sensor (a spy?), etc. In the following we will take $\mathcal{I}_3 = 0$.

You can look at invariants in this way: an invariant is a licence to eliminate one dimension (variable) from the model. In the case of WSN, it is easier done that explained how X_3, X_4 and X_5 may be eliminated:

$$X_3 = N_s - X_1 - X_2, \quad X_4 = N_g - X_2, \quad X_5 = X_2\,, \tag{5}$$

leaving us with only two independent variables, X_1 and X_2.

The general stoichiometry reduction can be summarized by the following

Rules of thumb of stoichiometry reduction

- verify that $\operatorname{rank} \mathsf{S} < \min(r, n)$. Stoichiometry reduction is only possible if S is rank-deficient;
- identify the redundant variables to be eliminated;
- work directly with transitions $\mathcal{T}$ of the PCTMC model, and not with derived models (equations). The number and 'label' of transitions must remain the same as in the original model;
- for each transition, eliminate the redundant variables in favor of the constant c_is;
- truncate the state space by leaving only non-redundant variables;
- pay attention to the domain of the reduced model: because of various constraints it may have a complicated shape.

Following these rules we get the transitions of the reduced model. Since in the following we will only look at the stoichiometry-reduced model, from now on we use the same letters τ_i, w_i, $\mathbf{X} = (X_1, X_2)^\intercal$ to denote the reduced quantities. So

$$\begin{aligned} \tau_1 = &\left\{\text{connect}, \quad \mathbf{s}_1 = (-1,+1)^\intercal, \quad w_1 = k_1 X_1 (N_g - X_2)/N_s \right\} \\ \tau_2 = &\left\{\text{disconnect}, \quad -\mathbf{s}_1 = (+1,-1)^\intercal, \quad w_2 = k_2 X_2 \right\} \\ \tau_3 = &\left\{\text{dispatch}, \quad \mathbf{s}_2 = (0,-1)^\intercal, \quad w_3 = k_3 X_2 \right\} \end{aligned} \tag{6}$$

3.2 Differential equation representations

3.2.1 The master equation

A probabilistic state of the WSN model is described by $P(X_1, X_2; t)$, which is a joint probability of there being an integer number X_1 of collecting sensors (SC) and X_2 of transmitting sensors (ST) at a time t, conditional to some unspecified initial state at $t = 0$. The probability at some later time is found by solving a system of linear ODEs, known as the 'Master equation' (ME) or 'Kolmogorov's equation'. The Master equation of the (reduced) WSN model reads

$$\begin{aligned} \partial_t P(\mathbf{X}) \;=\; & w_1(\mathbf{X} - \mathbf{s}_1) \cdot P(\mathbf{X} - \mathbf{s}_1) + w_2(\mathbf{X} + \mathbf{s}_1) \cdot P(\mathbf{X} + \mathbf{s}_1) \\ & + w_3(\mathbf{X} - \mathbf{s}_2) \cdot P(\mathbf{X} - \mathbf{s}_2) - \left[w_1(\mathbf{X}) + w_2(\mathbf{X}) + w_3(\mathbf{X})\right] \cdot P(\mathbf{X}) \end{aligned} \tag{7}$$

This equation is defined on a *discrete state space* in two dimensions (see figure 2) that is constrained by

$$0 \le X_1 + X_2 \le N_s, \quad 0 \le X_1 \le N_s, \quad 0 \le X_2 \le N_g. \tag{8}$$

3.2.2 The rate equations

We would like to take the $N_s, N_g \to \infty$ limit and to pass from the representation by the probability $P(\mathbf{X}; t)$ to a representation in terms of variables, and we would like these coordinates to have a non-trivial $N = \infty$ limit. For that purpose, define 'densities' or 'concentrations' of agents as

$$\begin{aligned} x_1 &= \langle X_1 \rangle (t)/N_s \\ x_2 &= \langle X_2 \rangle (t)/N_s \,, \end{aligned}$$

where the mean is defined by $\langle \mathbf{X} \rangle (t) = \sum_{\mathbf{X}} \mathbf{X} P(\mathbf{X}; t)$. If we derivate both sides of this expression with respect to time, insert the master equation on the right, and make a few simple manipulations within this equation, we can find an exact differential equation for the expectation values:

$$\frac{\mathrm{d} \langle \mathbf{X} \rangle}{\mathrm{d}t} = \mathbf{s}_1 \langle w_1(\mathbf{X}) \rangle - \mathbf{s}_2 \langle w_2(\mathbf{X}) \rangle + \mathbf{s}_2 \langle w_3(\mathbf{X}) \rangle \,. \tag{9}$$

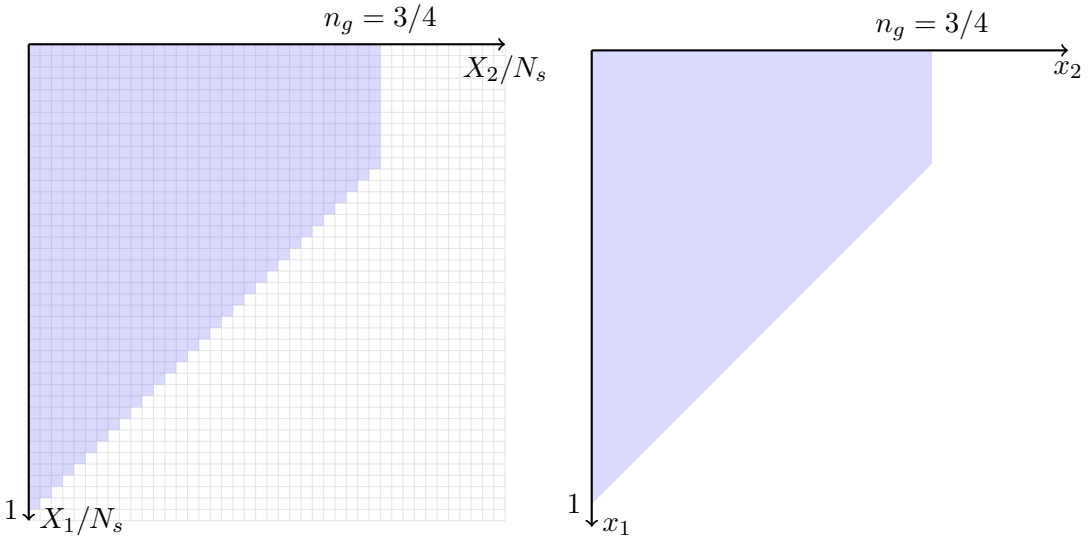


Figure 2: Domains of definition (shaded areas). The left panel represents a discrete state space of (7) with $N_s = 40$ and $N_g = 30$. For visualization purposes, each point (X_1, X_2) is represented as a little square, making the domain look as a filled area (a useful illusion). On the right, we have a continuous $\subset \mathbb{R}^2$ space of densities in the mean field limit. The two models cannot be compared directly, because the sets of parameters are different (the situation on the right-hand side corresponds to $N_s = \infty$ and $N_s = \infty$) – we would be comparing 'apples and apple jam'. We use a rather unusual orientation of axes in this and the subsequent phase diagrams; we believe that this arrangement reinforces a useful impression of X_1 and X_2 as being the row and column indices of a matrix (in a row-major format).

The mean field limit consists of replacing the averages of the form $\langle X_i X_j \rangle$ for all pairs i and j with 'uncorrelated' averages $\langle X_i \rangle \langle X_j \rangle$, and passing to the limit $N_s, N_g \to \infty$ (see [BHLM13] for the limiting procedure). The mean field limit results in the exact system of *rate equations*[2]

$$\begin{aligned} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} &= \mathbf{s}_1 w_1 - \mathbf{s}_1 w_2 + \mathbf{s}_2 w_3 && (10)\\ w_1 &= k_1 x_1 (n_g - x_2) && (11)\\ w_2 &= k_2 x_2 && (12)\\ w_3 &= k_3 x_3 && (13) \end{aligned}$$

or

$$\begin{aligned} \begin{pmatrix} \frac{\mathrm{d}x_1}{\mathrm{d}t} \\ \frac{\mathrm{d}x_2}{\mathrm{d}t} \end{pmatrix} &= \begin{pmatrix} -1 & +1 & 0 \\ +1 & -1 & -1 \end{pmatrix} \cdot \begin{pmatrix} k_1 x_1 (n_g - x_2) \\ k_2 x_2 \\ k_3 x_2 \end{pmatrix} \\ &= \begin{pmatrix} k_1 x_1 (x_2 - n_g) + k_2 x_2 \\ -k_1 x_1 (x_2 - n_g) - (k_2 + k_3) x_2 \end{pmatrix} \end{aligned} \qquad (14)$$

From (8), we get the analogous constraints for the continuous state space (see figure 2)

$$0 \le x_1 + x_2 \le 1, \quad 0 \le x_1 \le 1, \quad 0 \le x_2 \le n_s\,. \qquad (15)$$

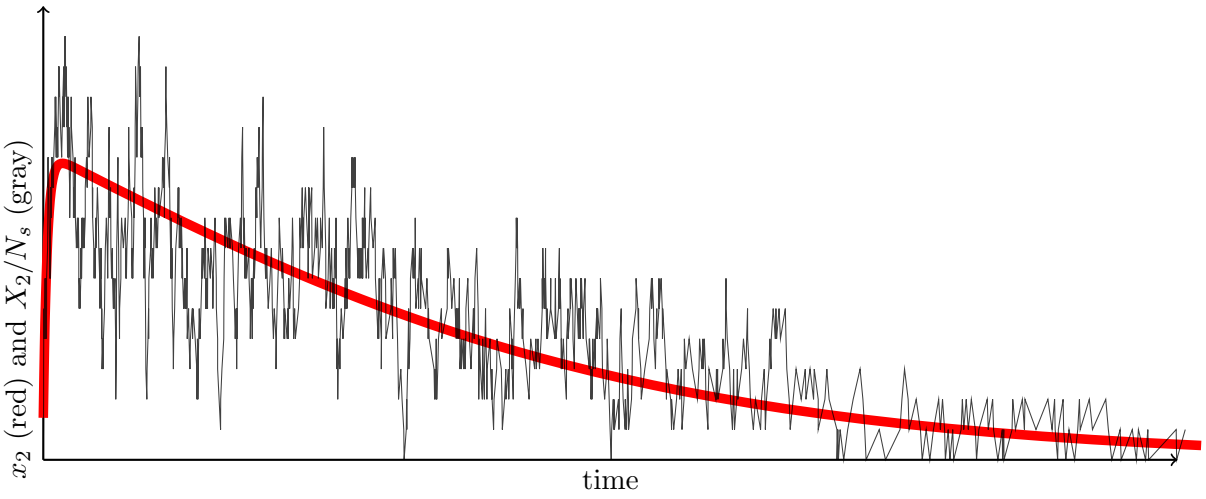


Figure 3: Stochastic time series of X_2 with $N_s = 40$ and $N_g = 30$ (gray) and the corresponding mean field curve, with rate constants as in the QE approximation (see section 5.1) and $T = 100$. The initial steep rise defines the fast time scale, which is more than $\times 10$ shorter than the slow scale.

4 Overview of multiple time scale reductions

The term *dimensional reduction* means that a model of interest may be replaced by another model with fewer degrees of freedom. The reduced model is expected to be either exactly equivalent to the original model, or a reasonably good approximation thereof. In order to be reducible in one way or another, a model must qualify for the particular kind of reduction. Here we will discuss the qualifying factors for a special class of reduction techniques, based on the hypothesis that the model describes processes with largely separated temporal scales. We should mention that there are also other reduction techniques, based on different principles. For example, any master equation qualifies for the exact mean field reduction, that is based on the central limit theorem. Stoichiometry reduction is another reduction, based on the rank-deficiency, as discussed in section 3.1.

The effects of multiple temporal scales can be visualized by plotting a curve of some coordinate, say, $x_2(t)$ vs. t, as in figure 3. The curve should have at least two clearly identifiable legs: a brief, steep initial rise, followed by a 'knee', followed by the second, slowly varying leg, extending to $t = T$. The temporal extents over which x_2 changes by roughly the same amount on the first and second legs are, respectively, the fast and slow time scales. In the example of figure 3, the fast scale is ≈ 10, and the slow scale is ≈ 100 units of time. The objective of multiple time scale reduction is to eliminate some variables, so that the second leg of the curve would still be reproducible.

We shall limit this section with discussing mean field reductions because stochastic reductions are still a developing field. In the following section, we will make some links between the two.

After reviewing some literature, we came up with the notion of there being four levels of multi-scale reductions:

- canonical form of the singular perturbation theory;
- Quasi-equilibrium approximations;
- Quasi-steady state approximations;
- Purely numerical methods.

[2]The general form of rate equations is given by (2).

4.1 Singular perturbation theory

The conceptually simplest case to describe is the *singular perturbation theory* [Ver05], applicable when the mean field ODE is of a special, 'canonical' form. We require three assumptions to be satisfied: 1) existence of a small, controllable parameter ϵ; 2) possibility to partition $\mathbf{x}$ in two groups of slow, $\mathbf{y}$, and fast, $\mathbf{z}$, variables, $\mathbf{x} = (\mathbf{y}, \mathbf{z})$; and 3) the ODE is in the form:

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\epsilon t} = \mathbf{g}(\mathbf{y}, \mathbf{z}) \,, \quad \epsilon \frac{\mathrm{d}\mathbf{z}}{\mathrm{d}\epsilon t} = \mathbf{h}(\mathbf{y}, \mathbf{z}) \,. \tag{16}$$

We used the 'slow time' $\tau = \epsilon t$ scaling, because the slow scale is usually of interest. In the limit $\epsilon = 0$, the second differential equation is killed[3] but the solution must satisfy

$$\mathbf{h}(\mathbf{y}, \mathbf{z}) = 0 \,. \tag{17}$$

In other words, the solution must lie on a surface, defined by (17), called the 'slow manifold'.

Let's assume that a unique solution, $\mathbf{z} = \mathbf{h}^{-1}(\mathbf{y})$, of (17), exists. We have to solve only the former differential equation[4], expressing the missing variable via the slow manifold equation,

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\tau} = \mathbf{g}(\mathbf{y}, \mathbf{h}^{-1}(\mathbf{y})) \,. \tag{18}$$

Just like the variables were separated into fast and slow ones, we can think of $\mathbf{h}$ as representing all the fast transitions, and $\mathbf{g}$ as representing all the slow transitions. Bringing rate equations to this form is the ultimate objective of the (mean field) multi-scale reduction. What we have described here goes by the name Tikhonov's theorem [Ver05, Thm 8.1]. Further analytic development of this theory involves rigorous estimates and approximations when $\epsilon > 0$ is not zero (not pursued in this article).

4.2 Quasi-Equilibrium approximation (QE)

We retain the small parameter ϵ. The quasi-equilibrium approximation (QE) consists of two conditions. First, we assume that certain rate functions are small because they are proportional to ϵ; say, the last $r - l$ rate functions are of the form $w_{l+1} = \epsilon \tilde{w}_{l+1}$, $w_{l+2} = \epsilon \tilde{w}_{l+2}$ etc. This suggests that as $\epsilon \to 0$, the original PCTMC with r transitions is equivalent, approximately, to a PCTMC with l transitions. At the level of rate equations, this means setting $\epsilon = 0$ in

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \sum_{i=1}^{l} \mathbf{s}_i w_i + \epsilon \sum_{j=l+1}^{r} \mathbf{s}_j \tilde{w}_j \,. \tag{19}$$

This is not yet a reduction because we haven't identified (and eliminated) the fast variables. To (19), we associate a $n \times r$ stoichiometry matrix $\mathsf{S} = (\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_r)$, which is of full rank by assumption. However if we set $\epsilon = 0$, then the stoichiometry matrix loses $n - l$ columns and as a result of this, may become rank deficient. Hence the second condition of the QE approximation:

$$\gamma = \mathrm{rank}(\mathbf{s}_1, \ldots, \mathbf{s}_l) < n \,. \tag{20}$$

If $\gamma < n$, then $n - \gamma$ variables can be eliminated as 'fast variables' using the stoichiometry reduction, described in section 3.1. It is not difficult to realize that it is possible to make a linear transformation of variables, that brings (19) to (16). For example, suppose that $\mathbf{I}_1$ is a vector that nullifies the truncated stoichiometry matrix $(\mathbf{s}_1, \ldots, \mathbf{s}_l)$. Define the new variable $y_1 = \mathbf{I}_1 \cdot \mathbf{x}$ and multiply (19) by $\epsilon^{-1}\mathbf{I}_1$ on the left. We get

$$\frac{\mathrm{d}y_1}{\mathrm{d}\epsilon t} = \sum_{i=l+1}^{r} \big(\mathbf{I} \cdot \mathbf{s}_i\big) \tilde{w}_i(\mathbf{x}) = g_1(\mathbf{x})$$

This equation is in the form of a first, the slow variable equation in (16). One such equation per each null vector $\mathbf{I}$ can be constructed, and the entire model can be brought to the canonical form.

[3] which is the reason why this particular perturbation is called *singular* perturbation

[4] leaving aside the question of the appropriate initial condition

4.3 Quasi-Steady State approximation (QSS)

At this point, we drop the small parameter assumption. The argument proposed in the following is based on a 'heuristic' *steady state assumption*. The target model is given in the 'as is' form – with some numerical rate constants, possibly of the same magnitude, and no clue about the separation of scales. Based on intuition or some available heuristic argument, we could 'declare' certain transitions as fast transitions, which means the following. We assume that a particular fast transition, say $w_a(\mathbf{x})$, reaches a steady state (or 'burns out' to use more colorful language) well before the remaining transitions have enough time to change significantly. An indication for identifying the fast transition can be had, if we look at a derivative of the rate functions. In general, we get a linear function with respect to w_a,

$$\frac{\mathrm{d}w_a}{\mathrm{d}t} = \frac{w_a^\infty(\mathbf{x}) - w_a(\mathbf{x})}{\tau_a(\mathbf{x})} \tag{21}$$

where $w_a^\infty \geq 0$ and $\tau_a > 0$ are known functions of $\mathbf{x}$. We can interpret τ_i as a 'rate of a rate'; then the w_a with a smallest τ_a may well be the best choice for a fast transition. Then, we can interpret w_a^∞ as a steady state value of this transition. The steady state declaration is thus equivalent to a statement that there exists a special, confining surface called the QSS equation of state, defined by

$$w_a(\mathbf{x}) = w_a^\infty(\mathbf{x})\,. \tag{22}$$

Hence the QSS approximation consists of two parts: the 'declaration' of a set of fast transitions, and of the QSS equation of state (22). The latter plays a role of the slow manifold, analogous to (17). If our declaration was any good, then the true (simulated) value of $\sup_{t\in[t_1,t_2)} |w_a(\mathbf{x}(t)) - w_a^\infty(\mathbf{x}(t))|$ is reasonably small for most $t_1 < t_2$, justifying the 'promotion' of this estimate to the status of a small parameter ϵ, and an expansion $w_a(\mathbf{x}) = w_a^\infty(\mathbf{x}) + \epsilon \ldots$.

It is never a bad idea to emphasize two differences between the QSS and QE approximations. The notion of a QSS is based on 'declarations' which are, if we wish to put it bluntly, speculations based on arguments that are correlated with the level of user's expertise and may end up in good or bad approximations. This fact appears in stark contrast with the rigor of SPT and QE approximations. One could loosen the rigor of the QE by 'declaring' that some transitions can be discarded, thus making it closer to a QSS approximation. The second point to be emphasized is a different terminology. The is a minor difference in the jargon that often leads to a lot of confusion. In QE, we declare the transitions as *slow* based on the *small* parameter ϵ of a rate function, whereas in QSS, we declare the transition as *fast* based on the fast equilibration of a corresponding rate function. It may happen, that a slow transition (QE) is equivalent to fast equilibration (QSS). An example of this is presented in section 5.

4.4 Numerical methods

By now one must have realized that the possibility of a reduction is closely related to identifying a special embedded surface. In section 4.1, this surface was given to us by the formulation of equations(17). In section 4.2 we had to make a linear transformation of variables $\mathbf{x}$ to achieve the same result. In section 4.3 a judgement call was made to declare the best candidate for such a surface (22). In the most general case or when numerical accuracy is relevant, one may opt for a completely numerical strategy. There are several strategies available for tackling mean field problems, and none for stochastic models. In our opinion, three mean field techniques deserve special attention: the *Computational Singular Perturbation* (CSP), the *Invariant Low Dimensional Manifold* method (ILDM), and the *Fraser's method* (FM), each of which is described in section 6.

5 Reductions of the WSN model

We will skip the singular perturbation theory as it would take us too far into analysis. We will also not mention anything about purely numerical methods because our model is too simple for that. However,

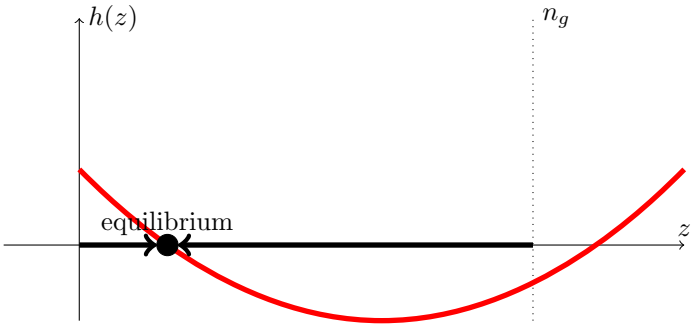


Figure 4: Illustration of the fast equilibration.

several interesting features of QE and QSS reductions can be illustrated using WSN as an example. To isolate these features, we identify three scenarios:

- Dispatch bottleneck scenario (DB);
- Gateway current bottleneck scenario (GCB);
- Hungry sensor scenario (HS);

To illustrate the QE approximation, we use the

Dispatch bottleneck scenario (DB) The dispatch transition τ_3 is the slow transition, because the *rate constant* k_3 (see (13)) is much smaller than the remaining ones: $k_3 \ll k_1, k_2$.

For simulations of the dispatch bottleneck, we use $k_1 = 1$, $k_2 = 1.38$, $k_3 = 0.1$, and $n_g = 0.75$.

5.1 Mean field QE approximation

We propose a new stoichiometric invariant $\mathbf{I} = (1, 1)$ which, we claim, brings the WSN to the canonical form (16). We also propose a second vector $(0, 1)$ in order to complete the coordinate transformation. Stacking these vectors as rows of the transformation matrix, we obtain the coordinate transformation,

$$\begin{pmatrix} y \\ z \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} x_1 + x_2 \\ x_2 \end{pmatrix} \tag{23}$$

Since $w_3 \sim k_3$ and k_3 is the small 'ϵ' parameter, we can recognize that dividing these equations by k_3, we obtain the (16). Multiplying the rate equations (10) by the transformation matrix, we get the equations for the new variables which we can write as

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -w_3 \tag{24}$$

$$\frac{\mathrm{d}z}{\mathrm{d}t} = w_1 - w_2 - w_3 \tag{25}$$

5.1.1 Fast variable equilibration

Let's focus now on the fast variable z. Substituting the coordinate transformations, we get

$$\frac{\mathrm{d}z}{\mathrm{d}t} = h(z; y) = n_g y - (y + n_g + a)z + z^2$$

where $a = (k_2 + k_3)/k_1$. In the extreme limit $k_3 = 0$, this equation can be solved independently from (24), because then y is simply a constant parameter. The idea of quasi-equilibrium is to assume that the same can be done, approximately, if k_3 is non-zero but small. This is the essence of the separation

of scales assumption: if the variable y varies a little during the interval of time that is sufficient for the variable z to equilibrate, then we may approximately freeze y for the same interval of time.

It is reasonable to expect that typical systems 'equilibrate' in some sense, meaning that their variables approach some stationary, 'equilibrium' values. Equilibration of the variable $z(t)$ towards the equilibrium z^* is illustrated in figure 4. The condition of equilibrium is $h(z^*; y) = 0$. Since $h(z; y)$ is a parabola with respect to z, there are two $h = 0$ solutions. In our case, these are two real roots: $z^* = \frac{1}{2}(y + n_g + a)\big(1 \pm \sqrt{1 - 4n_g y(y + n_g + a)^{-2}}\big)$. Looking at figure 4, all zs that have the parabola above the zero ordinate have positive velocity. Therefore, they move to the right, as indicated in the figure. Similarly, all zs with the parabola below the zero ordinate, move to the left (also indicated in the figure). So we have a situation where the lower root 'attracts' all zs from its immediate neighborhood, while the larger root 'repels' them. However, the larger root lies outside of the domain of z, so we don't need to worry about it. This argument and figure 4 prove that the smaller root (black dot) attracts from everywhere in the domain, so indeed, it is an equilibrium.

We can estimate the speed of attraction by approximating the parabola, close to the black dot, by a straight line. We get an exponential, $z(t) \sim z^* + A \exp(-k_{\text{fast}} t)$ where

$$k_{\text{fast}} = \left| \frac{\partial h(z; y)}{\partial z}(z^*) \right| = \sqrt{(y + a)^2 - 4y} \tag{26}$$

5.1.2 Slow variable equilibration

The relation between y and z, assuming that z is in local equilibrium as described in the preceding section, is called the quasi-equilibrium *equation of state*. Returning to the original variables (only for the smaller root of interest), the quasi-equilibrium equation of state is

$$\text{DB}: \quad x_2 = \frac{n_g k_1 x_1}{k_1 x_1 + k_2 + k_3} \tag{27}$$

The meaning of the equation of state is that it defines a curve (or more generally a surface) of slow equilibration the $(x_1,\ x_2)$ plane. If the initial conditions are not on this curve, then the fast equilibration 'adjusts' the trajectory, by quickly bringing it to the DB curve. The fast time scale is the time it takes for this 'adjustment'.

To describe the rate of slow equilibration (along the QE curve) we return to (24) where, on the right hand side we have $w_3 = k_3 z$. To eliminate z we have to be careful to use the lower root solution, $z^* = n_g y/(n_g + a + y) + o(y/(n_g + a + y))$. Approximating by the first term (for small y), we get

$$\frac{\mathrm{d}y}{\mathrm{d}t} = -\frac{k_3 k_1 n_g y}{k_1 n_g + k_2 + k_3} + O(y^2) \tag{28}$$

Also here, we find an exponential equilibration towards zero, $y(t) \sim B \exp(-k_{\text{slow}} t)$ with the rate

$$k_{\text{slow}} = \frac{k_3 k_1 n_g}{k_1 n_g + k_2 + k_3} \tag{29}$$

Since k_{slow} is proportional to a small parameter k_3, this equilibration is slower than the one described in the preceding section, as expected.

5.2 Stochastic QE approximation

Here we would like to translate the ideas of the mean field QE reduction, from section 5.1, to the master equation setting. The master equation of the WSN model is

$$\begin{aligned} \partial P(X_1, X_2) \;=\; & [k_1(X_1 + 1)(N_g - X_2 + 1)/N_s] \cdot P(X_1 + 1, X_2 - 1) \\ & + k_2(X_2 + 1) \cdot P(X_1 - 1, X_2 + 1) + k_3(X_2 + 1) \cdot P(X_1, X_2 + 1) \\ & - \big\{ k_1 X_1 (N_g - X_2)/N_s + (k_2 + k_3) X_2 \big\} \cdot P(X_1, X_2)\,. \end{aligned} \tag{30}$$

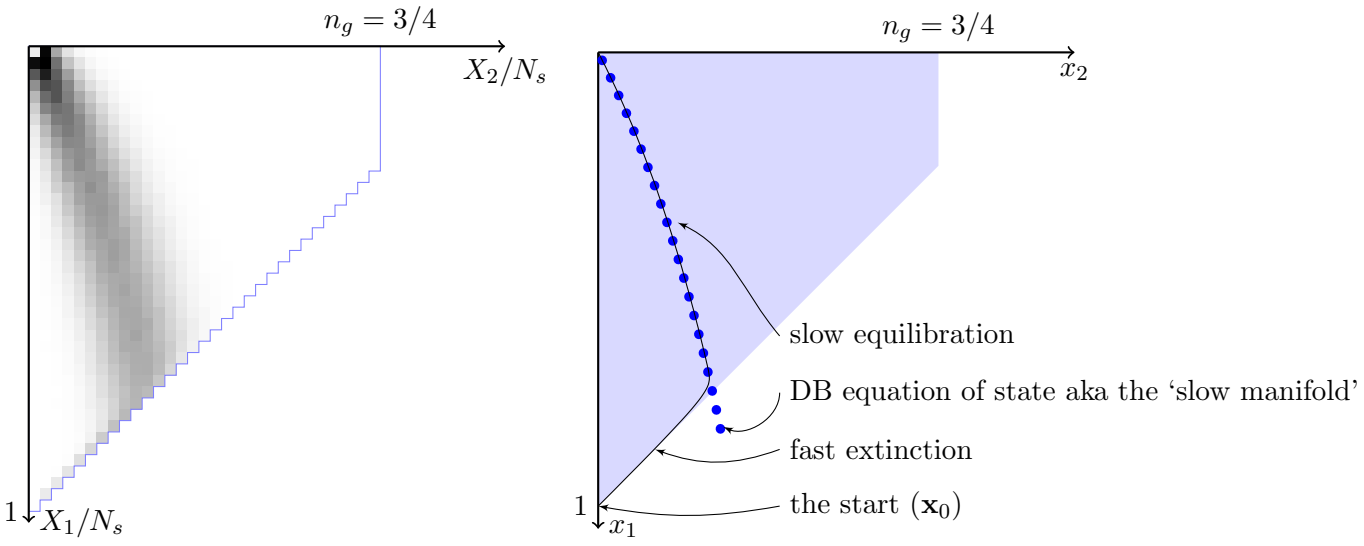


Figure 5: Left: Histogram of $P(\mathbf{X};t)$ with $P(\mathbf{X};0) = \delta(\mathbf{X} - \mathbf{D})$, $\mathbf{D} = (N_s, 0)$, obtained from 1000 stochastic trajectories. Right: a single mean field trajectory $\mathbf{x}(t)$ starting at $\mathbf{x}(0) = (1, 0)$. Main features of the trajectory and the QE equation of state (27) are also indicated.

5.2.1 Slow variable equilibration

An analogous coordinate transformation to (23) is $Y = X_1 + X_2$ and and $Z = X_2$. The probability of Y, X is $Q(Y, Z) = P(Y - Z, Z)$. Using this relation, we find, $P(X_1 + 1, X_2 - 1) = Q(Y, Z - 1)$, and $P(X_1 - 1, X_2 + 1) = Q(Y, Z + 1)$, and $P(X_1, X_2 + 1) = Q(Y + 1, Z + 1)$. Substituting into (30), we get

$$\begin{aligned} \partial_t Q(Y,Z) \;=\; & r_1(Y, Z-1)\cdot Q(Y, Z-1) - r_1(Y,Z)\cdot Q(Y,Z) \\ & r_2(Y, Z+1)\cdot Q(Y, Z+1) - r_2(Y,Z)\cdot Q(Y,Z) \\ & k_3\big\{(Z+1)\cdot Q(Y+1, Z+1) - Z\cdot Q(Y,Z)\big\}\,. \end{aligned} \tag{31}$$

The new rate functions r_1 and r_2 are: $r_1(Y, Z) = k_1(Y - Z)(N_g - Z)/N_s$, and $r_2(Y, Z) = k_2 Z$.

Summing over all possible values of a variable in a joint probability, eliminates that variable; so $Q(Y) = \sum_Z Q(Y, Z)$ is a probability of Y. Performing this summation in (31) eliminates the fast variable, and we get a master equation for the probability of a slow variable

$$\partial Q(Y;t) = \big[k_3 \mathrm{E}(Z|Y+1;t)\big]\cdot Q(Y+1;t) - \big[k_3\mathrm{E}(Z|Y;t)\big]\cdot Q(Y;t) \tag{32}$$

where

$$\mathrm{E}(Z|Y;t) = \sum_Z Z Q_c(Z|Y;t)\,.$$

is the mean of the fast variable, conditional on Y being held fixed, and $Q_c(Z|Y)$ is the 'Z conditional on Y' probability, found from a relation

$$Q(Y, Z; t) = Q(Y; t)\cdot Q_c(Z|Y;t)\,.$$

This equation describes a 'death process' with a transition

$$W(Y|Y') = \delta_{Y,Y'-1}\big[k_3\mathrm{E}(Z|Y')\big]$$

which is slow, indeed, because the rate is proportional to a small parameter k_3.

A couple of technical details are in place. An important property of r_1 and r_2 was exploited in arriving at (32). Note that r_1 and r_2 depend on Y without any ± 1 shifts. Upon summing (31), all terms containing r_1 and r_2, cancel out exactly[5]. If $\mathrm{E}(Z|Y)$ can be approximated by a constant, then 32 is solvable (see [Gar85, Chap I] and [Gil77]).

[5] In general this is an approximation because there may be non-zero contributions from the boundaries.

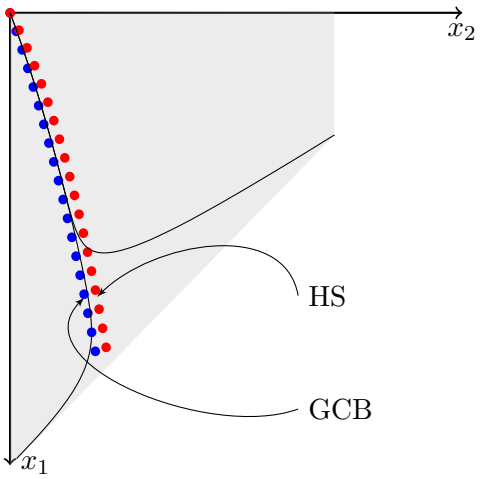


Figure 6: The mean field QSS reduction. Two sample trajectories stick together, therefore the model is reducible. The two equations of state, corresponding to GCB and HS scenarios are also shown.

5.2.2 Fast variable equilibration

The fast equilibration is described by the conditional probability $Q_c(Z|Y)$. The most important assumption of the stochastic QE approximation is the Markovian property of Q_c. With this assumption, Z becomes completely decoupled from the slow dynamics described in the preceding section. Neglecting the k_3 term, we get a 'birth-death' master equation for the fast variable,

$$\begin{aligned} \partial Q_c(Z|Y;t) &= \sum_{Z'} \Big[t^+(Z|Z') + t^-(Z|Z') \Big] \cdot Q_c(Z'|Y;t) \\ &\quad - \sum_{Z'} \Big[t^+(Z'|Z) + t^-(Z'|Z) \Big] \cdot Q_c(Z|Y;t) \end{aligned} \tag{33}$$

with the birth and death rates, respectively,

$$t^+(Z|Z') = \delta_{Z,Z'+1} r_1(Y,Z')\,, \qquad t^-(Z|Z') = \delta_{Z,Z'-1} r_2(Y,Z')\,.$$

Also here, the fact that Ys have no shifts in Q_c in (31) was an important technical detail of derivation.

5.3 Mean field QSS approximations

The general differences between the quasi-steady state and quasi-equilibrium approximations, and the main procedures of the QSS reduction were outlined in section 4. We have two options of declaring the fast transition here. We discuss both of them using the 'gateway current bottleneck' and 'hungry sensor' scenarios, below. Both scenarios are illustrated by a simulation with parameters $k_1 = 1$, $k_2 = 1.38$, $k_3 = 0.85$, $n_g = 0.75$, and two initial conditions, $\mathbf{x}(0) = (1, 0)$ and $\mathbf{x}(0) = (n_g, 1 - n_g)$. The results are shown in figure 6. By design, none of the ks could be identified as small or dominant. And yet, one can clearly see in the figure, that the QSS reduction is as effective as was the QE reduction, where the separation of scales was guaranteed by the smallness of k_3. Moreover, note that GCS and HS scenarios are not equivalent. GCS emerges clearly as a winner here. HS is less accurate, although it captures the overall trend.

Gateway current bottleneck scenario (GCB) Declaring either $w_2 = k_2 x_2$ or $w_3 = k_3 x_2$ as fast transitions, and proceeding as outlined in section 4, the QSS equation of state (the steady state condition) for the GCB scenario is found as $w_1 = w_2 + w_3$. Solving for the concentrations, we find

$$\text{GCB} \;:\quad x_2 = \frac{k_1 n_g x_1}{k_1 x_1 + k_2 + k_3}\,. \tag{34}$$

Note that it is the same equation as (27). It is an interesting result: the gateway current bottleneck scenario, i.e. assuming that w_2 and w_3 are fast transitions, is equivalent to the slow dispatch scenario. This result may strike as an odd one, because it would seem that the slow dispatch idea runs contrary to the assumption of a fast w_3. Nevertheless, the two are consistent: even if w_3 is considered as fast, it may result in a small steady current. Note also that in the QSS scenario, we didn't assume that k_3 is small. Rather, we can verify a-posteriori, that the approximation holds better if n_g is not too large, which in turn explains the title of this scenario.

Hungry sensor scenario (HS) If we declare that $w_1 = k_1 x_1 (n_g - x_2)$ is the fast transition, we are implying that sensors are somehow 'eager' to establish a contact with gateways, or we can think of them as being 'hungry'. The condition of a steady current is $0 = \dot{w}_1 = k_1(n_g - x_2)(w_1 - w_2) + k_1 x_1 (w_1 - w_2 - w_3)$. Rearranging terms, it is cast as a quadratic equation for x_2,

$$\text{HS} \;:\quad x_2^2(x_1+\kappa_2) - x_2\Big[(x_1+\kappa_2)(x_1+n_g) + x_1(\kappa_3+n_g)\Big] + n_g x_1 (x_1+n_g) = 0\,, \tag{35}$$

where $\kappa_2 = k_2/k_1$ and $\kappa_3 = k_3/k_1$. Note that by being a quadratic equation, it has two solutions, as in the QE case. Likewise, the + branch is discarded as being outside the domain of definition.

5.4 Stochastic QSS approximation

We have seen that the GCB scenario is equivalent to the DB scenario, therefore the stochastic approximation is the same. General methods for the stochastic QSS reductions do not exist yet.

6 Further reading: mean field model reductions

6.1 Singular perturbation theory

The analytic singular perturbation theory develops the picture of QE and QSS in detail, especially in its important aspects of convergence. This theory is complicated, and does not lead to any more reduction than what is allowed by Tikhonov theorem. However, we give a brief outline for completeness. For more details, we recommend an excellent book by Verhulst [Ver05].

A differential equation of the form

$$F(t, x, x', \dots; \epsilon) := F_0(t, x, x', \dots) + \epsilon F_1(t, x, x', \dots, \epsilon) = 0,$$

or a system of differential equations of the form

$$\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x}, \epsilon) := \mathbf{g}_0(\mathbf{x}) + \epsilon \mathbf{g}_1(\mathbf{x}, \epsilon),$$

where ϵ is a small parameter controlling the size of the perturbation, F_1 or $\mathbf{g}_1$, is called *singular* if one the following (equivalent) statements is true

- Differential equation $F_0 = 0$ is of lower order than $F_0 + \epsilon F_1 = 0$. This situation occurs when, for example, F_1 contains higher derivatives than F_0;
- The Jacobian matrix $\mathsf{J}(\mathbf{x}; \epsilon) = \frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\mathbf{x}; \epsilon)$ is singular (some eigenvalues are 0 or ∞) if $\epsilon = 0$, and non-singular otherwise;

In both scenarios of the WSN model a small parameter ϵ multiplies the first of the two equations, namely $y_1' = \mathcal{O}(\epsilon)$. If we set $\epsilon = 0$, the first row of the Jacobian matrix is full of zeros, therefore at least one of its eigenvalue is equal to zero; hence both WSN scenarios are singular. This holds as a general rule: QE and QSS approximations are called for whenever the analytic problem contains a singular perturbation.

Tikhonov theorem (1958) is often quoted as the analytic foundation of the QSS approximation. We present this theorem following [Ver05, Thm 8.1]. Consider the initial value problem

$$\begin{aligned} \frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} &= \mathbf{f}(\mathbf{x},\mathbf{y},t) + \epsilon\cdot\ldots, \quad x \in D \subset \mathbb{R}^n, t \geq 0, \\ \epsilon\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}t} &= \mathbf{g}(\mathbf{x},\mathbf{y},t) + \epsilon\cdot\ldots \quad y \in G \subset \mathbb{R}^m. \end{aligned} \tag{36}$$

Here $\mathbf{f}$ and $\mathbf{g}$ are sufficiently smooth vector functions in $\mathbf{x}$, $\mathbf{y}$, and t; the dots represent (smooth) higher-order terms in ϵ. Assume that,

1. a unique solution of the initial value problem exists
2. a unique solution of the reduced initial value problem exists; the reduced problem:

$$\begin{aligned} \dot{\mathbf{x}} = \mathbf{f}(\mathbf{x},\mathbf{y},t), \quad \mathbf{x}(0) = \mathbf{x}_0, \\ 0 = \mathbf{g}(\mathbf{x},\mathbf{y},t), \end{aligned}$$

   with solutions $\overline{\mathbf{x}}(t)$, $\overline{\mathbf{y}}(t)$.

3. Suppose that $0 = \mathbf{g}(\mathbf{x},\mathbf{y},t)$ is solved by $\mathbf{y} = \boldsymbol{\phi}(\mathbf{x},t)$ where $\boldsymbol{\phi}(\mathbf{x},t)$ is a continuous function and an isolated root. Also suppose that $\mathbf{y} = \boldsymbol{\phi}(\mathbf{x},t)$ is an asymptotically stable solution of the equation

$$\frac{\mathrm{d}\mathbf{y}}{\mathrm{d}\tau} = \mathbf{g}(\mathbf{x},\mathbf{y},t)$$

   that is uniform in the parameters $\mathbf{x} \in D$ and $t \in \mathbb{R}^+$.

4. $\mathbf{y}(0)$ is contained in an interior subset of the domain of attraction of $\overline{\mathbf{y}} = \boldsymbol{\phi}(\mathbf{x},t)$ in the case of the parameter values $\mathbf{x} = \mathbf{x}(0)$, $t = 0$.

Then we have

$$\begin{aligned} \lim_{\epsilon\to 0} \mathbf{x}_\epsilon(t) &= \overline{\mathbf{x}}(t), \qquad 0 \leq t \leq L, \\ \lim_{\epsilon\to 0} \mathbf{y}_\epsilon(t) &= \overline{\mathbf{y}}(t), \qquad 0 < d \leq t \leq L \end{aligned} \tag{37}$$

with d and L constants, independent of ϵ.

It is easy to show that both scenarios of the WSN model, given by (**??**)–(**??**) and (**??**)–(**??**) are in the Tikhonov form (36) if the time variable is scaled according to the formula $t_{\text{Tikhonov}} = \epsilon t_{\text{WSN}}$. The form (36) is more convenient for the statement of the theorem because it renders the initial transient infinitely fast in the limit $\epsilon = 0$.

In assumption (iii), t and $\mathbf{x}$ are parameters and not variables. The idea is that during the fast motion of the variable $\mathbf{y}$, the small variations of these parameters are negligible as long as the stability requirement holds.

Invariant manifold is an important concept which we briefly mention here. By the very meaning of invariance, any data, that is located on an invariant manifold initially, remains on the same manifold forever. There are many different flavors of invariant manifolds. A special role is attributed to a manifold with a property of being "attracting" or, equivalently "stable". The attracting property means that all trajectories within its "basin of attraction", approach the manifold (distance tends to zero), forward in time.

It is easy to check that the velocity field on the slow manifold is not parallel to the slow manifold, $\mathbf{y} = \boldsymbol{\phi}(\mathbf{x},t)$, discussed in section 6.1, whenever $\epsilon > 0$. Hence, the slow manifold is *not* an invariant manifold. Generic long-time behavior of a typical trajectory, in relation to the slow manifold, can be described by imprecise statements like "are approximately parallel to each other", "stay not far

from one another", "have locally similar speeds," but it is not correct to assume the slow manifold equation as a mathematical identity. For more details, the perturbation theory must be applied, which can be unfeasible for complex problems. Practitioners have long gone the numerical way, and there are very nice, geometry-inspired methodologies on the market. Virtually any advanced method builds on the concept of a slow manifold. We recommend, as a first reading, [Fra88], who explains beautifully the relationship between the slow manifold, the invariant (center) manifold, the QSS and QE approximations using mostly pictures. For the analytic theory an excellent book by Verhulst is recommended [Ver05].

For a discussion of Tikhonov's theorem in the context of SDEs, see an article by Berglund and Gentz [BG03].

Geometric perturbation theory We only mention that a theory by Fenichel [Fen79] shows that, under suitable conditions, there is an invariant manifold with a special meaning, called the center manifold, which is a very important object. It is located "not far" from the slow manifold. Advanced numerical methods mentioned below, attempt, directly or otherwise, to compute and quantify the center manifold.

6.2 Fraser's method

Fraser & co. have proposed a methodology, based on a very interesting geometric structure, discovered by the authors[RF90, NF89, Fra88]. They consider a system of n mean-field equations of the form $\dot{\mathbf{x}} = \mathbf{g}(\mathbf{x}, \epsilon)$ and use the concept of a *nullcline*. A nullcline $\mathcal{N}_i$ is an embedded $n-1$ dimensional surface, defined by the equation $g_i(\mathbf{x}, \epsilon) = 0$. A special geometric structure is present, authors claim, in all chemical reaction networks which, as authors put it, can be reduced to planar flows (with one-dimensional slow dynamics) [RF90]. To my understanding, this assumption is equivalent to the following statements:

- the model is degenerate in the limit $\epsilon = 0$ (an additional invariant exists);
- there is a global isolated equilibrium $\mathbf{x}_0$;
- the degenerate limit has a local equilibrium $\mathbf{x}_1$. The equilibrium forms a "one dimensional line" [RF90] as a consequence of degeneracy;
- some intuitive argument can be used to identify a one-dimensional slow variable.

Authors argue that the intersection $\cap\mathcal{N}_i$, splits the entire n-dimensional space in such a way that a compact disjoint subspace can be identified which, topologically, is a tube. The cross-section of this tube is a $n-1$-simplex, determined by the intersection. The tube should possess the following properties: 1) it contains $\mathbf{x}_0$ and $\mathbf{x}_1$ when $\epsilon > 0$ is small; 2) its boundary is, locally, a dynamical no-return surface (although this detail is not mentioned by the authors). The no-return property means that all trajectories, entering the tube, stay inside the tube. Hence, authors claim, the tube should contain the so-called *center manifold* (provided that such a manifold exists in a given system) and they propose a simple idea how to find it. They propose a variational equation, which they call a *functional equation*, that is simply the equation of the trajectory obtained from the system of differential equations by eliminating the slow variable. Suppose for simplicity that the slow variable is $s = x_n$. Then, in the $n-1$ dimensional vector space, $\mathbf{y} \in \mathbb{R}^{n-1}$, we may define this surface as a one dimensional curve $\mathbf{y}(s)$, as a solution of

$$\begin{aligned} &g_1(y_1, \ldots, y_{n-1}, s, \epsilon) - g_n(y_1, \ldots, y_{n-1}, s, \epsilon)\frac{\mathrm{d}y_1}{\mathrm{d}s} = 0\,, \\ &\vdots \\ &g_{n-1}(y_1, \ldots, y_{n-1}, s, \epsilon) - g_n(y_1, \ldots, y_{n-1}, s, \epsilon)\frac{\mathrm{d}y_{n-1}}{\mathrm{d}s} = 0\,, \end{aligned} \tag{38}$$

and authors propose to solve this equation discretizing $\mathbf{y}(s)$ in the domain of interest, setting for the initial condition $\mathbf{y}_0(s)$ one of the intersections between two arbitrarily chosen nullclines, improving the solution iteratively, using Newton's iterations.

Extensions to this method to the master equation has been studied [RZ04].

6.3 Intrinsic Low-Dimensional Manifold (ILDM)

The method proposed by Maas & Pope [MP92b, MP92a] is also a geometric method in the sense that it wants to find a certain *invariant manifold*, embedded in the phase space. We fix the dimension of the surface, k, to be an arbitrary integer, $0 < k < n$ and ask for a k-dimensional surface S, that best approximates the slow dynamics, with the expectation that any reaction trajectory, after a quick transient, should land on this surface. We would like to find this surface as a parameterization $S(c_1, \ldots, c_k) = 0$ by a k-dimensional grid with coordinates c_1, ..., c_k. This then corresponds to a parametrically defined coordinates $\mathbf{x} = \mathbf{f}(c_1, \ldots, c_k)$ for some $\mathbf{f}$.

The (local) $n \times n$ Jacobian matrix is defined by

$$\mathsf{J}(\mathbf{x}, \epsilon) = \frac{\partial \mathbf{g}}{\partial \mathbf{x}}(\mathbf{x}, \epsilon)\,. \tag{39}$$

It is characterized by n eigenvalues $\lambda_i(\mathbf{x}, \epsilon)$, and n eigenvectors $\mathbf{v}_i(\mathbf{x}, \epsilon)$. The adjective 'local' refers to the fact that all of these quantities are defined with respect to a the reference point $\mathbf{x}$. We may assume without loss of generality, that all eigenvectors and eigenvalues are ordered in non-decreasing order by the real part of the corresponding eigenvalue λ_i, i.e. $\Re\lambda_i \leq \Re\lambda_{i+1}$.

The main hypothesis is that a trajectory $\mathbf{x}(t)$ at a point $\mathbf{x}$ accelerates most in the direction of the eigenvector with the largest real eigenvalue, that is in the direction of $\mathbf{v}_n$; a bit less in the direction of $\mathbf{v}_{n-1}$, and so on, down to the slowest acceleration in the direction of $\mathbf{v}_1$[6]. The idea is to define the ILDM surface locally, by requiring that the vector field on the surface $\tilde{\mathbf{g}}(c_1, \ldots, c_k) = \mathbf{g}(\mathbf{f}(c_1, \ldots, c_k))$ is normal to $n - k$ fastest eigenvectors of J. This results in a system of equations

$$\begin{aligned}
&\mathbf{v}_{k+1}^{\intercal} \cdot \tilde{\mathbf{g}}(c_1, \ldots, c_k) = 0 \\
&\vdots \\
&\mathbf{v}_{n}^{\intercal} \cdot \tilde{\mathbf{g}}(c_1, \ldots, c_k) = 0 \\
&A_1(c_1, \ldots, c_k) = 0 \\
&\vdots \\
&A_k(c_1, \ldots, c_k) = 0\,,
\end{aligned} \tag{40}$$

which is augmented by k additional constraints $A_1 \ldots A_k$ to make the solution unique. The higher is the k, the shorter is the transient time but on the flip side, a bigger grid is needed to tabulate a surface with larger k. Most common numbers for k used in the literature are 1 or 2. In case of $k = 1$, the ILDM surface should be the same as found by the Fraser method, given the same boundary conditions.

We mention two important terms used in numerical methods: 'stiffness' of differential equations, and 'condition number' of matrices. We claim that, multiple time scales, stiff differential equations, and badly conditioned matrices, are different facets of the same reducibility phenomenon; a stiff differential equation implies a badly conditioned Jacobian matrix, which implies multiple time scales and vice versa. In particular, to describe the characteristic features of Jacobians of reaction networks, we quote Maas, who describes a model of methane-air combustion [MP92b, p. 248]:

If we look at real chemical systems, sample calculations ¡...¿ show the following: Some eigenvalues are zero ¡...¿. Almost all other eigenvalues are negative with values from typically -10^{-2} *to* $-10^{7}\mathrm{s}^{-1}$.

[6] This hypothesis is to be attributed to Lam and Goussis who call it "speed ranking" [LG94]. It is implicit in the work by Maas and Pope.

Only very few eigenvalues are positive (in the sample calculations at most one eigenvalue was positive). At the equilibrium point all eigenvalues are zero or negative. ¡...¿ due to poor conditioning (roughly speaking,"nearly linearly-dependent eigenvectors"), numerical difficulties could arise.

To avoid the issues with badly conditioned matrices Maas proposes to replace the eigenvectors $\mathbf{v}$ with vectors of the Schur basis $\mathbf{q}$ which are insensitive to conditioning numbers. in [MP92b] the following computational strategy is proposed

- Start with an initial guess of the state variables $\mathbf{x} = \mathbf{f}(c_1, \ldots, c_k)$ that fulfills the algebraic equations $A_1 = 0, \ldots, A_k = 0$.
- For i=0,1,2, ..., compute the Jacobian J and the Real Schur decomposition $\mathsf{Q}^\intercal \mathsf{J} \mathsf{Q} = \mathsf{T}$, referring to the old variable $\mathbf{x}^{\text{old}}$. Solve

$$\mathsf{Q}_L^\intercal \frac{\mathrm{d}\mathbf{x}^{\text{new}}}{\mathrm{d}t} = \mathsf{Q}_L^\intercal \tilde{\mathbf{g}} \tag{41}$$

$$\mathbf{A} = 0 \tag{42}$$

  until a steady state is obtained.
- If $||\mathbf{x}^{\text{new}} - \mathbf{x}^{\text{old}}|| < \mu$ accept solution, set $\mathbf{x} = \mathbf{x}^{\text{new}}$ and leave the loop.

Maas & Pope propose to approximate the solution of the differential-algebraic equation (41)–(42) by the extrapolation method.

ILDM has become quite popular in chemistry, and various special case studies and extensions exist. For example ILDM can be coupled with transport and diffusion equations or temperature, to account for non-ideally mixed reactions [BM07]. Coupling with spatial degrees require other techniques to be used to handle the added complexity. In this context, Principal component analysis (PCA) and multivariate adaptive spline regression (MARS) have been used in conjunction with ILDM[YPC13]. Overview of other applications and recent developments can be found here [KSR+11].

6.4 Computational Singular Perturbation (CSP)

Computational Singular Perturbation by Lam & Goussis [LG94, Lam93, LG91, LG89] is a numerical technique to do practical computations with singularly perturbed systems. The final version of the method [LG94] is widely cited within the numerical combustion community, but is also used outside of it. For an example of its use outside of chemistry, circadian rhythms in Drosophila, see [GN06]. Authors consider CSP as, simply, an integrator for stiff differential equations (by repeatedly pointing out that their method eliminates the 'stiffness' feature). Authors skillfully manage the subtleties of the special structure of equations of reaction kinetics to craft a theoretically sound strategy. Their superior craftsmanship shows also where it is mostly useful to the reader: presenting intuitive interpretation of the intermediate steps, and making analogies with simpler reaction systems and analytic computations. This method is our favorite pick of the section.

It is based on the availability of the Jacobian matrix, and requires the ability to compute some of its eigenvectors, and eigenvalues. Dense matrix inversion is a prerequisite as well, but hopefully inversion does not have to be performed very often. The original exposition is quite heuristic; for a more formal presentation and error analysis [ZKK04b, ZKK04a] could be consulted as well. See [Lam13] for a non-numerical application. Although conceived for the numerical combustion problems, it seems to be applicable to general reaction network models. Of possible interest is [WQH13], where software is presented, written for a special case study of CSP reduction. It is possible that *ChemSuite* (?) implements CSP [references needed].

The starting point is a chemical reaction network of N nonlinear ODEs

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{g}(\mathbf{x}) = \sum_{i=1}^{R} \mathbf{s}_i r^i(\mathbf{x}) \tag{43}$$

where in the last expression, $\mathbf{s}_r$ are constant stoichiometric vectors, and r^i are rate functions. Since $\mathbf{g}$ can be expanded in a basis of "trial" column basis vectors $\mathbf{a}_i$, and row basis vectors $\mathbf{b}_i$. The two sets are adjoint by the orthogonality $\mathbf{b}^i \cdot \mathbf{a}_j = \delta^i_j$. The velocity field is expanded as follows

$$\mathbf{g} = \sum_{i=1}^{N} \mathbf{a}_i f^i, \qquad f^i = \mathbf{b}^i \cdot \mathbf{g} \tag{44}$$

The expansion coefficients f^i are called "modes". Mode dynamics is governed by

$$\frac{\mathrm{d}f^i}{\mathrm{d}t} = \sum_{j=1}^{N} \Lambda^i_j f^j \tag{45}$$

Two equivalent representations of the matrix (Λ^i_j) can be found by differentiating (44) with respect to time;

$$\Lambda^i_j = \big(\frac{\mathrm{d}\mathbf{b}^i}{\mathrm{d}t} + \mathbf{b}^i \cdot \mathsf{J}\big) \cdot \mathbf{a}_j = \mathbf{b}^i \cdot \big(- \frac{\mathrm{d}\mathbf{a}_j}{\mathrm{d}t} + \mathsf{J} \cdot \mathbf{a}_j\big) \tag{46}$$

Consider the ideal picture of a linear model: if $\mathbf{b}^i$ and $\mathbf{a}_i$ are left and right eigenvectors, respectively, of J, then $\Lambda^i_j = \mathrm{diag}(\lambda_1, \ldots, \lambda_N)$ and so mode dynamics is decoupled. Lam and Goussis propose a to proceed as follows. Le's suppose we have information that Λ^i_j is diagonal initially, so that its eigenvalues are ordered in decreasing order.

A certain threshold time scale τ is defined by the user. All time scales, smaller than τ are declared to be "fast", equivalently, all eigenvalues of Λ that are greater than $1/\tau$ are declared to be "large". Suppose there are M such eigenvalues, clearly $M < N$. If initially the eigenvalues are arranged in decreasing order, then the upper $M \times M$ diagonal block contains the Jordan form of the "fast" subspace, the bottom $(N - M) \times (N - M)$ diagonal block contains the Jordan form of the "slow" subspace. Ideally, the fast and slow subspaces are decoupled, so the upper-right and lower-left corners of Λ are zero-matrices. We show the structure of Λ in (47)

$$\Lambda = \left(\begin{array}{c|c} \tilde{\Lambda} & B \\ \hline C & D \end{array}\right), \qquad \mathbf{f} = \left(\begin{array}{c} \tilde{f} \\ \hline \overline{f} \end{array}\right) \tag{47}$$

In real situations B and C are not exactly zero. We have

$$\frac{\mathrm{d}\tilde{f}}{\mathrm{d}t} = \tilde{\Lambda}\tilde{f} + B\overline{f} = \tilde{\Lambda}\big(\tilde{f} - \tilde{f}_\infty\big) \tag{48}$$

where $\tilde{f}_\infty = -\tilde{\Lambda}^{-1}B\overline{f}$. Authors claim that $\tilde{f} - \tilde{f}_\infty$ is a "smaller" quantity that $\tilde{f}$. If we consider a new variable $\tilde{f}_{\mathrm{new}} = \tilde{f} - \tilde{f}_\infty$ then, we have

$$\frac{\mathrm{d}\tilde{f}^{\mathrm{new}}}{\mathrm{d}t} = \tilde{\Lambda}\big(\tilde{f}^{\mathrm{new}} - \tilde{f}^{\mathrm{new}}_\infty\big) \tag{49}$$

where $\tilde{f}^{\mathrm{new}}_\infty = \tilde{\Lambda}^{-1}\frac{\mathrm{d}\tilde{f}_\infty}{\mathrm{d}t}$. Because $\tilde{\Lambda}$ is "big", $\tilde{f}^{\mathrm{new}}_\infty$ is supposed to be "smaller" (in appropriate limit) than $\tilde{f}_\infty$. And so on... In the last version of the method, authors propose to use the following iterative "refinement" scheme, acting on the basis vectors:

- First step:

$$\tilde{\mathbf{b}}^{\mathrm{new}} = \tilde{\Lambda}^{-1}\big(\frac{\mathrm{d}\tilde{\mathbf{b}}}{\mathrm{d}t} + \widetilde{\mathbf{b} \cdot \mathsf{J}}\big), \quad \tilde{\mathbf{a}}_{\mathrm{new}} = \tilde{\mathbf{a}} \tag{50}$$

- Second step:

$$\tilde{\mathbf{b}}^{\text{new}} = \tilde{\mathbf{b}}, \quad \tilde{\mathbf{a}}^{\text{new}} = \big(- \frac{\mathrm{d}\tilde{\mathbf{a}}}{\mathrm{d}t} + \widetilde{\mathsf{J} \cdot \mathbf{a}} \big) \tilde{\Lambda}^{-1} \tag{51}$$

Here, $\mathbf{a}$ is a matrix whose columns are column vectors $\mathbf{a}_i$, $\mathbf{b}$ is a matrix, whose rows are row-vectors $\mathbf{b}^j$. Authors claim that these two steps may be performed singly, or in tandem in any order, or recursively any number of times – provided the most current $\tilde{\mathsf{T}}$ is used always. Authors acknowledge that this iteration scheme is an extension of the *Mises Power Method* to compute the eigenvectors, corresponding to the largest eigenvalues of a matrix. The time derivatives of $\mathbf{a}$ and $\mathbf{b}$ can be computed using their expressions in terms of $\mathbf{x}$ and the time derivative $\dot{\mathbf{x}} = \mathbf{g}$. The iterations maintain orthogonality of $\mathbf{a}_i$ and $\mathbf{b}^i$. The first step "depresses" the block B, while the second step "depresses" the block C.

After the modes are decoupled to a satisfactory accuracy, we stop. This procedure provides two interesting results. First, it gives a new system of differential equations in which the fast and slow degrees are decoupled, and the equation for the slow degrees only, can be approximated as

$$\frac{\mathrm{d}\mathbf{x}}{\mathrm{d}t} = \mathbf{g}(\mathbf{x}) \approx \sum_{J=M+1}^{N} \mathbf{a}_J^{\text{new}} f_{\text{new}}^J \tag{52}$$

The second result is M approximate equation of state,

$$f_{\text{new}}^m = \mathbf{b}_{\text{new}}^m \cdot \mathbf{g}(\mathbf{x}) \approx 0 \tag{53}$$

6.5 Polynomial reaction rates

Gorban, Karlin & associates have developed the *Method of Invariant Manifolds* (MIM) [GK92, GK03]. The fundamental additional assumption about the rate functions is

$$r_i(\mathbf{x}) = k_i^+ \prod_{j=1}^{d} x_j^{\alpha_{ij}} - k_i^- \prod_{j=1}^{d} x_j^{\beta_{ij}} \tag{54}$$

The polynomial rate functions (54) is indeed to most common assumption in chemical reaction modeling, but their relevance for our project needs to be verified.

With (54) in place, chemical reaction kinetics could be reformulated using certain "potentials", like the Lyapunov function, and quantities like temperature and other "global" quantities. Gorban and Kaitlin advocate a thesis whose gist is as follows. Chemical reaction dynamics is fuzzier, and fundamentally more inaccessible to verification, than one would wish. As a result, creative reaction paths, out-of-the-hat reaction rates and similar "frivolities" are commonplace in the business of chemical reaction modeling. An alternative theory would be superior to a detailed, but too "creative" reaction kinetics, if it could somehow manage to operate in terms of fewer quantities and those quantities could be measured unambiguously from observations (thermodynamic potentials are candidates for such quantities)[GK03]. It is a kind of reduction, in the sense that authors advocate a reductionist view of chemical reaction networks. It seems to me that authors to fall short of providing concrete strategies of how such a theory could be implemented in practice. Instead they fall back to reformulating the reaction rate kinetics using thermodynamic potentials, which ends up being an alternative formulation of the same theory. Be it as it may, the idea is an interesting one.

Polynomial rates are ubiquitous and quite a few researchers have dabbled with them. Already Fenichel [Fen79] has described a method of computing the slow manifold with polynomial rates (see also [NW11].) Nicolini et al. mention normal forms in the polynomial bases [NF13a, NF13b]. See also [vdSRJ13] for a recent article about polynomial rates.

6.6 Sensitivity analysis

Sensitivity analysis is a very old method to gain some information about the system [REFERENCES?]. It is essentially linear analysis, where derivatives of product concentrations with respect to rates are assembled into a matrix, and this matrix is played with. It seems like the methods discussed up to now are superior to sensitivity analysis. Model reduction methods, based on dynamic sensitivities from the impulse parametric sensitivity analysis (iPSA) and the Green's function matrix (GFM) analysis have been developed by [PMKTG13]

6.7 A mix combinations and extensions of the previous methods

Geometric interpretation has been pursued by [NF13a, NF13b] where QSSA ideas of Fraser, Maas and other are further developed. In particular Nicolini & associates introduce a concept of "attractiveness" of certain slow manifolds. Acknowledging the relevance of slow manifolds, Bykov et. al. focus on the details of fast manifolds [BG13]. Hybridization of the QSSA-ILDM and CSP methods is proposed in [ZAI13]. A very interesting discussion of the fast-slow manifolds and the associated degeneracy problems is presented by Goussis in [Gou13] where the van der Pol oscillator is discussed, whose equilibrium state is oscillatory rather than a stagnation point. Lebedev & co. have developed an advanced tool for reduction of detailed kinetic mechanisms with a 'minimal human effort' [LOC$^+$12]. The tool includes 10 reduction and 2 analysis methods which are based on the results of zero-dimensional modeling. The methods can be combined and applied in sequence. The reduction tool has been implemented as a part the *Chemical Workbench* computational package and has been tested for a number of large kinetic mechanisms of gas-phase processes. Equivalence of the leading order asymptotic analysis and the geometric approximations (QE, QSS) is discussed in [Gou12]. Chiavazzo & Karlin claim to have developed a fully adaptive methodology for reducing the complexity of large dissipative systems. Accurate reduced description is achieved, by construction of a hierarchy of slow invariant manifolds, with an embarrassingly simple implementation in any dimension [CK11]. Possibly interesting ideas concerning algorithmic implementation are [BM11]. A very efficient computational technique, based on the ILDM method and in situ adaptive tabulation (ISAT) of the accessed region of the composition space is presented by Pope [Pop97]. Test results show excellent control of errors; and a speed-up factor of about 1000 compared to the direct approach of numerically integrating the reaction equations. This is a very popular computational technique in combustion community.

7 Further reading: stochastic model reduction

The oldest reduction technique based on the separation of scales is the so-called *adiabatic elimination of fast variables*, applied to the Fokker-Planck equation (FPE). A classic example is the Smoluchowski equation for the Brownian motion of a light particle[Gar85, Chap 6]. The ambient space of a particle, the probability $P(x, v; t)$, and the full FPE is six dimensional (three space coordinates x, and three velocity coordinates v). However, its Langevin equation ($\dot{x} = v$, $\mu\dot{v} = -\lambda v + F(x) + \gamma\xi(t)$) is in a Tikhonov form with respect to the small mass parameter μ so, setting $\mu\dot{v} \to 0$ we can eliminate the velocity and, consequently, reduce the dimensionality of the ambient space. Formulating a new FPE for a probability $P_{\text{Smol}}(x; t)$ in the reduced space, we obtain a simpler model because it contains half the dimensions of the original model. Rigorous derivation of this reduction is a classical subject, for which a classic text by Gardiner is recommended[Gar85].

7.1 Langevin equation

The idea of building reduction techniques, starting from the Langevin equation representation of a Markov chain is being advocated by Kurtz & co. [BKPR06, KK13]. Authors introduce the system size parameter N, and a set of exponents γ, α_1,..., α_s, and β_1, ..., β_r by which the time, each of

the s species, and each of r reaction rates are scaled as follows: $t \to \tau = tN^{\gamma}$, $X_i \to Z_i = N^{-\alpha_i} X_i$, and $k_j \to k'_j = N^{\beta_j} k_j$. The 'classical' RRE are obtained if $\gamma = 0$, $\alpha_i = 1$ and $\beta_i = 0\,(-1)$ for unary (binary) reactions. By considering N and all the exponents as free parameters, authors show that, in the limit $N \to \infty$, approximate or exact reduced models can be obtained, which involve different subsets of chemical species, depending on the different choices of exponents.

7.2 QSS and QE

Stochastic partial equilibrium assumption (PEA) is a multi-scale stochastic simulation algorithm (MSSA) which makes use of Gillespie's stochastic simulation algorithm (SSA) together with a new stochastic formulation of the partial equilibrium assumption (PEA) [CGP05b]. I think it is also referred to as the "slow-scale" simulation. Pioneering works in model reduction, applied to a class of 'stochastic Michaelis-Menten' models by Rao & Arkin [RA03], Haseltine & Rawlings [HR02, HR05], Cao et al. [CGP05b], Goutsias [Gou05] integrate the Gillespie algorithm at some stage of approximation. The prerequisite is a model in a standard QSS form. Then the joint probability for all variables $P(x, y)$ is split as a product of the slow variable probability $P(x)$ and the 'fast conditional on slow' probability $P_c(y|x)$ by a relation $P(x, y; t) = P(x)P(y|x)$. The master equation for the slow variable is found by marginalizing over the fast variables. The result is, typically, a death or a birth-death process master equation, whose rates are expressed in terms of averages over the conditional probability P_c . The fast variables are treated by making the assumption that P_c is Markovian. This requirement implies that P_c depends on the slow variable as if it were a time-dependent parameter $x(t)$, i.e. $P_c(y|x) = \tilde{P}(y; x(t))$. Then, Rao & Arkin [RA03] propose to approximate P_c by a stationary distribution (followed by further approximations needed to find such a stationary distribution), and then apply the Gillespie algorithm to the slow rate equation. Haseltine & Rawlings[HR02] propose to use the deterministic approximation for the fast variables.

7.3 Gillespiada

The stochastic simulation algoritm (SSA) aka 'the Gillespie algorithm' an exact algoritm to simulate the Langevin equation. It is simple to code, simple to parallelize and, indeed, has become the method of choice to simulate reasonably simple stochastic models. Gillespie shares his take on stochastic chemical kinetics, paying particular attention to numerical simulation algorithms [GHP13]; a slightly older review on a similar subject is [Gil07].

Tau leaping or τ-leaping, is a framework for approximation and solution of multi-scale systems, where the scales are over a range of populations of chemical species from moderate to very large numbers [Gil01, GP03, RPCG03, RPCG05]. The stochastic version of the Michaelis-Menten has been benchmarked with τ-leaping here [WFCP11].

Slow-scale + SSA = ssSSA Cao et. al. develop a systematic approximate theory that allows one to stochastically advance the system in time by simulating the firings of only the slow reaction events. When it works, very substantial increases in simulation speed can be realized. Some examples [CGP05a] and recent developments [CP08a]. The Michaelis-Menten model, stochastified using the ssSSA procedure, and it's deterministic progenitor, are compared in [GSP11].

Software: StochKit, StochKit2, StochMA, etc. Gillespie, Petzold & co. have been busy with creating useful software for SDEs. Their first version of the popular "StochKit software toolkit" [CP08b], and the follow-up major upgrade [SWR+11]. The next version is interesting because of its model-reducing capabilities: Automatic model analysis algorithm using an adaptively weighted Petri net to dynamically identify opportunities for model reductions for both the stochastic simulation algorithm and tau-leaping simulation, with no requirement of expert knowledge input [WFLP12].

QSS in evolutionary biology Quasi stationary distributions (QSD) in CTMCs are reviewed in [vDP13]. A typical example: biological systems where populations are certain to "die out" eventually, but appear to be stationary over any reasonable time scale.

References

References

[BG03] Nils Berglund and Barbara Gentz. Geometric singular perturbation theory for stochastic differential equations. *Journal of Differential Equations*, 191(1):154, 2003.

[BG13] V. Bykov and V. Gol'dshtein. Fast and slow invariant manifolds in chemical kinetics. *Computers & Mathematics with Applications*, 65(10):1502–1515, May 2013.

[BHLM13] Luca Bortolussi, Jane Hillston, Diego Latella, and Mieke Massink. Continuous approximation of collective systems behaviour: a tutorial. *Performance Evaluation*, 2013.

[BKPR06] Karen Ball, Thomas G. Kurtz, Lea Popovic, and Greg Rempala. Asymptotic analysis of multiscale approximations to reaction networks. *The Annals of Applied Probability*, 16(4):1925–1961, November 2006.

[BM07] V. Bykov and U. Maas. Extension of the ILDM method to the domain of slow chemistry. *Proceedings of the Combustion Institute*, 31(1):465–472, January 2007.

[BM11] Viatcheslav Bykov and Ulrich Maas. Hierarchy analysis and reduction of reacting flow systems. In Krause, E and Shokin, Y and Resch, M and Kroner, D and Shokina, N, editor, *Computational Science and High Performance Computing IV*, volume 115, pages 233–252, Berlin, 2011. Springer-Verlag.

[BO91] I. Babuška and J. Osborn. Eigenvalue problems. In *Handbook of Numerical Analysis*, volume 2, pages 641–787. Elsevier, 1991.

[CGP05a] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. Accelerated stochastic simulation of the stiff enzyme-substrate reaction. *The Journal of Chemical Physics*, 123(14):144917, 2005.

[CGP05b] Yang Cao, Daniel T. Gillespie, and Linda R. Petzold. The slow-scale stochastic simulation algorithm. *The Journal of Chemical Physics*, 122(1):014116, 2005.

[CK11] Eliodoro Chiavazzo and Ilya Karlin. Adaptive simplification of complex multiscale systems. *Physical Review E*, 83(3), March 2011.

[CP08a] Yang Cao and Linda Petzold. Slow-scale tau-leaping method. *Computer Methods in Applied Mechanics and Engineering*, 197(43-44):3472–3479, August 2008.

[CP08b] Yang Cao and Linda Petzold. Slow-scale tau-leaping method. *Computer Methods in Applied Mechanics and Engineering*, 197(43-44):3472–3479, August 2008.

[Fen79] Neil Fenichel. Geometric singular perturbation theory for ordinary differential equations. *Journal of Differential Equations*, 31(1):5398, 1979.

[Fra88] Simon J. Fraser. The steady state and equilibrium approximations: A geometrical picture. *The Journal of chemical physics*, 88:4732, 1988.

[Gar85] Crispin W. Gardiner. *Handbook of stochastic methods*, volume 3. Springer Berlin, 1985.

[GHP13] Daniel T. Gillespie, Andreas Hellander, and Linda R. Petzold. Perspective: Stochastic algorithms for chemical kinetics. *The Journal of Chemical Physics*, 138(17):170901, 2013.

[Gil77] Daniel T. Gillespie. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361, 1977.

[Gil01] Daniel T. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115:1716, 2001.

[Gil07] Daniel T. Gillespie. Stochastic simulation of chemical kinetics. *Annu. Rev. Phys. Chem.*, 58:3555, 2007.

[GK92] Alexander N. Gorban and Iliya V. Karlin. Thermodynamic parameterization. *Physica A: Statistical Mechanics and its Applications*, 190(3):393404, 1992.

[GK03] Alexander N. Gorban and Iliya V. Karlin. Method of invariant manifold for chemical kinetics. *Chemical Engineering Science*, 58(21):47514768, 2003.

[GN06] Dimitris A. Goussis and Habib N. Najm. Model reduction and physical understanding of slowly oscillating processes: The circadian cycle. *Multiscale Modeling & Simulation*, 5(4):1297–1332, January 2006.

[Gou05] John Goutsias. Quasiequilibrium approximation of fast reaction kinetics in stochastic biochemical systems. *The Journal of chemical physics*, 122:184102, 2005.

[Gou12] Dimitris A. Goussis. Quasi steady state and partial equilibrium approximations: their relation and their validity. *Combustion Theory and Modelling*, 16(5):869–926, October 2012.

[Gou13] Dimitris A. Goussis. The role of slow system dynamics in predicting the degeneracy of slow invariant manifolds: The case of vdP relaxationoscillations. *Physica D: Nonlinear Phenomena*, 248:16–32, April 2013.

[GP03] Daniel T. Gillespie and Linda R. Petzold. Improved leap-size selection for accelerated stochastic simulation. *The Journal of Chemical Physics*, 119(16):8229, 2003.

[GSP11] D.T. Gillespie, K.R. Sanft, and L.R. Petzold. Legitimacy of the stochastic Michaelis-Menten approximation. *IET Systems Biology*, 5(1):58–69, January 2011.

[GVL12] G. H. Golub and C. F. Van Loan. *Matrix computations*, volume 3. JHU Press, 2012.

[HR02] Eric L. Haseltine and James B. Rawlings. Approximate simulation of coupled fast and slow reactions for stochastic chemical kinetics. *The Journal of Chemical Physics*, 117(15):6959, 2002.

[HR05] Eric L. Haseltine and James B. Rawlings. On the origins of approximations for stochastic chemical kinetics. *The Journal of chemical physics*, 123:164115, 2005.

[JG11] Kenneth A. Johnson and Roger S. Goody. The original Michaelis constant: Translation of the 1913 Michaelis–Menten paper. *Biochemistry*, 50(39):8264–8269, October 2011.

[Kat95] Tosio Kato. *Perturbation theory for linear operators*, volume 132. springer, 1995.

[KK13] Hye-Won Kang and Thomas G. Kurtz. Separation of time-scales and model reduction for stochastic reaction networks. *The Annals of Applied Probability*, 23(2):529–583, April 2013.

[KSR+11] E. Krause, Y. Shokin, M. Resch, D. Kroner, and N. Shokina, editors. *The 4th Russian-German Advanced Research Workshop, Freiburg, Germany, October 12 to 16, 2009.* Springer, Berlin; Heidelberg, 2011.

[Lam93] S. H. Lam. Using CSP to understand complex chemical kinetics. *Combustion Science and Technology*, 89(5-6):375–404, 1993.

[Lam13] Sau H. Lam. Model reductions with special CSP data. *Combustion and Flame*, 160(12):2707–2711, December 2013.

[LG89] S. H. Lam and D. A. Goussis. Understanding complex chemical kinetics with computational singular perturbation. In *Symposium (International) on Combustion*, volume 22, pages 931–941, 1989.

[LG91] S. H. Lam and D. A. Goussis. Conventional asymptotics and computational singular perturbation for simplified kinetics modeling. In M. D. Smooke, editor, *Reduced Kinetic Mechanisms and Asymptotic Approximations for Methane-Air Flames*, number 384 in Lecture Notes in Physics, pages 227—242. Springer-Verlag, 1991.

[LG94] S. H. Lam and D. A. Goussis. The CSP method for simplifying kinetics. *International Journal of Chemical Kinetics*, 26(4):461–486, 1994.

[LOC+12] A. V. Lebedev, M. V. Okun, V. A. Chorkov, P. M. Tokar, and M. Strelkova. Systematic procedure for reduction of kinetic mechanisms of complex chemical processes and its software implementation. *Journal of Mathematical Chemistry*, 51(1):73–107, August 2012.

[MP92a] Ulrich Maas and Stephen B. Pope. Implementation of simplified chemical kinetics based on intrinsic low-dimensional manifolds. In *Symposium (International) on Combustion*, volume 24, pages 103–112, 1992.

[MP92b] Ulrich Maas and Stephen B. Pope. Simplifying chemical kinetics: intrinsic low-dimensional manifolds in composition space. *Combustion and Flame*, 88(3):239–264, 1992.

[NF89] An Hoang Nguyen and Simon J. Fraser. Geometrical picture of reaction in enzyme kinetics. *The Journal of Chemical Physics*, 91(1):186—193, 1989.

[NF13a] Paolo Nicolini and Diego Frezzato. Features in chemical kinetics. i. signatures of self-emerging dimensional reduction from a general format of the evolution law. *The Journal of Chemical Physics*, 138(23):234101, 2013.

[NF13b] Paolo Nicolini and Diego Frezzato. Features in chemical kinetics. II. a self-emerging definition of slow manifolds. *The Journal of Chemical Physics*, 138(23):234102, 2013.

[NW11] Lena Noethen and Sebastian Walcher. Tikhonovs theorem and quasi-steady state. *Discrete and Continuous Dynamical Systems B*, 16(3):945961, 2011.

[PMKTG13] Thanneer Malai Perumal, Sashikant Madgula Krishna, Sai Sandeep Tallam, and Rudiyanto Gunawan. Reduction of kinetic models using dynamic sensitivities. *Computers & Chemical Engineering*, 56:37–45, September 2013.

[Pop97] S. B. Pope. Computationally efficient implementation of combustion chemistry using in situ adaptive tabulation. *Combustion Theory and Modelling*, 1:41—63, 1997.

[RA03] Christopher V. Rao and Adam P. Arkin. Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the gillespie algorithm. *Journal of Chemical Physics*, 118(11):49995010, 2003.

[RF90] Marc R. Roussel and Simon J. Fraser. Geometry of the steady-state approximation: Perturbation and accelerated convergence methods. *The Journal of chemical physics*, 93(2):1072–1081, 1990.

[RPCG03] Muruhan Rathinam, Linda R. Petzold, Yang Cao, and Daniel T. Gillespie. Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. *The Journal of Chemical Physics*, 119:12784, 2003.

[RPCG05] Muruhan Rathinam, Linda R. Petzold, Yang Cao, and Daniel T. Gillespie. Consistency and stability of tau-leaping schemes for chemical reaction systems. *Multiscale Modeling & Simulation*, 4(3):867–895, January 2005.

[RZ04] Marc R. Roussel and Rui Zhu. Reducing a chemical master equation by invariant manifold methods. *The Journal of Chemical Physics*, 121(18):8716–8730, October 2004.

[Ste90] G. W. Stewart. *Matrix perturbation theory*. Computer science and scientific computing. Academic Press, Boston, 1990.

[Ste01] G. W Stewart. *Matrix algorithms*. SIAM, Philadelphia, 2001.

[SWR+11] K. R. Sanft, S. Wu, M. Roh, J. Fu, R. K. Lim, and L. R. Petzold. StochKit2: software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics*, 27(17):2457–2458, July 2011.

[vDP13] Erik A. van Doorn and Philip K. Pollett. Quasi-stationary distributions for discrete-state models. *European Journal of Operational Research*, 230(1):1–14, October 2013.

[vdSRJ13] Arjan van der Schaft, Shodhan Rao, and Bayu Jayawardhana. On the mathematical structure of balanced chemical reaction networks governed by mass action kinetics. *SIAM Journal on Applied Mathematics*, 73(2):953–973, April 2013.

[Ver05] F. Verhulst. *Methods and applications of singular perturbations: boundary layers and multiple timescale dynamics*. Number 50 in Texts in applied mathematics. Springer, New York, 2005.

[VK92] Nicolaas Godfried Van Kampen. *Stochastic processes in physics and chemistry*, volume 1. Access Online via Elsevier, 1992.

[WFCP11] Sheng Wu, Jin Fu, Yang Cao, and Linda Petzold. MichaelisMenten speeds up tau-leaping under a wide range of conditions. *The Journal of Chemical Physics*, 134(13):134112, 2011.

[WFLP12] Sheng Wu, Jin Fu, Hong Li, and Linda Petzold. Automatic identification of model reductions for discrete stochastic simulation. *The Journal of Chemical Physics*, 137(3):034106, 2012.

[WQH13] Zuozhu Wu, Xinqi Qiao, and Zhen Huang. A criterion based on computational singular perturbation for the construction of reduced mechanism for dimethyl ether oxidation. *Journal of the Serbian Chemical Society*, 78(8):1177–1188, 2013.

[YPC13] Yue Yang, Stephen B. Pope, and Jacqueline H. Chen. Empirical low-dimensional manifolds in composition space. *Combustion and Flame*, 160(10):1967–1980, October 2013.

[ZAI13] Shuliang Zhang, Ioannis P. Androulakis, and Marianthi G. Ierapetritou. A hybrid kinetic mechanism reduction scheme based on the on-the-fly reduction and quasi-steady-state approximation. *Chemical Engineering Science*, 93:150–162, April 2013.

[ZKK04a] Antonios Zagaris, Hans G. Kaper, and Tasso J. Kaper. Analysis of the computational singular perturbation reduction method for chemical kinetics. *Journal of Nonlinear Science*, 14(1):59–91, 2004.

[ZKK04b] Antonios Zagaris, Hans G. Kaper, and Tasso J. Kaper. Fast and slow dynamics for the computational singular perturbation method. *Multiscale Modeling & Simulation*, 2(4):613–638, 2004.