Journal of Computational Physics 323 (2016) 265–282

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp

Constrained approximation of effective generators for multiscale stochastic reaction networks and application to conditioned path sampling

Simon L. Cotter [1]

School of Mathematics, University of Manchester, Oxford Road, Manchester, M13 9PL, United Kingdom

ARTICLE INFO

Article history:
Received 22 June 2015
Received in revised form 10 March 2016
Accepted 26 July 2016
Available online 29 July 2016

Keywords:
Stochastic
Multiscale
Chemical kinetics
Constrained dynamics

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.

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

E-mail address: simon.cotter@manchester.ac.uk.

[1] 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.

http://dx.doi.org/10.1016/j.jcp.2016.07.035

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 [7,17].

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 approach zero, and a number of different methods for overcoming this have been presented [2,31].

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.

1.1. 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