Constrained Approximation of Effective Generators for Multiscale Stochastic Reaction Networks and Application to Conditioned Path Sampling

Simon L. Cotter

School of Mathematics, University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom; e-mail: simon.cotter@manchester.ac.uk. *SC was funded by First Grant Award EP/L023989/1 from EPSRC. The author would like to thank the Isaac Newton Institute for Mathematical Sciences, Cambridge, for support and hospitality during the programme Stochastic Dynamical Systems in Biology: Numerical Methods and Applications, where work on this paper was undertaken. This work was supported by EPSRC grant no EP/K032208/1.*

Abstract

Efficient analysis and simulation of multiscale stochastic systems of chemical kinetics is an ongoing area for research, and is the source of many theoretical and computational challenges. In this paper, we present a significant improvement to the constrained approach, which is a method for computing effective dynamics of slowly changing quantities in these systems, but which does not rely on the quasi-steady-state assumption (QSSA). The QSSA can cause errors in the estimation of effective dynamics for systems where the difference in timescales between the "fast" and "slow" variables is not so pronounced.

This new application of the constrained approach allows us to compute the effective generator of the slow variables, without the need for expensive stochastic simulations. This is achieved by finding the null space of the generator of the constrained system. For complex systems where this is not possible, or where the constrained subsystem is itself multiscale, the constrained approach can then be applied iteratively. This results in breaking the problem down into finding the solutions to many small eigenvalue problems, which can be efficiently solved using standard methods.

Since this methodology does not rely on the quasi steady-state assumption, the effective dynamics that are approximated are highly accurate, and in the case of systems with only monomolecular reactions, are exact. We will demonstrate this with some numerics, and also use the effective generators to sample paths of the slow variables which are conditioned on their endpoints, a task which would be computationally intractable for the generator of the full system.

Keywords: Stochastic, multiscale, chemical kinetics, constrained dynamics

1. Introduction

Understanding of the biochemical reactions that govern cell function and regulation is key to a whole range of biomedical and biological applications and understanding mathematical modelling of gene regulatory networks has been an area of huge expansion over the last half century. Due to the low copy numbers of some chemical species within the cell, the random and sporadic nature of individual reactions can play a key part in the dynamics of the system, which cannot be well approximated by ODEs[13]. Methods for the simulation of such a system, such as Gillespie's stochastic simulation algorithm (SSA)[18], (or the similar Bortz-Kalos-Lebowitz algorithm[5] specifically for Ising spin systems), have been around for some decades. Versions which are more computationally efficient have also been developed in the intermediate years[17, 7].

Unfortunately, their application to many systems can be very computationally expensive, since the algorithms simulate every single reaction individually. If the system is multiscale, i.e. there are some reactions (fast reactions) which are happening many times on a timescale for which others (slow reactions) are unlikely to happen at all, then in order for us to understand the occurrences of the slow reactions, an unfeasible number of fast reactions must be simulated. This is the motivation for numerical methods which allow us to approximate the dynamics of the slowly changing quantities in the system, without the need for simulating all of the fast reactions.

For systems which are assumed to be well-mixed, there are many different approaches and methods which have been developed. For example the τ-leap method[20] speeds up the simulation by timestepping by an increment within which several reactions may occur. This can lead to problems when the copy numbers of one or more of the species approaches zero, and a number of different methods for overcoming this have been presented[31, 2].

Several other methods are based on the quasi steady-state assumption (QSSA). This is the assumption that the fast variables converge in distribution in a time which is negligible in comparison with the rate of change of the slow variable. Through this assumption, a simple analysis of the fast subsystem yields an approximation of the dynamics of the slow variables. This fast subsystem can be analysed in several ways, either through analysis and approximation[6], or through direct simulation of the fast subsystem[11].

Another approach is to approximate the system by a continuous state-space stochastic differential equation (SDE), through the chemical Langevin equation (CLE)[19]. This system can then be simulated using numerical methods for SDEs. An alternative approach is to approximate only the slow variables by an SDE. The SDE parameters can be found using bursts of stochastic simulation of the system, initialised at a particular point on the slow state space[15], the so-called "equation-free" approach. This was further developed into the constrained multiscale algorithm (CMA)[9], which used a version of the SSA which also constrained the slow variables to a particular value. Using a similar approach to [6], the CMA can similarly be adapted so that approximations of the invariant distribution of this constrained system can be made without the need

for expensive stochastic simulations[10]. However, depending on the system, as with the slow-scale SSA, these approximations may incur errors. Work on how to efficiently approximate the results of multiscale kinetic Monte Carlo problems is also being undertaken in many different applications such as Ising models and lattice gas models[24].

Analysis of mathematical models of gene regulatory networks (GRNs) is important for a number of reasons. It can give us further insight into how important biological processes within the cell, such as the circadian clock[33] or the cell cycle[23] work. In order for these models to be constructed, we need to observe how these systems work in the first place. Many of the observation techniques, such as the DNA microarray[27], are notoriously subject to a large amount of noise. Moreover, since the systems themselves are stochastic, the problem of identifying the structure of the network from this data is very difficult. As such, the inverse problem of characterising a GRN from observations is a big challenge facing our community[21].

One popular approach to dealing with inverse problems, is to use a Bayesian framework. The Bayesian approach allows us to combine prior knowledge about the system, complex models and the observations in a mathematically rigorous way[29]. In the context of GRNs, we only have noisy observations of the concentrations of species at a set of discrete times. As such, we have a lot of missing information. This missing data can be added to the state space of quantities that we wish to infer from the data that we do have. This complex probability distribution on both the true trajectories of the chemical concentrations, and on the network itself, can be sampled from using Markov chain Monte Carlo (MCMC) methods, in particular a Gibb's sampler[16]. Within this Gibb's sampler, we need a method for sampling a continuous path for the chemical concentrations given a guess at the reaction parameters, and our noisy measurements. Exact methods for sampling paths conditioned on their endpoints have been developed [16, 25].

In other applications, methods for path analysis and path sampling have been developed, for example discrete path sampling databases for discrete time Markov chains[32], or where the probability of paths, rather than that of trajectories of discrete Markov processes can be used to analyse behaviour[30]. In [12], a method for transition path sampling is presented for protein folding, where the Markov chain has absorbing states. Other approaches for coarse-graining transition path sampling in protein folding also exist[3]. Other methods also exist for the simulation of rare events where we wish to sample paths transitioning from one stable region to another[4].

The problems become even more difficult when, as is often the case, the systems in question are also multiscale. This means that these inverse problems require a degree of knowledge from a large number of areas of mathematics. Even though many of the approaches that are being developed are currently out of reach in terms of our current computational capacity, this capacity is continually improving. In this paper we aim to progress this methodology in a couple of areas.

[1] Define a dominating process to have transition rates given by the matrix $\mathcal{M} = \frac{1}{\rho}\mathcal{G} + I$.
[2] This process has uniformly distributed reaction events on the time interval $[t_0, t_1]$. The number r of such events is given by (1).
[3] Once $r = \hat{r}$ has been sampled, the type of each event must be decided, by sampling from the distribution (2), starting with the first event. An event which corresponds to rate $m_{i,i}$ indicates that no reaction event has occurred at this event.
[4] Once all event types have been sampled, we have formed a sample from the conditioned path space.

Table 1: *A summary of the methodology presented in [16], for sampling paths of Markov-modulated Poisson processes, conditioned on their endpoints.*

1.1. Conditioned path sampling methods

We will briefly review the method presented in [16] for the exact sampling of conditioned paths in stochastic chemical networks. Suppose that we have a Markov jump process, possibly constructed from such a network, with a generator $\mathcal{G}$. The generator of such a process is the operator $\mathcal{G}$ such that the master equation of the system can be expressed as

$$\frac{d\mathbf{p}}{dt} = \mathcal{G}\mathbf{p},$$

where $\mathbf{p}$ is the (often infinite dimensional) vector of probabilities of being in a particular state in the system. We wish to sample a path, conditioned on $X(t_0) = x_0$ and $X(t_1) = x_1$. Such a path can be found by creating a dominating process (i.e. a process whose rate is greater than the fastest rate of any transitions of the original system) with a uniform rate.

We define the rate to be greater than the fastest rate of the process with generator $\mathcal{G}$, so that

$$\rho > \max_i \mathcal{G}_{i,i}.$$

Then we define the transition operator of the dominant process by:

$$\mathcal{M} = \frac{1}{\rho}\mathcal{G} + I.$$

We can then derive the number of reaction events N_U of the dominating process in the time interval $[t_0, t_1]$ by:

$$\mathbb{P}(N_U = r) = \frac{\exp(-\rho t)(\rho t)^r / r! [\mathcal{M}^r]_{x_0,x_t}}{[\exp(\mathcal{G}t)]_{x_0,x_t}}. \tag{1}$$

Here the notation $[\cdot]_{a,b}$ denotes the entry in the matrix with coordinates $(a, b) \in \mathbb{N}^2$. A sample is taken from this distribution. The times $\{t_1^*, t_2^*, \ldots t_r^*\}$ of all of

the r reaction events can then be sampled uniformly from the interval $[t_0, t_1]$. The only thing that then remains is to ascertain which reaction has occurred at each reaction event. This can be found by computing, starting with $X(t_0) = x_0$, the probability distribution defined by:

$$\mathbb{P}(X(t_j^*)) = x | X(t_{j-1}^*) = x_{j-1}^*, X(t_1) = x_1) = \frac{[\mathcal{M}]_{x_{j-1}^*, x} [\mathcal{M}^{r-j}]_{x, x_1}}{[\mathcal{M}^{r-j+1}]_{x_{j-1}^*, x_1}}. \quad (2)$$

This method, summarised in Table 1, exactly samples from the desired distribution, but depending on the size and sparsity of the operator $\mathcal{G}$, it can also be very expensive. In the context of multiscale systems with a large number of possible states of the variables, the method quickly becomes computationally intractable.

1.2. Summary of Paper

In Section 2, we introduce a version of the Constrained Multiscale Algorithm (CMA), which allows us to approximate the effective generator of the slow processes within a multiscale system. In particular, we explore how stochastic simulations are not required in order to compute a highly accurate effective generator. In Section 3, we consider the differences between the constrained approach, and the more commonly used quasi-steady state assumption (QSSA). In Section 4, we describe how the constrained approach can be extended in an iterative nested structure for systems for whose constrained subsystem is itself a large intractable multiscale system. By applying the methodology in turn to the constrained systems arising from the constrained approach, we can make the analysis of highly complex and high dimensional systems computationally tractable. In Section 5, we present some analytical and numerical results, aimed at presenting the advantages of the CMA over other approaches. This includes some examples of conditioned path sampling using effective generators approximated using the CMA. Finally, we will summarise our findings in Section 6.

2. The Constrained Multiscale Algorithm

The Constrained Multiscale Algorithm was originally designed as a multiscale method which allowed us to compute the effective drift and diffusion parameters of a diffusion approximation of the slow variables in a multiscale stochastic chemical network. The idea was simply to constrain the original dynamics to a particular value of the slow variable. This can be done through a simple alteration of the original SSA by Gillespie[18]. First, a (not necessarily orthogonal) basis is found for the system in terms of "slow" and "fast" variables, $[\mathbf{S} = [S_1, S_2, \ldots], \mathbf{F} = [F_1, F_2, \ldots]]$. Slow variables are not affected by the most frequently firing reactions in the system. Then, as shown in [9], the SSA is computed as normal, until one of the slow reactions (a reaction which alters the value of the slow variable(s)) occurs. After the reaction has occurred, the slow variable is then reset to its original value, in such a way that the fast variables are not affected. This is equivalent to projecting the state of the system,

after each reaction, back to the desired value of the slow variable, whilst also preserving the value(s) of the fast variable(s). The constrained SSA is given in Table 2. Here the $\alpha_i(\mathbf{X}(t))$ denote the propensity of the reaction R_i when the system is in state $\mathbf{X}(t) = [X_1(t), X_2(t), \ldots]$, where $\Delta t \alpha_i(\mathbf{X}(t)$ is the probability that this reaction will fire in the infinitesimally small time interval $(t, t + \Delta t)$ with $1 \gg \Delta t > 0$. The stoichiometric vectors $\boldsymbol{\nu}_i$ denote the change in the state vector $\mathbf{X}(t)$ due to reaction R_i firing.

In order to describe the constrained approach, we first introduce some definitions that will be helpful.

Definition 2.1. *Constrained Projector: Given a basis of the state space* $\mathbf{X} = [X_1, X_2, \ldots, X_N]$ *with* N_f *fast variables* $\mathbf{F} = [F_1, F_2, \ldots, F_{N_f}]$ *and* N_s *slow variables* $\mathbf{S} = [S_1, S_2, \ldots, S_{N_s}]$*, the constrained projector* $\mathcal{P}_{\mathbf{S}} : \mathbb{N}_0^N \to \mathbb{N}_0^N$ *for a given value of* $\mathbf{S}$ *preserves the values of the fast variables, whilst mapping the values of the slow variables to* $\mathbf{S}$*:*

$$\mathcal{P}_{\mathbf{S}}([\hat{\mathbf{S}}, \hat{\mathbf{F}}]) = [\mathbf{S}, \hat{\mathbf{F}}] \qquad \forall([\hat{\mathbf{S}}, \hat{\mathbf{F}}]) \in \mathbb{N}_0^N. \tag{3}$$

Definition 2.2. *Constrained Stoichiometric Projector: Given a basis of the state space* $\mathbf{X} = [X_1, X_2, \ldots, X_N]$ *with* N_f *fast variables* $\mathbf{F} = [F_1, F_2, \ldots, F_{N_f}]$ *and* N_s *slow variables* $\mathbf{S} = [S_1, S_2, \ldots, S_{N_s}]$*, the constrained stoichiometric projector* $\mathcal{P} : \mathbb{N}_0^N \to \mathbb{N}_0^N$ *maps any non-zero elements of the slow coordinates to zero, whilst preserving the values of the fast coordinates:*

$$\mathcal{P}([\mathbf{S}, \mathbf{F}]) = [\mathbf{0}, \mathbf{F}] \qquad \forall([\mathbf{S}, \mathbf{F}]) \in \mathbb{N}_0^N. \tag{4}$$

Definition 2.3. *Constrained Subsystem: Given a system with* N_R *reactions* $R_1, R_2, \ldots, R_{N_R}$ *with propensity functions* $\alpha_i(\mathbf{X})$ *and stoichiometric vectors* $\boldsymbol{\nu}_i \in \mathbb{N}_0^N$*, the constrained subsystem is the system that arises from applying the constrained projector* $\mathcal{P}_{\mathbf{S}}$ *to the state vector after each reaction in the system. This is equivalent to applying the constrained stoichiometric projector* $\mathcal{P}$ *to each of the stoichiometric vectors in the system. This may leave some reactions with a null stoichiometric vector, and so these reactions can be removed from the system. This projection can lead to aphysical systems where one or more variables may become negative; in these cases we set the propensities of the offending reactions at states where a move to a negative rate is possible, to zero.*

Let us illustrate this using an example which we shall be using later in the paper.

$$\begin{aligned} R_1 &: \quad \emptyset \xrightarrow{k_1} X_1, \\ R_2 &: \quad X_2 \xrightarrow{k_2} \emptyset, \\ R_3 &: \quad X_1 \xrightarrow{k_3} X_2, \\ R_4 &: \quad X_2 \xrightarrow{k_4} X_1. \end{aligned} \tag{5}$$

[1] Define a basis of the state space in terms of slow and fast variables.

[2] Initialise the value of the state, $\mathbf{X}(t_0) = \mathbf{x}$.

[3] Calculate propensity functions at the current state $\alpha_i(\mathbf{X}(t))$.

[4] Sample the waiting time to the next reaction in the system

$$\tau = -\frac{\log(u)}{\alpha_0(\mathbf{X}(t))}, \quad \text{where} \quad \alpha_0(\mathbf{X}(t)) = \sum_{k=1}^{M} \alpha_k(\mathbf{X}(t)), \qquad u \sim U([0,1]).$$

[5] Choose one $j \in \{1, \ldots, M\}$, with probability α_j/α_0, and perform reaction R_j, with stoichiometry which has been projected using the constrained stoichiometric projector:

$$\mathbf{X}(t+\tau) = \mathbf{X}(t) + \mathcal{P}(\nu_j).$$

[6] Repeat from step **[3]**.

Table 2: *The* Constrained *Stochastic Simulation Algorithm (CSSA) using the constrained stoichiometric projector given in Definition 2.2. Simulation starts with $S = s$ where s is a given value of the slow variable.*

In certain parameter regimes, this system is multiscale, with reactions R_3 and R_4 occurring many times on a time scale for which reactions R_1 and R_2 are unlikely to happen at all. The variable $S = X_1 + X_2$ is unaffected by these fast reactions, and as such is a good candidate for the slow variable which we wish to analyse. A discussion about how the fast and slow variables could be identified is given in Section 6. We have two choices for the fast variable, either $F = X_1$ or $F = X_2$, in order to form a basis of the state space along with the slow variable S. As detailed in [9], it is preferable (although not essential) to pick fast variables that are not involved in zeroth order reactions. Therefore, in this case, we choose $F = X_2$. Following the projection of the stoichiometric vectors using the constrained projector, the constrained system can be written in the following way:

$$\begin{aligned} C_1 &: \quad X_1 + X_2 = S, \\ R_2 &: \quad X_2 \xrightarrow{k_2} X_1, \\ R_3 &: \quad X_1 \xrightarrow{k_3} X_2, \\ R_4 &: \quad X_2 \xrightarrow{k_4} X_1. \end{aligned} \tag{6}$$

Note that reaction R_1 has disappeared completely, since only involves changes to the slow variable, and as such after projection, the stoichiometric vector is null, and the reaction can be removed. The stoichiometry of reaction R_2 has been altered as it involves a change in the slow variables. If this reaction occurs, the slow variable is reduced by one. We are not permitted to change the fast variable X_2 in order to reset the slow variable to its original value, and therefore

we must increase X_1 by one, giving us a new stoichiometry for this reaction.

In the original CMA, statistics were taken regarding the frequency of the slow reactions, at each point of the slow domain, and were used to construct the effective drift and diffusion parameters of an effective diffusion[9, 8] process. However, this constrained approach can also be used to compute an effective generator for the discrete slow process, as we will now demonstrate. The CMA can be very costly, due to the large computational burden of the stochastic simulations of the constrained system. In this section, we will introduce a method for avoiding the need for these simulations, whilst also significantly improving accuracy.

The constrained systems can often have a very small state space (which we will denote $\Gamma(s)$), since they are constrained to a single value of the slow variables. For example, for the constrained system (6), there are only $\left\lfloor \frac{S}{2} \right\rfloor$ possible states. Such a system can easily be fully analysed. For example, the invariant distribution can be found by characterising the one-dimensional null space of the generator matrix of the constrained process. For small to medium-sized systems, this is far more efficient than exhaustive Monte Carlo simulations. For other systems with larger constrained state spaces, stochastic simulation may still be the best option, although in Section 4 we show how the constrained approach can be applied iteratively until the constrained subsystem is easily analysed.

Suppose that we have a constrained system with N_f fast variables, $F_1, F_2, \ldots, F_{N_f}$. The generator for the constrained system with $S = s$ is given by $\mathcal{G}_F(s)$. Since the system is ergodic, there is a one-dimensional null space for this generator. This can be found by using standard methods for identifying eigenvectors, by searching for the eigenvector corresponding to the eigenvalue equal to zero. Krylov subspace methods allow us to find these eigenvectors with very few iterations. Suppose we have found such a vector $\mathbf{v} = [v_1, v_2, \ldots]$, such that

$$\mathcal{G}_F(s)\mathbf{v} = 0.$$

Then our approximation to the invariant distribution of this system is given by the discrete probability distribution represented by the vector

$$\mathbf{p}(s) = [p_1(s), p_2(s), \ldots] = \frac{\mathbf{v}}{\sum v_i}.$$

Our aim is now to use this distribution to find the effective propensities of the slow reactions of the original system.

Suppose that we have N_s slow reactions in the original system. Each has an associated propensity function $\alpha_1(S, F), \alpha_2(S, F), \ldots, \alpha_{N_s}(S, F)$. We now simply want to find the expectation of each of these propensity functions with respect to the probability distribution $\mathbf{p}(s)$:

$$\mathbb{E}(\alpha_i(S, \cdot)) = \sum_i p_i(s)\alpha_i(S, f). \tag{7}$$

[1] For each value of the slow variable $S = s \in \Omega$, compute the generator $\mathcal{G}_s$ of the constrained subsystem.
[2] Find the zero eigenvector $\mathbf{v} = [v_1, v_2, \ldots]$ of $\mathcal{G}_s$, and let $\mathbf{p}(s) = \frac{\mathbf{v}}{\sum v_i}$.
[3] Approximate the effective propensities at each point $s \in \Omega$ using (7).
[4] Construct an effective generator $\mathcal{G}$ of the slow processes of the system using these effective propensities.

Table 3: *The CMA approach to approximating the effective generator $\mathcal{G}$ of the slow variables on the (possibly truncated) domain $S \in \Omega$, without the need for stochastic simulations.*

Having computed this expectation for all of the slow propensities, over all required values of the slow variable, then an effective generator for the slow variable can be constructed.

3. Comparing the CMA and QSSA approaches

A very common approach to approximating the dynamics of slowly changing quantities in multiscale systems, is to invoke the quasi steady-state assumption (QSSA). The assumption is that the fast and slow variables are operating on sufficiently different time scales that it can be assumed that the fast subsystem enters equilibrium instantaneously following a change in the slow variables, and therefore is unaffected by the slow reactions. This assumption means that if the fast subsystem's invariant distribution can be found (or approximated), then the effective propensities of the slow reactions can be computed. However, as demonstrated in [8], this assumption incurs an error, and for systems which do not have a large difference in time scales between the fast and slow variables, this error can be significant.

The CMA does not rely on the QSSA, and is able to take into account the effect that the slow reactions have on the invariant distribution of the fast variables, conditioned on a value of the slow variables. In a true fast-slow system, this will yield the same results as the QSSA, but for most systems of interest, the constrained approach will have a significant increase in accuracy. If we follow the approach outlined in Table 3, we don't even need to conduct any stochastic simulations to approximate the effective dynamics.

The assumptions for the CMA are weaker than the QSSA, namely that we assume that the dynamics of the slow variable(s) can be approximated by a Markov-modulated Poisson process, independently of the value of the fast variables. This means that we have made the assumption that the current value of the fast variables has no effect on the transition rates of the slow variables once a slow reaction has occurred. This is subtly weaker than the QSSA, and importantly the effect of the slow reactions on the invariant distribution of the fast variables is accounted for. Note that this may necessitate a slow variable which has more than one dimension, for example in oscillating systems for which

the effective dynamics cannot be approximated by a one dimensional Markov process. Consideration of such systems is an area for future work.

4. The Nested CMA

There will be many systems for which the constrained subsystem is itself a highly complex and multiscale system. In this event, it will not be feasible to find the null space of a sensibly truncated generator for the constrained subsystem. Therefore, we need to consider how we might go about approximating this. Fortunately, we already have the tools to do this, since we can iteratively apply the CMA methodology to this subsystem. This is analogous to the nested strategy proposed in the QSSA-based nested SSA[11].

This nested approach allows us to reduce much more complex systems in an accurate, computationally tractable way. The problem of finding the null space of the first constrained subsystem is divided into finding the null space of many small generators, through further constraining. An example of this nested approach will be demonstrated in Section 5.3.

5. Examples

In this section we will present some analytical and numerical results produced using the CMA approach for three different examples. In order to give an indication of the computational cost of the algorithms, we include the runtime of certain operations. All numerics were performed using MATLAB on a mid-2014 MacBook Pro. Disclaimer: the implementations used are not highly optimised, and these runtimes are purely given as an indication of the true costs of a well implemented version.

5.1. A Simple Linear System

First we consider a simple linear system, in order to demonstrate that the CMA approximation of the effective generator of the slow variable is exact in the case of systems with only monomolecular reactions, which is in contrast to the approximation found using a more standard QSSA-based approach. Let us illustrate this by returning to the example given by the linear system (5), first analysing it using the QSSA.

$$\begin{aligned} R_1 &: \quad \emptyset \xrightarrow{k_1} X_1, \\ R_2 &: \quad X_2 \xrightarrow{k_2} \emptyset, \\ R_3 &: \quad X_1 \xrightarrow{k_3} X_2, \\ R_4 &: \quad X_2 \xrightarrow{k_4} X_1. \end{aligned}$$

We will consider this system in the following parameter setting:

$$k_1 V = 20, \qquad k_2 = 1, \qquad k_3 = 5, \qquad k_4 = 5. \tag{8}$$

Here V denotes the volume of the well-mixed thermally-equilibrated reactor.

5.1.1. QSSA-based analysis

The QSSA tells us that the fast subsystem (made up of reactions R_3 and R_4) reaches probabilistic equilibrium on a timescale which is negligible in comparison with the timescale on which the slow reactions are occurring. Therefore we may treat this subsystem in isolation with fixed S:

$$X_1 \underset{k_4}{\overset{k_3}{\rightleftarrows}} X_2, \qquad S = X_1 + X_2.$$

This is a very simple autocatalytic reaction system, for which a great deal of analytical results are available. For instance, we can compute the invariant distribution for this system[22], which gives us that X_2 is a binomial random variable

$$X_2 \sim \mathcal{B}\left(\cdot, S, \frac{k_3}{k_3 + k_4}\right).$$

Therefore, we can compute the conditional expectation $\mathbb{E}(X_2|S) = \frac{k_3 S}{k_3+k_4}$ in this fast subsystem, and use this to approximate the effective rate of reaction R_2. Therefore, the effective slow system is given by the reactions:

$$\emptyset \xrightarrow{\hat{k}_1} S \xrightarrow{\hat{k}_2} \emptyset, \tag{9}$$

where

$$\hat{k}_1 = k_1, \qquad \hat{k}_2 = \frac{k_2 \mathbb{E}(X_2)}{S} = \frac{k_2 k_3}{k_3 + k_4}.$$

We can compute the invariant distribution for this effective system[22], which in this instance is a Poisson distribution:

$$S \sim \mathcal{P}\left(\frac{k_1 V(k_3 + k_4)}{k_2 k_3}\right). \tag{10}$$

We can quantify the error we have made in using the quasi-steady state assumption by, for example, comparing this distribution with the true invariant distribution. Once again, using the results of [22], we can compute the invariant distribution of the system (5), which is a multivariate Poisson distribution:

$$[X_1, X_2] \sim \mathcal{P}(\bar{\lambda}_1, \bar{\lambda}_2),$$

where $\bar{\lambda}_1 = \frac{k_1 V(k_2+k_4)}{k_2 k_3}$, and $\bar{\lambda}_2 = \frac{k_1 V}{k_2}$. Trivially one can compute the marginal distribution on the slow variable S:

$$\begin{aligned} \mathbb{P}(S = s) &= \sum_{n=0}^{s} \frac{\bar{\lambda}_1^n}{n!} \frac{\bar{\lambda}_2^{s-n}}{(s-n)!} \exp(-(\bar{\lambda}_1 + \bar{\lambda}_2)), \\ &= \frac{(\bar{\lambda}_1 + \bar{\lambda}_2)^s}{s!} \exp(-(\bar{\lambda}_1 + \bar{\lambda}_2)). \end{aligned}$$

Therefore S is also a Poisson variable with intensity $\lambda = \bar{\lambda}_1 + \bar{\lambda}_2 = \frac{k_1 V(k_2+k_3+k_4)}{k_2 k_3}$, which differs from the intensity approximated invariant density (10) by $\frac{k_1 V}{k_3}$.

Note that k_3 is one of the fast rates, and $k_1 V$ is one of the slow rates, and therefore as the difference in timescales of the fast and slow reactions increases, this error decreases to zero, so that the QSSA gives us an asymptotically exact approximation of the slow dynamics. For systems with a finite timescale gap, the QSSA approximation will incur error over and above the error incurred in any approximation of the marginalised slow process by a Markov process.

5.1.2. CMA analysis

For comparison, let us compute approximations of the effective slow rates by using the CMA. The CMA for this system tells us that we need to analyse the constrained system (6).

$$\begin{aligned} C_1 &: & X_1 + X_2 &= S, \\ R_2 &: & X_2 &\xrightarrow{k_2} X_1, \\ R_3 &: & X_1 &\xrightarrow{k_3} X_2, \\ R_4 &: & X_2 &\xrightarrow{k_4} X_1. \end{aligned}$$

The constrained system in this example only contains monomolecular reactions, and as such can be analysed using the results of [22]. The invariant distribution for this system is a binomial, such that

$$X_2 \sim \mathcal{B}\left(\cdot, S, \frac{k_3}{k_2 + k_3 + k_4}\right).$$

Using this, we can compute the effective propensity of reaction R_2,

$$\bar{\alpha}_2(S) = k_2 \mathbb{E}(X_2 | S) = \frac{k_2 k_3 S}{k_2 + k_3 + k_4},$$

giving us the effective rate $\bar{k_2} = \frac{k_2 k_3}{k_2 + k_3 + k_4}$. The invariant distribution of (9) with this effective rate for $\bar{k_2}$ is once again a Poisson distribution with intensity

$$\lambda = \frac{k_1 V (k_2 + k_3 + k_4)}{k_2 k_3},$$

which is *identical* to the intensity of the true distribution on the slow variables. In other words, for this example, the CMA produces an approximation of the effective dynamics of the slow variables for this system, whose invariant distribution is identical to the marginal invariant distribution of the slow variables in the full system. The constrained approach corrects for the effect of the slow reactions on the invariant distribution of the fast variables. In this and other examples of systems with monomolecular reactions, the constrained approach gives us a system whose invariant distribution is exactly equal to the marginal distribution on the slow variables for the full system. Another example is presented in Section 5.3, for which the constrained system is itself too large to easily compute expectations directly through its generator, and requires another iteration of the CMA to be applied.

For this example, we did not even need to compute the invariant distributions of the constrained systems numerically. In Section 5.2, we will come across a system for which it is necessary to numerically compute the invariant distribution of the constrained system.

5.1.3. Comparison of approximation of invariant densities

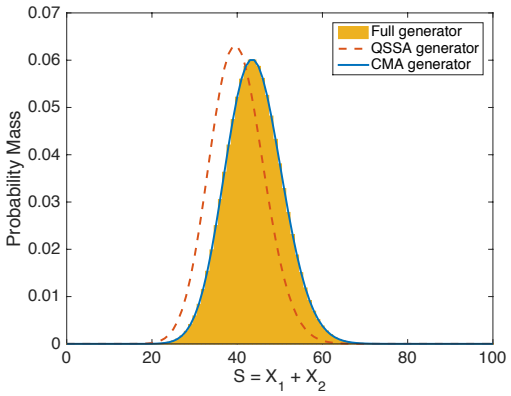


Figure 1: *Approximations of the invariant distribution of the slow variable $S = X_1 + X_2$ of system* (5) *with parameters* (8) *through marginalisation of the distribution of the full system (histogram), of the effective generator computed using the CMA (solid line) and computed using the QSSA (dashed line).*

Figure 1 shows the invariant distributions of the slow variables $S = X_1 + X_2$ in the parameter regime (8), computed by marginalising the invariant distribution of the full system, and from the CMA and QSSA as outlined above. The CMA exactly matches the true distribution, as both are Poisson distributions with rate $\lambda = 44$. The QSSA incorrectly approximates the effective rate of R_4, and as such is a Poisson distribution with rate $\lambda = 40$. The relative error of the CMA for this problem is zero, and for the QSSA is 4.322×10^{-1}.

5.1.4. Conditioned path sampling using effective generators

The approaches described in Section 1.1 hit problems when the system for which we are trying to generate a conditioned path is multiscale. In a multiscale system, the rate ρ of the dominating process will be very large, and as such the number of reaction events will be large, even if the path we are trying to sample is short. Therefore M^r is likely to be a full matrix, and the number of calculations of (2) will be large. Moreover, the size of M is also likely to be large, since for each value $S = s$ of the slow variable, there are many states, one for each possible value of the fast variable. All of these factors make the problem of computing a conditioned path in such a scenario computationally intractable.

Considering once more the system (5), naturally we cannot store the actual generator of this system, since the system is open and as such the generator is an infinite dimensional operator. However, the state space can be truncated carefully in such a way that the vast majority of the states with non-negligible invariant density are included, but an infinite number of highly unlikely states are presumed to have probability zero. Note that this means that we are effectively sampling paths satisfying $S(t_0) = s_1$, $S(t_1) = s_2$ conditioned on $S(t) \in \Omega \quad \forall t$. However, even with careful truncation the number of states can be prohibitively large.

Suppose that we truncate the domain for this system to

$$\Omega = \{[X_1, X_2] | X_1, X_2 \in \{0, 1, \ldots, 200\}\,.$$

This truncated system has $201^2 = 40401$ different states, and therefore the generator $\mathcal{G} \in \mathbb{R}^{40401\times 40401}$. Although this matrix is sparse, the matrix exponential required in (1) is full, as is M^r for moderate $r \in \mathbb{N}$. A full matrix of this size stored at double precision would require over 13GB of memory. So even for this system, the most simple multiscale system that one could consider, the problem of sampling conditioned paths is computationally intractable.

In comparison, suppose that we use a multiscale method such as the CMA to approximate the effective rates of the slow reactions. Then, for the same Ω, we only have 401 possible states of the slow variable, a reduction of 99.25%. The effective generator $\mathcal{G} \in \mathbb{R}^{401\times 401}$ would then only require 1.29MB to be stored as a full matrix in double precision. The dominating process for this system must now have rate $\rho > 201.4$, instead of $\rho > 1220$, which is over 6 times bigger. This means far fewer calculations of (2). What is more, as we saw in Section 5.1.2, for some systems the CMA *exactly* computes the effective dynamics of the slow variables, with no errors.

The system (5), in order to highlight more effectively the differences between the CMA and a QSSA-based approach, is in a parameter range (8), for which the difference in time scales between the "fast" and "slow" variables is relatively small, and of course for systems with larger timescale difference, the difference in ρ between the full and effective generators would be far larger.

Naturally, this approach only allows us to sample the paths of the slow variables. However, the fast variables, if required, can easily be sampled after the fact, using an adapted Gillespie approach which samples the fast variables given a trajectory of the slow variables.

As we have just demonstrated in the previous section, the CMA can be used to compute an effective generator for the slow variable $S = X_1 + X_2$ in the system (5), with parameters (8), whose invariant distribution is exactly that of the slow variable in the full system without the multiscale reduction. Moreover, this can be achieved with no Monte Carlo simulations, since the constrained subsystem contains only monomolecular reactions, and as such its invariant distribution can be exactly computed[22].

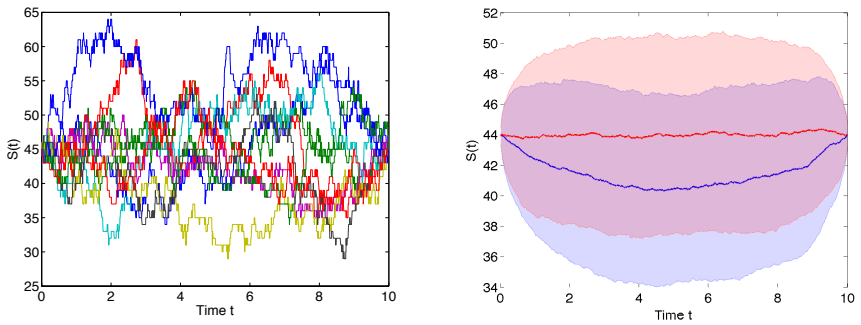
At this juncture, we simply need to apply the method of Fearnhead and Sherlock[16] to the computed effective generator in order to be able sample paths

conditioned on their endpoints. Suppose we wish to sample paths conditioned on $S(t_0 = 0) = 44 = S(t_1 = 10)$. The invariant distribution of this system, as shown previously in this paper, is a Poisson distribution with mean $\lambda = \frac{k_1 V(k_2+k_3+k_4)}{k_2 k_3} = 44$. Therefore, we are attempting to sample paths which start and finish at the the mean of the invariant distribution, which in itself is not a particularly interesting thing to do, but it will allow us to highlight again the advantages of using the CMA over QSSA-based approaches.

Since the system is open, we are required to truncate the domain in order to be able to store and manipulate the effective generator. We truncate the domain to $\Omega = \{[X_1, X_2] | S = X_1 + X_2 \leq 400\}$. Therefore we aim to sample paths

$$\{S(t),\, t \in [0, 10] \,|\, S(0) = 44 = S(10),\, S(t) \in \Omega \,\forall t \in [0, 10]\}.$$

As the number of possible states of the slow variable is relatively small, it was possible to compute and store full matrices for M^r as required in (1) and (2) for $r \in 1, 2, \ldots, 2369$. r has an upper bound of 2369 as the cumulative mass function for the probability distribution (1) is within machine precision of one at $r = 2369$. Storing all powers of the matrices is clearly not the most efficient way to implement this algorithm, but for this example was possible without any intensive computations, and with minimal numerical error. We will present a more efficient approach in the next section.



(a) (b)

Figure 2: *(a) 10 trajectories of the slow variable conditioned on $S(0) = 44 = S(10)$, sampled using the CMA approximate effective generator. (b) Mean and standard deviation of 1000 trajectories of the slow variable conditioned on $S(0) = 44 = S(10)$, sampled using the approximate effective generator from both the QSSA (blue plots) and CMA (red plots).*

Figure 5.1.4 (a) shows 10 example trajectories sampled using the the conditioned path sampling algorithm with the CMA approximation of the effective generator of the slow variable. We also implemented exactly the same approach using the QSSA approximation of the effective generator. The mean and standard deviation of 1000 paths sampled using both methods is plotted in Figure 5.1.4 (b). Since the paths are conditioned to start and finish at the mean of

the system's monomodal invariant distribution, we would expect the mean to converge to a constant $S = 44$ as we sample more paths.

This appears to be the case for the paths sampled using the CMA effective generator, which is what we would hope since this generator preserves the true mean of the slow variables, as demonstrated in the previous section.

The QSSA, as has also been demonstrated in Section 5.1.1, does not correctly preserve the invariant distribution of the slow variables, and underestimates the mean value of the invariant distribution. This can be seen in 5.1.4 (right), where the mean value of the path dips in the middle of the trajectory as it reverts to the mean of the invariant distribution of the QSSA approximation, before increasing towards the end of the trajectory in order to satisfy the condition $S = 44$.

This demonstrates that the accuracy of the approximation of the effective dynamics has a knock-on impact, as one would expect, to the accuracy of the conditioned path sampling. It would be preferable, naturally, if we could compare path statistics of the multiscale approaches to that of conditioned paths statistics of the full system. However, this is simply not feasible, due to the size of the matrices, even for the truncated domain Ω. Instead, this does succeed in demonstrating that these methods make conditional path sampling of the slow variables a possibility, where it was computationally intractable previously. Since the rates could be explicitly calculated for this simple example, the effective generators could be produced in the order of 10^-3 seconds for the domain $S \in \{0, 1, \ldots, 400\}$. The set up process for the path sampling, involving finding the probabilities in (1) and computing the required powers of $\mathcal{M}$ took around 90 seconds. After this, each path took a third of a second to sample.

5.2. A Bistable Example

Sampling of paths conditioned on their endpoints is an integral part of some approaches to Bayesian inversion of biochemical data. A Gibb's sampler can be used to alternately update the network structure and system parameters, and the missing data (i.e. the full trajectory), sampled for example using the method found in [16]. However, efficient methods to sample paths of multiscale systems may also be useful in other areas. For instance, it may allow us to sample paths which make rare excursions, or large deviations from mean behaviour. This forms part of the motivation for considering the next example.

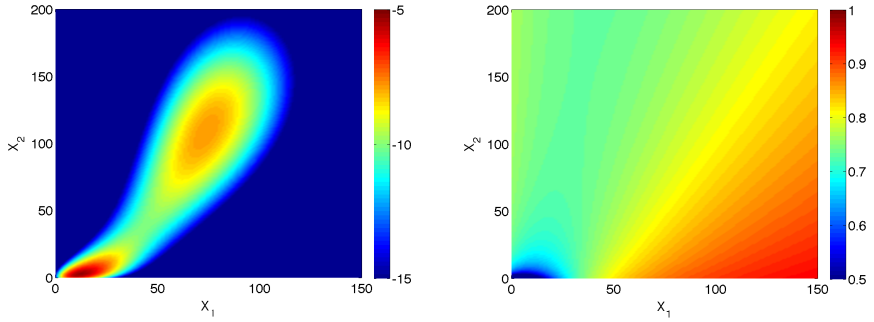
Let us consider the following chemical system, which in certain parameter regimes exhibits bistable behaviour.

$$\begin{aligned} R_1, R_2 \quad &: \quad X_2 \underset{k_2}{\overset{k_1}{\rightleftarrows}} X_1 + X_2, \\ R_3, R_4 \quad &: \quad \emptyset \underset{k_4}{\overset{k_3}{\rightleftarrows}} X_1, \\ R_5, R_6 \quad &: \quad X_1 + X_1 \underset{k_6}{\overset{k_5}{\rightleftarrows}} X_2, \\ R_7 \quad &: \quad X_2 \xrightarrow{k_7} \emptyset. \end{aligned} \tag{11}$$

In particular, we consider parameter regimes where the occurrence of reactions R_5 and R_6 are on a relatively faster timescale than the other reactions. The following is just such a parameter regime:

$$k_1 = 142, \qquad \frac{k_2}{V} = 1, \qquad k_3 V = 880, \tag{12}$$
$$k_4 = 92.8, \qquad \frac{k_5}{V} = 10, \qquad k_6 = 500, \qquad k_7 = 6.$$

As before, V denotes the volume of the well-mixed thermally-equilibrated reactor.



(a) (b)

Figure 3: *(a) A log plot of an approximation π_Ω of the invariant distribution on the slow variable $S = X_1 + 2X_2$ of system* (11) *with parameters* (5.2)*, demonstrating the bistable nature of the system. Approximation was computed by finding the null space of the full generator of the system on the truncated domain $\{0, 1, \ldots, 800\} \times \{0, 1, \ldots, 1200\}$. (b) Proportion of total propensity $P_{R_5,R_6}(X_1, X_2)$ attributed to the fast reactions R_5 and R_6, given by* (13)*.*

That said, this parameter regime is one in which the use of the QSSA will create significant errors, since the timescale gap is not very large in all parts of the domain as demonstrated in Figure 3. Figure 3 (a) shows a highly accurate approximation of the invariant distribution of the full system, found by computing the null space of the full generator for the system truncated to the finite domain $\Omega = \{(x_1, x_2) \in \{0, 1, \ldots, 800\} \times \{0, 1, \ldots, 1200\}$. The zero eigenvector of this truncated generator was found using standard eigenproblem solvers, then normalised. Since this system has 2nd order reactions, its invariant density cannot in general be written in closed form, and as such, we could use this approximation on the truncated domain in order to quantify the accuracy of the multiscale approaches. This plot demonstrates the bistable nature of this system, which can take a long time to switch between the two favourable regions. This example has been chosen in order that such an approximation can still be computed in order to check the accuracy of the approach.

Figure 3 (b) shows the proportion of the total propensity for each state which

is attributed to the fast reactions, R_5 and R_5, given by:

$$P_{R_5,R_6}(X_1,X_2) = \frac{\alpha_5(X_1,X_2)+\alpha_6(X_1,X_2)}{\alpha_0(X_1,X_2)} = \frac{\alpha_5(X_1,X_2)+\alpha_6(X_1,X_2)}{\sum_{i=1}^{M}\alpha_i(X_1,X_2)}. \tag{13}$$

This proportion, which is a measure of the gap in timescales between the "fast" reactions R_5 and R_6, and the rest of the reactions, varies across the domain. We can approximate the expected proportion of propensity attributed to the fast reactions:

$$\mathbb{E}(P_{R_5,R_6}) = \sum_{(X_1,X_2)\in\Omega} P_{R_5,R_6}(X_1,X_2)\pi_\Omega(X_1,X_2),$$

where π_Ω is the approximate invariant density of the full generator on the truncated domain Ω. In this system with parameters (5.2), $\mathbb{E}(P_{R_5,R_6}) = 0.6941$, i.e. the expected proportion of all reactions which are either of type R_5 or R_6 is 69.41%. As such, although reactions R_5 and R_6 are occurring more frequently than other reactions, there is not a stark difference in timescales, as we might expect in a system for which the QSSA yields a good approximation. The "fast" reactions in this example are reactions R_5 and R_6, and as such, $S = X_1 + 2X_2$ is a good choice of slow variable, since this quantity is invariant to these fast reactions.

5.2.1. The QSSA Approach

By applying the QSSA to the system (11), we can approximate the effective rates of the slow variables by considering the fast reactions in isolation. The fast subsystem is given by the reactions R_5 and R_6:

$$\begin{aligned} C_1 \quad &: \qquad X_1 + 2X_2 = S, \\ R_5, R_6 \quad &: \qquad X_1 + X_1 \underset{k_6}{\overset{k_5}{\rightleftarrows}} X_2. \end{aligned} \tag{14}$$

Lines denoted by C_i in this and what follows denotes a constraint. It is important to keep a track of these constraints, since each one reduces the dimension of the state space by one.

For a fixed value of $S = X_1 + 2X_2 \in \{0, 1, \ldots, S_{\max}\}$, we wish to find the generator for the process X_2 (or equivalently $X_1 = S - 2X_2$) within this fast subsystem. The generator can be found by considering the master equation for each state $X_2 = i$:

$$\begin{aligned} \frac{dp_i}{dt} &= -(\alpha_5(S-2i,i)+\alpha_6(S-2i,i))p_i + \alpha_5(S-2(i-1),i-1)p_{i-1} \\ &+ \alpha_6(S-2(i+1),i)p_{i+1}, \end{aligned}$$

where $p_i(t)$ is the probability of $X_2(t) = i$. Putting this set of differential equations into vector form gives us:

$$\frac{d\mathbf{P}}{dt} = \mathcal{G}\mathbf{P},$$

where $\mathcal{G}$ is the generator of the fast subsystem (14). Note that since we are restricted to states such that $X_1 + X_2 = S$ for some value of S, there are only $\lfloor \frac{S}{2} \rfloor$ possible different states, and as such $\mathcal{G} \in \mathbb{R}^{\lfloor \frac{S}{2} \rfloor \times \lfloor \frac{S}{2} \rfloor}$. Even for moderately large values of S, the one-dimensional null space of such a sparse matrix is not computationally expensive to find, and when normalised gives us the invariant density of X_2 (and therefore X_1 if required). This invariant density can then be used to compute the expectation of the propensities of the slow reactions of the system for the state S as in (7), and in turn be entered into the (truncated) effective generator for the slow variable.

5.2.2. The Constrained Approach

When using the CMA, the methodology is largely the same as was described for the QSSA-based approach in the last section. The only real difference lies in the subsystem which is analysed in order to compute the invariant distribution of the fast variables conditioned on the value of the slow variable. As we have done previously, we will consider each of the reactions in the system in turn, constraining the value of the slow variable to a particular value, whilst being sure not to change the value of the fast variables. There are two choices for the fast variable, in order to form a basis of the state space along with the slow variable S, but as explained in detail in [9], $F = X_2$ is the best choice, since there is a zeroth order reaction involving X_1, which can lead to an unphysical constrained subsystem, if this is chosen as the fast variable.

With this choice of fast variables, the first four reactions all disappear in the constrained subsystem. This is because none of these reactions alter the fast variable, and as such the constrained stoichiometric projector maps their stoichiometric vectors to zero, and therefore reactions R_1, R_2, R_3, R_4 have no net effect on the constrained subsystem.

Reaction R_7 differs in that it causes a change in the fast variable X_2. The projector in this case maps the stoichiometric vector to $[-2, 1]^T$ and therefore the net effect of reaction R_7 is equivalent to $X_2 \xrightarrow{k_7} X_1 + X_1$. This leads to the following constrained system:

$$\begin{aligned} C_1 \quad &: \qquad X_1 + 2X_2 = S, \\ R_5, R_6 \quad &: \qquad X_1 + X_1 \underset{k_6}{\overset{k_5}{\rightleftarrows}} X_2, \\ R_7 \quad &: \qquad X_2 \xrightarrow{k_7} X_1 + X_1. \end{aligned}$$

Note that since reactions R_6 and R_7 have the same stoichiometry, this system can be simplified by removing R_7 and adding its rate to R_6:

$$\begin{aligned} C_1 \quad &: \qquad X_1 + 2X_2 = S, \\ R_5, R_6 \quad &: \qquad X_1 + X_1 \underset{k_6+k_7}{\overset{k_5}{\rightleftarrows}} X_2. \end{aligned} \tag{15}$$

For every fixed value of $S \in \{0, 1, \ldots, S_{\max}\}$, the generator for (15) can be found following the same approach as in the previous section, the only difference

	QSSA	CMA	π_Ω
Relative l^2 difference	6.347×10^{-1}	1.796×10^{-2}	-
LH peak position	20	20	20
LH peak height	5.378×10^{-3}	1.591×10^{-2}	1.582×10^{-2}
RH peak position	309	295	295
RH peak height	6.192×10^{-2}	4.060×10^{-3}	4.006×10^{-3}

Table 4: *Differences in the accuracy of the QSSA and CMA approximations of the invariant density of S, with respect to the approximation π_Ω.*

being the altered rate for reaction R_6. Following this methodology, an effective generator $\mathcal{G}$ can be computed.

5.2.3. Comparison of approximation of invariant densities

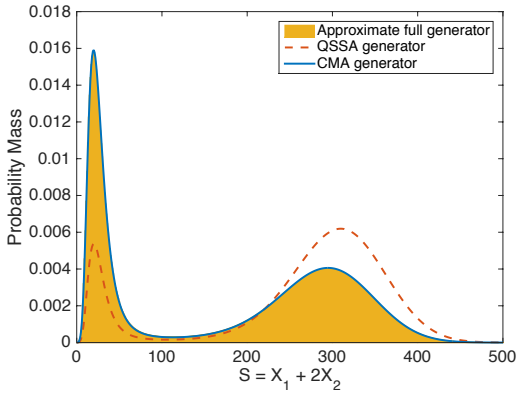


Figure 4: *Approximations of the invariant distribution of the slow variable $S = X_1 + 2X_2$ of system* (11) *with parameters* (5.2)*, through computing the null space of the truncated generator of the full system (histogram), of the effective generator computed using the CMA (solid line) and computed using the QSSA (dashed line).*

One approach to quantifying the accuracy of these two methods of approximating effective generators of the slow variable, is to compare the invariant distributions of the two systems with that of the marginalised density of the slow variable in the full system. We consider the approximation π_Ω of the invariant density of the full system, truncated to the region $\Omega = \{(x_1, x_2) \in \{0, 1, \ldots, 800\} \times \{0, 1, \ldots, 1200\}$, as shown in Figure 3 (a). We can marginalise this density to find an approximation of the invariant density of the slow variable, as is shown by the histogram in Figure 4.

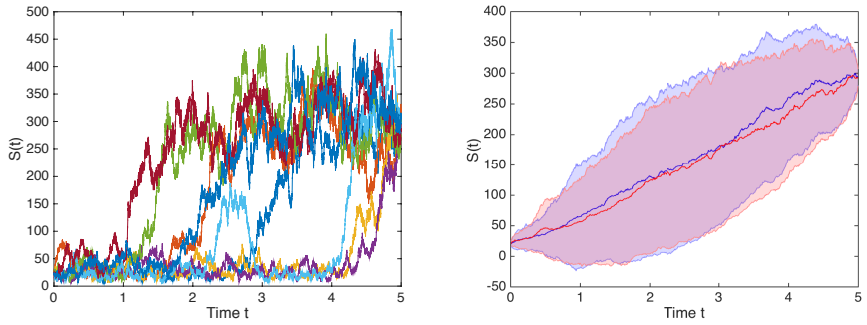
The CMA approximation of the invariant density of the slow variable is indistinguishable by eye from the highly accurate approximation computed in this manner, as shown in Figure 4. The QSSA approximation, on the other hand, incorrectly approximates both the placement and balance of probability

mass of the two peaks in the distribution. The difference in the quality of these approximations is stark. This example is an extreme one, as the parameters have been chosen to demonstrate how far apart these two approximations can be, but since the CMA has no additional costs associated with it, the advantages of this approach are significant. The relative l^2 errors of these two approaches, when compared with the approximate density π_Ω, are given in Table 4, along with the position and heights of the two local maxima in the densities.

The CMA computed the generator on the domain $S \in [0, 2000]$ in around 55 seconds, and the eigensolver took less than a tenth of a second to find the null space to approximate the invariant density. This is negligible in comparison with the cost of exhaustive stochastic simulation of the full system.

5.2.4. Conditioned path sampling using effective generators

Given an approximation of the effective generator of the slow variables, computed using the CMA or the QSSA, we can now employ the methodology of [16], as summarised in Section 1.1, to sample paths conditioned on their endpoints. This time, a full eigenvalue decomposition of the matrix $\mathcal{M} = \frac{1}{\rho}\mathcal{G} + I$ was computed, so that matrices V and D could be found with V unitary and D diagonal, with $\mathcal{M} = V^{-1}DV$. Then rows of $\mathcal{M}^r = V^{-1}D^rV$ can be efficiently and accurately computed, as required in (1) and (2).



(a) (b)

Figure 5: *(a) 8 trajectories of the slow variable $S = X_1 + 2X_2$ sampled conditioned on $S(0) = 20$, $S(10) = 195$, $S(t) \in \Omega = \{0, 1, \ldots, 500\} \forall t \in [0, 5]$ for the system (11) with parameters (5.2), using the CMA approximation of the effective generator. (b) The means and standard deviations of 100 paths sampled using the QSSA (blue plots) and CMA (red plots).*

Figure 5 presents results using this approach. An effective generator for the system (11) was computed for the domain $X_1 + 2X_2 = S \in \Omega = \{0, 1, \ldots, 500\}$, using both the QSSA and CMA, and then fed into the conditioned path sampling algorithm. Figure 5 (a) shows 8 samples of conditioned paths approximated using the CMA. Notice that as the transition time between the two favourable regions is relatively short compared with the length of the simulation, the time

of the transition varies greatly between the different trajectories. This indicates that we are producing trajectories with a fair reflection of what happens in a transition between these regions. Figure 5 (b) shows the means and standard deviations of 100 paths sampled for both methods of computing the effective generator. The QSSA, which overestimates the value of the second peak in the invariant density, has a higher mean than the CMA. This demonstrates again that errors in approximating the effective generator has a knock-on affect to applications such as conditioned path sampling.

The effective generator was computed on the domain $S \in [0, 500]$ for the path sampling, which took the CMA close to 5 seconds to approximate. The calculation of the probabilities in (1), and the full eigenvalue decomposition of the generator matrix on this domain, took around 50 seconds. After this, each path took around 350 seconds to sample.

5.3. An Example of the Nested CMA Approach

In this section, we will illustrate how the nested approach outlined in Section 4 can be applied. We will consider an example for which we know the invariant distribution of the slow variables. This gives us a way of quantifying any errors that we incur by applying the nested CMA and QSSA approaches.

$$\begin{aligned} R_1 &: \quad \emptyset \xrightarrow{k_1} X_1, \\ R_2 &: \quad X_3 \xrightarrow{k_2} \emptyset, \\ R_3 &: \quad X_1 \xrightarrow{\kappa} X_2, \\ R_4 &: \quad X_2 \xrightarrow{\kappa} X_1, \\ R_5 &: \quad X_2 \xrightarrow{\gamma} X_3, \\ R_6 &: \quad X_3 \xrightarrow{\gamma} X_2. \end{aligned} \tag{16}$$

We will consider this system in the following parameter regime:

$$k_1 V = 20, \qquad k_2 = 1, \qquad \kappa = 100, \qquad \gamma = 10. \tag{17}$$

As before, V denotes the volume of the well-mixed thermally-equilibrated reactor. In this regime, there are multiple different time scales on which the reactions are occurring. This is demonstrated in Figure 6, where there is a clear gap in the frequency of reactions R_1 and R_2 (the slowest), R_5 and R_6 (fast reactions) and R_3 and R_4 (fastest reactions).

This system was chosen as we are able to, using the results in [22], find the exact invariant distribution of the slow variable $S_1 = X_1 + X_2 + X_3$. In this instance, it is a Poisson distribution with mean parameter

$$\lambda = \frac{k_1 V}{k_2 \gamma \kappa} \left(\gamma k_2 + 3\gamma\kappa + 2k_2\kappa\right) = 64.2.$$

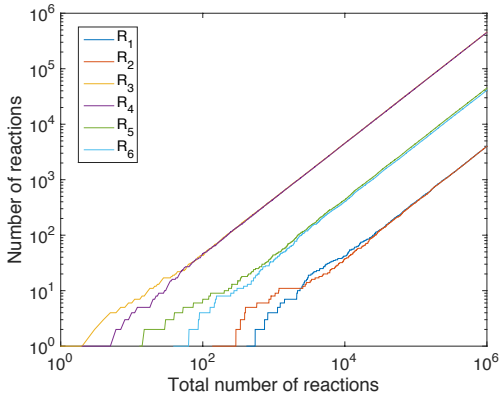


Figure 6: *Relative occurrences of the reactions R_1-R_6, for the system* (16) *with parameters* (17). *The most frequent reactions are reactions R_3 and R_4, reactions R_5 and R_6 are the next most frequent, with reactions R_1 and R_2 being the least frequent.*

5.3.1. QSSA-based analysis

One method to analyse such a system would be a nested QSSA-based analysis, similar to that which is suggested in [11]. In this paper the authors consider systems with reactions occurring on multiple timescales. If at first we consider all reactions R_3-R_6 to be fast reactions, then by applying the QSSA we are interested in finding the invariant distribution of the following fast subsystem:

$$\begin{aligned} C_1 &: \quad X_1 + X_2 + X_3 = S_1, \\ R_3 &: \quad X_1 \xrightarrow{\kappa} X_2, \\ R_4 &: \quad X_2 \xrightarrow{\kappa} X_1, \\ R_5 &: \quad X_2 \xrightarrow{\gamma} X_3, \\ R_6 &: \quad X_3 \xrightarrow{\gamma} X_2. \end{aligned} \tag{18}$$

Note that the quantity $S_1 = X_1 + X_2 + X_3$ is a conserved quantity with respect to these reactions, and as such is the slow variable in this system. This is in itself also a system with more than one timescale, and as such, we may want to iterate again and apply a second QSSA assumption, based on the fact that reactions R_3 and R_4 are fast reactions in comparison with reactions R_5 and R_6. This leads to a second fast subsystem:

$$\begin{aligned} C_1 &: \quad X_1 + X_2 + X_3 = S_1, \\ C_2 &: \quad X_1 + X_2 = S_2, \\ R_3 &: \quad X_1 \xrightarrow{\kappa} X_2, \\ R_4 &: \quad X_2 \xrightarrow{\kappa} X_1. \end{aligned}$$

Note that the quantity $S_2 = X_1 + X_2 = S_1 - X_3$ is a conserved quantity with respect to these reactions, and as such is the slow variable in this system. At this point in [11], the authors simulate the system using the Gillespie SSA. We could adopt the approach that we used in Section 5.2, in which we find the invariant distribution of the system by constructing its generator and then finding the normalised eigenvector corresponding to the null space of that operator. This would not be expensive since there are only S_2 different states. However, as in Section 5.1, as this system only contains monomolecular reactions, we can exactly find its invariant distribution. In this case, X_1 and X_2 follow a binomial distribution with mean $\bar{X_1}, \bar{X_2} = \frac{S_2}{2}$. This can then be used to compute the effective rate of reaction R_5 in the first subsystem (18), $\alpha_5(X_1, X_2) \approx \gamma \bar{X_2} = \frac{\gamma}{2} S$. This fast subsystem is then reduced to the following:

$$\begin{aligned} C_1 &: \quad X_1 + X_2 + X_3 = S_1 = S_2 + X_3, \\ C_2 &: \quad X_1 + X_2 = S_2, \\ R_5 &: \quad S_2 \xrightarrow{\gamma/2} X_3, \\ R_6 &: \quad X_3 \xrightarrow{\gamma} S_2. \end{aligned}$$

Note that we have completely eliminated the fast variables X_1 and X_2, and instead consider the slower variable $S_2 = X_1 + X_2$, with effective rate for R_5 given by the analysis above. This system is exactly solvable, and its invariant distribution is a gamma distribution with means given by $\bar{X_3} = \frac{S_1}{3}$ and $\bar{S_2} = \frac{2S_1}{3}$, found by computing the steady states of the mean field ODEs[22]. This in turn can be used to compute the effective rate of reaction R_2 in the full system, where we now lose all of the fast variables X_1, X_2, X_3 and instead wish to understand the dynamics of the slow variable $S_1 = X_1 + X_2 + X_3$, which is only altered by reactions R_1 and R_2. This system is given by the following:

$$\begin{aligned} R_1 &: \quad \emptyset \xrightarrow{k_1} S_1, \\ R_2 &: \quad S_1 \xrightarrow{k_2/3} \emptyset. \end{aligned}$$

Here the effective rate for R_2 has been found by using the approximation of the effective rate $\alpha_2(S_1) = k_2 \bar{X_3} = \frac{k_2}{3} S$.

5.3.2. CMA-based analysis

We will now go through the same procedure, but this time using the constrained subsystems instead of the fast subsystems as used in the previous section. There are 3 choices for the fast reactions, each involving two out of X_1, X_2 and X_3. Since X_1 is the product of a zeroth order reaction, it is preferable not to include this as one of the fast variables, and so we pick $\mathbf{F}_1 = [X_2, X_3]$. We then construct the constrained subsystem for this choice of slow and fast

variables:

$$\begin{aligned}
C_1 \quad &: \qquad X_1 + X_2 + X_3 = S_1, \\
R_2 \quad &: \qquad X_3 \xrightarrow{k_2} X_1, \\
R_3 \quad &: \qquad X_1 \xrightarrow{\kappa} X_2, \\
R_4 \quad &: \qquad X_2 \xrightarrow{\kappa} X_1, \\
R_5 \quad &: \qquad X_2 \xrightarrow{\gamma} X_3, \\
R_6 \quad &: \qquad X_3 \xrightarrow{\gamma} X_2.
\end{aligned} \tag{19}$$

Note that R_1 is removed, since it does not change the fast variables. R_2 is the only other reaction which has changes to its stoichiometry due the constrained stoichiometric projector. We have reduced the dimension of the system (due to the constraint $X_1 + X_2 + X_3 = \sigma$ for some $\sigma \in \mathbb{N}$), but we are still left with a multiscale system, which in theory could be computationally intractable for us to find the invariant distribution for, through funding the null space of its generator. Therefore, we can apply another iteration of the CMA to this constrained system.

Reactions R_3 and R_4 are the fastest reactions in the system, and therefore we pick our next slow variable that we wish to constrain to be $S_2 = X_1 + X_2$, which is invariant with respect to these reactions. Due to the previous constraint $S_1 = X_1 + X_2 + X_3$, we are only required to define one fast variable for this system. Both choices $F_2 = X_1, X_2$, are essentially equivalent, and so we pick $F_2 = X_1$. These choices lead us to the following second constrained system:

$$\begin{aligned}
C_1 \quad &: \qquad X_1 + X_2 + X_3 = S_1, \\
C_2 \quad &: \qquad X_1 + X_2 = S_2, \\
R_2 \quad &: \qquad \alpha_2(\mathbf{X}) = \begin{cases} k_2 X_3, & \text{if } X_2 > 0, \\ 0 & \text{otherwise,} \end{cases} \\
&\qquad\quad \boldsymbol{\nu}_2 = [1, -1, 0]^T, \\
R_3 \quad &: \qquad X_1 \xrightarrow{\kappa} X_2, \\
R_4 \quad &: \qquad X_2 \xrightarrow{\kappa} X_1.
\end{aligned} \tag{20}$$

Here $\boldsymbol{\nu}_i$ denotes the stoichiometric vector associated with reaction R_i, i.e. the vector which is added to the state $\mathbf{X}(t)$ if reaction R_i fires at time t. Notice that we now have two separate constraints, and as such reactions R_5 and R_6 now have zero stoichiometric vectors. Moreover, these constraints lead us to a somewhat unphysical reaction for R_2. The reactant for this reaction is X_3, but only X_2 and X_1 are affected by this altered reaction. In system (19) when reaction R_2 fires, we lose one X_3, and gain X_1. Therefore, both constraints within (20) have been violated. In order to reset these constraints, without changing the fast variable $F = X_3$, we arrive at the stoichiometry presented in (20). Note that we add the condition that this reaction can only happen if $X_2 > 0$, as we cannot have negative numbers of any species.

This is a closed system, with a very limited number of different states. Therefore, it is computationally cheap to construct its generator, and to find that generator's null space. Our aim with this system, is to find the invariant distribution of the fast variable given particular values for the constraints C_1 and C_2. This distribution will then allow us to compute the expectation of the reaction R_4 within the constrained system (6), which is the only reaction which is dependent on the results of the second constrained system (since $X_3 = S_1 - S_2$). Once the invariant distribution has been found, this can be used to find the effective propensity of reaction R_5 given values of $S_1 = X_1 + X_2 + X_3$ and $S_2 = X_1 + X_2$. In turn, the constrained system (19) can then be solved to find the invariant distribution on X_3 given a value of S_1. Finally, this leads us to the construction of an effective generator for the slow variable S_1.

Since this final constrained subsystem is aphysical, we cannot use the results of [22] to find the invariant distribution, and as such we must approximate them through finding the null space of the generator, as we did in Section 5.2

5.3.3. Comparison of approximation of invariant densities

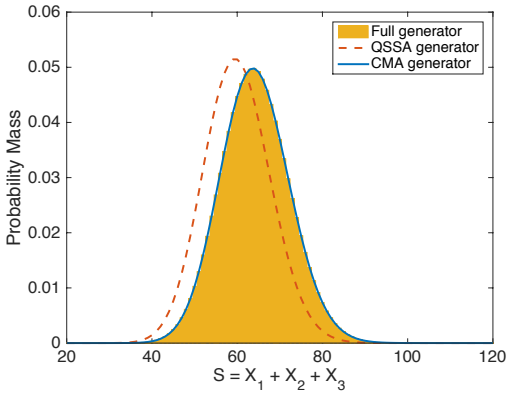


Figure 7: *Approximations of the invariant distribution of the slow variable $S = X_1 + X_2 + X_3$ of system* (16) *with parameters* (17)*, through marginalisation of the invariant distribution of the full system (histogram), of the effective generator computed using the CMA (solid line) and computed using the QSSA (dashed line).*

Figure 7 shows the invariant distributions of the slow variables $S = X_1 + X_2 + X_3$ computed by marginalising the invariant distribution of the full system, and from the CMA and QSSA as outlined above. The distribution computed using the CMA is indistinguishable by eye from the true distribution, and has a relative error of 5.936×10^{-12}, which can be largely attributed to rounding errors and error tolerances in the eigenproblem solvers. The QSSA approximation, on the other hand, has a significant relative error of 3.739×10^{-1}. This demonstrates again the substantial improvements in accuracy that we gain in using the constrained approach rather than one based on the QSSA. This is

delivered at no substantial additional computational effort. As in the previous two examples, the highly accurate effective generator approximated using the CMA can be used in a host of applications where the full generator could not, such as conditioned path sampling.

The CMA is more expensive in this example than the previous ones, as there are a very large number of small eigenvalue problems to solve. This is due to the fact that there are reactions of three species occurring on three different time scales. The generation of the CMA approximation of the effective generator took around 1240 seconds, and the subsequent approximation of the invariant distribution of the slow variables took just over half a second. This still pales into comparison with the cost of exhaustive stochastic simulation of the system. The savings would be even more pronounced in systems with multimodal invariant distributions where switches between the modes are rare.

6. Conclusions

In this paper, we presented a significant improvement and extension to the original constrained multiscale algorithm (CMA). Through constructing and finding the null space of the generator of the constrained process, we can find its invariant distribution without the need for expensive stochastic simulations. The CMA in this format can also be used not just to approximate the parameters of an approximate diffusion, but to approximate the rates in an effective generator for the slow variables.

In this paper we have not discussed how the slow and fast variables in these systems can be identified. In the simple examples presented, this is relatively straightforward. However in general, this is far from the case. If the high probability regions in the statespace are known a priori, or possibly identified through short simulations of the full system, then it is possible to identify which are the fast reactions in the system, and therefore what good candidates for the slow variable(s) could be. Other more sophisticated approaches exist, for example methods for automated analysis to identify the slow manifold[14, 28, 26]. One relatively ad hoc approach might be to briefly simulate the full system using the Gillespie SSA, which can give a good indication as to which the fast reactions are. Good candidates for slow variables are often linear combinations of the species who are invariant to the stoichiometry of the fast reactions, as we have seen in this paper. If the regions which the system is highly likely to spend the majority of its time are known, then looking at the relative values of the propensity functions, as we did in Figure 3 (b), can lead to an understanding of which reactions are fast and which are slow.

Through iterative nesting, the CMA can be applied to much more complex systems, as it can be applied repeatedly if the resulting constrained system is itself multiscale. This makes it a viable approach for a bigger family of (possibly biologically relevant) systems. This nested approach breaks up the original task of solving an eigenvalue problem for one large matrix per row of the effective generator, down into many eigenvalue solves for significantly smaller generators

for smaller dimensional problems, making the overall problem computationally tractable.

In the first example, we demonstrated that the CMA produces an approximation of the dynamics of the marginalised slow process in the system which is exact, at least by the measures that we have applied thus far, in the case of systems of monomolecular reactions. Since such systems are well understood, we were also able to compare this with the accuracy of the equivalent QSSA-based method, which incurred significant errors. We then applied the method of Fearnhead and Sherlock[16] to the approximate effective generators of the two approaches, in order to approximately sample conditioned paths of the slow variables. This task would be computationally intractable to attempt with the full generator for this system. This also demonstrated how the accuracies of the two approximations can impact the accuracy of any application for which they may be used.

In the second example, a more complex bistable system was also analysed using the CMA, and the invariant distribution of the computed effective generator was shown to be very close to the best approximation that we could make of the invariant distribution of the slow variables, using the null space of the original generator with as little truncation as we could sensibly manage with our computational resources. This was in stark contrast with the poor approximation which was computed using the equivalent QSSA-based approach. This highlighted again the improvement, at no or little extra cost, of using the constrained approach as opposed to the QSSA.

In the final example, we demonstrated how the constrained approach might be applied to a more complex example with multiple timescales. The algorithm can be applied iteratively in order to reduce the constrained subsystems themselves into a collection of easily solved one-dimensional problems. When comparing the invariant distributions of the approximate processes computed using the two approaches, the QSSA once again was incorrectly approximating the distribution of the fast variables conditioned on the slow variables, and so incurred significant errors. In contrast, the CMA produced an approximation to the invariant measure which was accurate up to 12 digits.

We showed how these effective generators can be used in the sampling of paths conditioned on their endpoints. Such an approach could be employed as a method to sample missing data within a Gibb's sampler when attempting to find the structure of a network that was observed[16]. This approach could also be used simply to simulate trajectories of the slow variables, in the same vein as [6] or [11]. In this instance, it would only be necessary to compute the column of the effective generator corresponding to the current value of the slow variables.

The constrained approach consistently significantly outperforms approximations computed using the more standard QSSA-based approach, and at negligible additional cost. Furthermore, in the limit of large separation of timescales, the constrained approach asymptotically approaches the QSSA approximation.

The computational savings that we make in using the CMA depends on the application with which we wish to use the effective generators. Similarly, if we wish to approximate the invariant distribution of the slow variables, then the

CMA will always be less costly than exhaustive stochastic simulation. This is because we are able to directly compute the invariant distribution, whereas in the simulation setting, to obtain the same statistics we would be required to compute a very long simulations.

If, on the other hand, we simply wish to use the CMA to compute a trajectory of the slow variables, then the savings will vary, based on the size of the chosen domain, and the relative differences in propensity of the fast and slow reactions in the relevant regions. If our aim is only to produce one relatively short trajectory, then it is possible that stochastic simulation will be more efficient than using the CMA. However this is such a trivial task, that any modeller wishing to do so what not consider invoking any approximations such as the QSSA or CMA.

There are many avenues for future work in this direction, not least its application to more complex biologically relevant systems. In particular, the treatment of systems where the effective behaviour of the slow variable(s) cannot be well approximated by a one-dimensional Markov process need to considered, for example systems which exhibit oscillations. Automated detection of appropriate fast and slow variables, and statistical tests for the validity of the approximation for different systems would be hugely beneficial. In the case of constrained systems which are deficiency zero and weakly reversible, using the results of [1] we can find the invariant distributions without even constructing the generator, and this could be a good direction to investigate.

Acknowledgements: The author would like to thank Kostas Zygalakis for useful conversations regarding this work. This work was funded by First Grant Award EP/L023989/1 from EPSRC.

[1] D. Anderson, G. Craciun, and T. Kurtz. Product-form stationary distributions for deficiency zero chemical reaction networks. *Bulletin of mathematical biology*, 72(8):1947–1970, 2010.

[2] A. Auger, P. Chatelain, and P. Koumoutsakos. R-leaping: Accelerating the stochastic simulation algorithm by reaction leaps. *The Journal of chemical physics*, 125(8):084103, 2006.

[3] A. Berezhkovskii, G. Hummer, and A. Szabo. Reactive flux and folding pathways in network models of coarse-grained protein dynamics. *The Journal of chemical physics*, 130(20):205102, 2009.

[4] P. Bolhuis and C. Dellago. 3 trajectory-based rare event simulations. *Reviews in Computational Chemistry*, 27:111, 2011.

[5] A. Bortz, M. Kalos, and J. Lebowitz. A new algorithm for Monte Carlo simulation of Ising spin systems. *Journal of Computational Physics*, 17(1):10–18, 1975.

[6] Y. Cao, D. Gillespie, and L. Petzold. The slow-scale stochastic simulation algorithm. *The Journal of chemical physics*, 122(1):014116, 2005.

[7] Y. Cao, H. Li, and L. Petzold. Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *The journal of chemical physics*, 121(9):4059–4067, 2004.

[8] S. Cotter and R. Erban. Error analysis of diffusion approximation methods for multiscale systems in reaction kinetics. *SIAM journal on Scientific Computing*, submitted.

[9] S. Cotter, K. Zygalakis, I. Kevrekidis, and R. Erban. A constrained approach to multiscale stochastic simulation of chemically reacting systems. *The Journal of chemical physics*, 135(9):094102, 2011.

[10] M. Cucuringu and R. Erban. Adm-cle approach for detecting slow variables in continuous time markov chains and dynamic data. *arXiv preprint arXiv:1504.01786*, 2015.

[11] W. E, D. Liu, and E. Vanden-Eijnden. Nested stochastic simulation algorithm for chemical kinetic systems with disparate rates. *The Journal of chemical physics*, 123(19):194107, 2005.

[12] N. Eidelson and B. Peters. Transition path sampling for discrete master equations with absorbing states. *The Journal of chemical physics*, 137(9):094106, 2012.

[13] R. Erban, S. Chapman, I. Kevrekidis, and T. Vejchodsky. Analysis of a stochastic chemical system close to a SNIPER bifurcation of its mean-field model. *SIAM Journal on Applied Mathematics*, 70(3):984–1016, 2009.

[14] R. Erban, T. Frewen, X. Wang, T. Elston, R. Coifman, B. Nadler, and I. Kevrekidis. Variable-free exploration of stochastic models: a gene regulatory network example. *The Journal of chemical physics*, 126(15):155103, 2007.

[15] R. Erban, I. Kevrekidis, D. Adalsteinsson, and T. Elston. Gene regulatory networks: A coarse-grained, equation-free approach to multiscale computation. *Journal of Chemical Physics*, 124:084106, 2006.

[16] P. Fearnhead and C. Sherlock. An exact Gibbs sampler for the Markov-modulated Poisson process. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):767–784, 2006.

[17] M. Gibson and J. Bruck. Efficient exact stochastic simulation of chemical systems with many species and many channels. *Journal of Physical Chemistry A*, 104:1876–1889, 2000.

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

[19] D. Gillespie. The chemical langevin equation. *The Journal of Chemical Physics*, 113(1):297–306, 2000.

[20] D. Gillespie. Approximate accelerated stochastic simulation of chemically reacting systems. *The Journal of Chemical Physics*, 115(4):1716–1733, 2001.

[21] A. Golightly and D. Wilkinson. Bayesian inference for markov jump processes with informative observations. *Statistical Applications in Genetics and Molecular Biology*, 2014.

[22] T. Jahnke and W. Huisinga. Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of mathematical biology*, 54(1):1–26, 2007.

[23] S. Kar, W. Baumann, M. Paul, and J. Tyson. Exploring the roles of noise in the eukaryotic cell cycle. *Proceedings of the National Academy of Sciences*, 106(16):6471–6476, 2009.

[24] M. Novotny. Monte carlo algorithms with absorbing markov chains: Fast local algorithms for slow dynamics. *Physical review letters*, 74(1):1, 1995.

[25] V. Rao and Y.W. Teh. Fast MCMC sampling for Markov jump processes and extensions. *The Journal of Machine Learning Research*, 14(1):3295–3320, 2013.

[26] M. Sarich and C. Schütte. Approximating selected non-dominant timescales by markov state models. *Comm. Math. Sci*, 10(3):1001–1013, 2012.

[27] M. Schena, D. Shalon, R. Davis, and P. Brown. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, 270(5235):467–470, 1995.

[28] A. Singer, R. Erban, I. Kevrekidis, and R. Coifman. Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *Proceedings of the National Academy of Sciences*, 106(38):16090–16095, 2009.

[29] A. Stuart. Inverse problems: a bayesian perspective. *Acta Numerica*, 19:451–559, 2010.

[30] S. Sun. Path summation formulation of the master equation. *Physical review letters*, 96(21):210602, 2006.

[31] T. Tian and K. Burrage. Binomial leap methods for simulating stochastic chemical kinetics. *The Journal of chemical physics*, 121(21):10356–10364, 2004.

[32] S. Trygubenko and D. Wales. Graph transformation method for calculating waiting times in markov chains. *The Journal of chemical physics*, 124(23):234110, 2006.

[33] J. Vilar, Hao Y. Kueh, N. Barkai, and S. Leibler. Mechanisms of noise-resistance in genetic oscillators. *Proceedings of the National Academy of Sciences*, 99(9):5988–5992, 2002.