High-Dimensional Problems in Stochastic Modelling of Biological Processes

Shuohao Liao
St Hugh's College
University of Oxford

A thesis submitted in partial fulfilment of
the requirements for the degree of
Doctor of Philosophy
in the subject of
Mathematics

Hilary Term 2017

To my families

for their unending support and encouragement

Acknowledgments

I am deeply indebted to my supervisor, Professor Radek Erban, for his unwavering support and guidance. To Radek for introducing me to the world of stochastic modelling of biological processes as a master student, for suggestions with the mathematics, for confidence in my research, and for support in my career development. It has been an honour and pleasure working with him on the various papers we have produced and I hope we will continue to work together in the future.

The research leading to these results has received funding from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement no. 239870, and from Radek's Philip Leverhulme Prize. I 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 part of the work in this thesis was undertaken. This was supported by EPSRC grant no. EP/K032208/1. I would also acknowledge the China Scholarship Council for the 2015 Chinese Government Award for Outstanding Self-financed Students Abroad.

I would like to express my gratitude to Dr Tomáš Vejchodský who acted as a second supervisor to me during his one-year visit to Oxford. He has always been amazingly patient, knowledgeable and highly professional. I thank him for his generous hospitality during my visit to Prague, and helping me mature both academically and personally. I would also like to show my appreciation to Professor Philip Maini and Professor Ruth Baker for their discussions and suggestions at the early stage of my DPhil life.

Special thanks must go to the mathematicians who have contributed, directly or indirectly, to the final results of this thesis. In particular, Dr Ramon Grima and Dr Andrew Duncan for the enjoyable and immense collaborations on the topic of comparing the discrete and continuous stochastic models, Dr Simon Cotter for interesting discussions on the topic of the most-likely oscillation path and offering me access to his code, Dr Dmitry Savostyanov and Dr Sergey Dolgov for their helpful comments and suggestions in the practicalities of the tensor methods. I also take this opportunity to thank Professor Mike Giles and Dr.Sci. Boris Khoromskij for agreeing to examine my thesis.

Thank you to everyone at the WCMB, my cohort, colleagues and the staff for making the DPhil experience more fun and manageable. Especially, my officemates Aaron, Jess, Jochen, Linus and Paul for frequent tea breaks and all the moments of fun and relaxation. I also extend my gratitude to all my friends from Oxford and from home for being such an important source of distraction from the work.

Finally, I would like to thank my family for always being there at the end of the day, through highs and lows.

Abstract

Stochastic modelling of gene regulatory networks provides an indispensable tool for understanding how random events at the molecular level influence cellular functions. A common challenge of stochastic models is to calibrate a large number of model parameters against the experimental data. Another difficulty is to study how the behaviour of a stochastic model depends on its parameters, i.e. whether a change in model parameters can lead to a significant qualitative change in model behaviour (bifurcation). This thesis addresses such computational challenges by a tensor-structured computational framework.

After a background introduction in Chapter 1, Chapter 2 derives the order of convergence in volume size between the stationary distributions of the exact chemical master equation (CME) and its continuous Fokker-Planck approximation (CFPE). It also proposes the multi-scale approaches to address the failure of the CFPE in capturing the noise-induced multi-stability of the CME distribution. Chapter 3 studies the numerical solution of the high-dimensional CFPE using the tensor train and the quantized-TT data formats. In Chapter 4, the tensor solutions are applied to study the parameter estimation, robustness, sensitivity and bifurcation structures of stochastic reaction networks.

A Matlab implementation of the proposed methods/algorithms is available at http://www.stobifan.org.

Contents

Listing of figures

"For it is simply a fact of observation that the guiding principle in every cell is embodied in a single atomic association existing only in one (or sometimes two) copy – and a fact of observation that it results in producing events which are a paragon of orderliness [...] the situation is unprecedented, it is unknown anywhere else except in living matter."

What is Life? by Schrodinger (1944)

1 Introduction

1.1 Stochasticity in biological systems

1.1.1 Origin of stochasticity

Randomness is ubiquitous throughout biological processes. Intracellular biochemical processes ultimately boil down to enzyme-catalysed reactions as a result of intermolecular collisions with the right directions (orientation) and sufficient energy (Bowsher & Swain, 2012). These collisions are effectively unpredictable because the irregular diffusive moment of individual molecules are stochastic. It follows that the randomness or fluctuations in a system, from a fundamental result of theoretical statistical physics, is inversely proportional to the square root of the number of particles (Schrodinger, 1944). Thus, processes involving low copy numbers, such as transcription and translation, inherit high fluctuation. This noise at the molecular

level may propagate up to and influence functioning cells, tissues and organisms. This has been proved, with the help of numerous advances in experimental techniques, by observations of unavoidable randomness following gene expression, which reveals cell-to-cell heterogeneity in even isogenic populations (Raj & van Oudenaarden, 2008). Such heterogeneity due to the stochastic nature of molecular interactions is known as intrinsic noise (Raser & O'Shea, 2005).

Another source of heterogeneity can be loosely labelled environmental, or extrinsic (Volfson et al., 2006; Shahrezaei et al., 2008). It is less specific and usually refers to fluctuations in the amounts or states of other cellular components leading indirectly to variation in the expression of a particular gene. Examples of extrinsic source of noise include the number of RNA polymerase or ribosome, the stage in the cell cycle, the quantity of the protein, mRNA degradation machinery, and the cell environment. In general, fluctuations in upstream cellular components transmit as extrinsic noise in the expression of downstream genes.

The functional roles of noise are critical in biological processes. A direct function is that the biological noise drives the divergence of cell fates and enables cells to randomly transit between different states (Balázsi et al., 2011), like forming bacterial persistence (Balaban et al., 2004). Other examples of noise phenomena include stochastic resonance, where biological signal is periodically boosted by resonating with added noise (Douglass et al., 1993), and stochastic focusing, where cells utilize the noise to enhance sensitivities in a gradual response mechanism (Paulsson et al., 2000). Of course, noise also generates errors in DNA replications and thus allowing mutation and evolution to occur (Pennington & Rosenberg, 2007).

Yet cellular functions and processes are ordered and precisely regulated. So how do we explain the use, filtration and robustness to noise that is found in biological systems in the presence of random interactions among cellular components? The answers, based on a close examination of sources and consequences of noise, are

being sought through the use of stochastic models.

1.1.2 Stochastic modelling of biological systems

Stochastic modelling and simulation of biological systems have received increasing interest in the last decade, which broadly divide into two categories: spatial-temporal reaction-diffusion models and non-spatial (well-mixed) models. The spatial-temporal models emphasize on the systems involving spatial transport and biochemical reactions, which often span multiple time and length scales (Murray, 2001). Examples include the molecular diffusion within cellular micro-domains (Truskey et al., 2004), transportation of signalling molecules from cell membrane to nucleus (Simons & Toomre, 2000; Takahashi et al., 2010), and cellular migration in tumour invasion (Jiang et al., 2005). The microscopic molecular-based approaches seek to track individual positions and and trajectories of the molecules on a continuous domain (Erban & Chapman, 2009). The mesoscopic compartment-based approaches discretise the spatial domain into sub-volumes, or compartments, assuming spatial homogeneity within each compartment, and diffusion is modelled as reaction events between adjacent compartments (Fange et al., 2010). There are also spatially hybrid models that couple the molecular-based and the compartment-based models (Flegg et al., 2011).

The non-spatial stochastic models, the main interest of this thesis, rely on the "dilute" and "well-mixed" hypothesis (Gillespie et al., 2013). The well-mixed assumption requires the diffusive processes to be much faster than reactive ones, meaning that an overwhelming amount of elastic molecular collisions have taken place before any transitions of substances into products (Gillespie, 1976). The dilute assumption means that the reactant molecules are well separated with average distance sufficiently large compared to their diameters. To such extent, the position of molecules become uniformly randomized in the cell, and hence we can safely ignore the spatial organisation and concern ourselves only with the molecular populations (Gillespie,

2007).

One starting point of the non-spatial models, primarily through the work of McQuarrie (1967), is the Markov jump process formulated in the chemical master equation (CME). The CME consists of a system of coupled differential equations, and captures the variation of the probabilities that the system occupies each possible state. The all-encompassing nature of the equation does not lend itself easily to numerical solutions. A breakthrough was made by Gillespie (1976) for his Stochastic Simulation Algorithm (SSA), a Monte Carlo formulation that generates statistically exact trajectories identical to the original Markov process. The SSA simulates stochastic realisations through recursively generating the next reaction time and the next reaction, while the efficiency does not scale well with the complexity of the system. Variants of the SSA have been developed to improve the computational speed, including the next reaction method (Gibson & Bruck, 2000) and tau-leap method (Gillespie, 2001).

In the limit of large numbers of molecules, an approximated stochastic differential equation of the Langevin type can be derived from the SSA using the central limit theorem (Gillespie, 2000). The chemical Langevin equation (CLE) is much less formidable than the CME, because numerical solution of the CLE allows it to step over many reaction events in the SSA. The probabilistic feature of the CLE is characterised by the chemical Fokker-Planck equation (CFPE), that is obtained from the truncated Kramers-Moyal expansion at the second derivative term (Kramers, 1940; Moyal, 1949b). The CFPEs are generally more analytically tractable than the CME (Risken, 1984), and have been applied to analyse the bifurcation features of the stochastic models (Erban et al., 2009). Another widely-used mesoscopic description is the linear noise approximation (LNA) of the CME using the system-size expansion developed by van Kampen (Van Kampen, 1992). Although its accuracy positions midway between the CLE/CFPE and the deterministic models (Grima et al., 2011),

the LNA allows all statistical moments to be computed in close form and is suitable to study parameter dependence in the stochastic models (Komorowski et al., 2011). There are also many different multiscale methods tailored to simulate specific real applications (Pahle, 2009; Crudu et al., 2009; Liu et al., 2012).

Along with improving robustness and efficiency of the SSAs, a number of available self-contained simulation packages have been developed, including, StochKit (Sanft et al., 2011), INA (Thomas et al., 2012), COPASI (Hoops et al., 2006), STOCKS (Kierzek, 2002) and BioNetS (Adalsteinsson et al., 2004). However, the entire attention of these existing packages has centred on the Monte Carlo formalism of the non-spatial stochastic models, generating statistics from large ensembles of realisations. As such, features of the full probability distribution, as well as its parameter dependence, are not fully explored.

1.1.3 Parametric analysis of stochastic models

Stochastic models usually involve a large number of model parameters that need to be calibrated against the experimental data. This requires recursively evaluating the parameter dependence of the model behaviours, and comparing the model predictions with the experimental data. In the deterministic settings, predictions of a stochastic model are mainly incorporated within its probability distribution, and therefore, efficient simulations for the probability distributions over ranges of parameter values are favourable in many applications.

Large parameter sweeps of the full probability distribution, however, have been regarded as computational impractical, because both state space and parameter space are high-dimensional. In a model of stochastic networks, the dimension of the state space, $\Omega_{\boldsymbol{x}}$, is equal to the number of reacting molecular species, denoted by N. When an algorithm, previously working with deterministic steady states, is extended to stochastic setting, its computational complexity is typically taken to the power N.

Moreover, the exploration of the parameter space, $\Omega_{\boldsymbol{k}}$, introduces another multiplicative exponential complexity. Given a system that involves K uncertain parameters, the 'amount' of parameter combinations to be characterized scales equally with the volume of $\Omega_{\boldsymbol{k}}$, i.e. it is taken to the power K (Hey & Nielsen, 2007).

Such exponentially scaled complexity in dimensions largely constrains the analysis of the stochastic models, especially when the deterministic framework is routinely applied in the stochastic context. Parametric analysis, such as sensitivity analysis, parameter inference and bifurcation analysis, is usually, if not all, performed by analysing the mean behaviours through repeating intensive Monte Carlo simulations (Komorowski et al., 2011). Another consequence is that, since the master equation was set up for chemical systems (Delbrück, 1940; McQuarrie, 1963), much of the following research has been focused on reducing its all-encompassing nature rather than incorporating more realistic features (Shahrezaei et al., 2008).

Herein, this thesis propose a tensor framework that avoids the exponential scaling of complexity in dimensions when analysing the features of the full probability distributions of stochastic models. The tensor framework has a complexity which grows linearly with respect to the total number of dimensions. With such framework, parameter sweeps can be treated as auxiliary parametric dimensions and performed in a single tensor computation.

1.2 Tensor-based computations

The exponential scaling in dimensions, together with multiple phenomena discussed above, is known as the *curse of dimensionality* (Bellman, 1961). It refers many things including a universal feature of classic matrix-vector-based data format that the memory requirements and computational cost of basic arithmetic grow exponentially with the volume of the space (Oseledets & Tyrtyshnikov, 2009). Thus, if matrices and

vectors constitute the underlying infrastructure for solving and analysing the probability distributions of stochastic models, the problem of the curse of dimensionality is naturally inherited. To tackle this, our work discards the traditional matrix-vector representation, and applies the low-parametric, separable tensor representation in analysing stochastic models.

The terminology, tensor, is commonly employed by the numerical analysis community to specify a multidimensional array, upon which algebraic operations generalizing matrix operations can be performed. Although sharing the same name, this notation should not be confused with stress tensors in physics (Narasimhan, 1993) or tensor fields in mathematics (Sharafutdinov, 1994). An Nth-order tensor is an element of a tensor product of N vector spaces, i.e., the order of a tensor is the number of dimensions. In this representation, the entries in a Nth-order tensor are identified by a N-tuple of indices, e.g., $x_{i_1,i_2,\ldots,i_N}$. A zero-th order tensor is a scalar, a first-order tensor is a vector, a second-order tensor is a matrix.

Tensors are not new in stochastic analysis. When a probability distribution function f, say in dimension $N = 3$, is discretised on a grid, it naturally yields an third-order tensor $\boldsymbol{f}$, and the entry f_{i_1,i_2,i_3} houses the value of function f at nodal point $(x_{i_1},x_{i_2},x_{i_3})$. Traditionally, this full tensor has been usually stacked, or reshaped, into a long vector for storage and calculation purposes, but this approach has been dubbed the curse of dimensionality. While in tensor-based algorithms, the full tensor needs to be decomposed into and saved as single-dimensional components (factors), and operations on the original tensor are performed through manipulating the one-dimensional components (Kolda & Bader, 2009). In this way, the high cost of working in high dimensions is avoided, and the efficiency would vary with different tensor decompositions.

1.2.1 Tensor decompositions

The idea of tensor decompositions is built upon the concept of separation of variables (Bellman, 1961), and the idea has been successfully applied to solve multi-dimensional stationary and time-dependent partial differential equations (PDEs), see reviews in Khoromskij (2015); Khoromskaia & Khoromskij (2015). If a PDE in a function $f(x_1, x_2, \ldots, x_N)$ has the separable coefficients and the separable right-hand side, the technique of the separation of variables allows one to make a substitution of the form

$$f(x_1, x_2, \ldots, x_N) = f_1(x_1) f_2(x_2) \cdots f_N(x_N), \tag{1.1}$$

breaking the original PDE into N weakly coupled ordinary differential equations for functions $\{f_i(x_i)\}_{i=1}^N$. But not all classes of functions can be represented efficiently as a direct product. When the above form is not applicable, one may consider a natural extension via a combination of direct products as

$$f(x_1, x_2, \ldots, x_N) = \sum_{r=1}^{R} f_1^{(r)}(x_1) f_2^{(r)}(x_2) \cdots f_N^{(r)}(x_N). \tag{1.2}$$

The above formulation is usually referred as the *separated representation*, and the number of summands, R, controls the accuracy of the decomposition (1.2). By increasing R, the separated expansion could theoretically achieve arbitrary accuracy. Such separable form allows one to manipulate with the approximation of f with linear complexity with respect to the dimension.

The discrete counterpart of the separated representation (1.2) is known as canonical polyadic (CP) decomposition or CANDECOMP/PARAFAC (Hitchcock, 1927; Carroll & Chang, 1970). Given a full tensor, the CP decomposition is constructed via a method that generalises the matrix singular value decomposition in a way that

higher-order tensors are represented as sums of rank-one tensors. For low separation ranks, the CP representation is very attractive for its linear scaling of complexity in dimensions. On the downside, each linear algebra operation results in the same object in the CP representation but with larger separation rank, and computational cost grows. Minimizing the separation rank after each operation is essential to avoid the rank growing uncontrollable and the representation becoming untenable. However, the existing techniques for the rank truncation of the CP representation are not stable, and in fact, computing the CP rank is NP-hard (Håstad, 1990; Hillar & Lim, 2013) and the problem of finding the best rank-R approximation can be ill-posed (De Silva & Lim, 2008). Thus, although conceptually easy, the CP representation could be problematic in practice.

The Tucker format (Tucker, 1966) is an alternative for the canonical decomposition. It decomposes a tensor into a core tensor multiplied by a factor matrix in each dimension. Truncating the separation ranks in the Tucker format is based on the higher-order singular value decomposition (HOSVD) (De Lathauwer et al., 2000a) developed in statistics, and thus the format is more stable. One extension of the algorithm, the so-called reduced HOSVD (RHOSVD), introduced by Khoromskij & Khoromskaia (2009), also provides the stable rank truncation algorithm for the class of canonical tensors. Both numerical and analytical approximations of multivariate functions and operators using low-rank Tucker format were discovered based on the polynomial interpolation, see Khoromskij (2006). Such functional insight has motivated the application of the Tucker decomposition in various computational techniques, for example the multi-grid scheme (Khoromskij & Khoromskaia, 2009) and numerical solution of the large scale PDEs (Khoromskij, 2009b; Khoromskij et al., 2011). Despite the smaller size compared to the original tensor, the resulting core tensor in the Tucker decomposition still has a hyper-cube of parameters, and thus suffers from the curse of dimensionality. So the format is more suitable for small or

moderate dimensions.

An alternative approach is to recursively split out the dimensions of a tensor resulting in single-dimensional components. If the construction is progressed linearly, it results in the matrix product states (MPS) (Fannes et al., 1992a; Klümper et al., 1993; Perez-García et al., 2007) and the density matrix renormalisation group (DMRG) (White, 1992, 1993; Schollwöck, 2011b) in large scale quantum chemistry and theoretical physics computations. The procedure can also be conducted in accordance with a binary tree, yielding the so-called hierarchical Tucker (HT) decomposition (Oseledets & Tyrtyshnikov, 2009; Hackbusch & Kühn, 2009; Grasedyck, 2010), and more generally, a tensor tree network can be constructed (Fannes et al., 1992b,a; Hübener et al., 2010).

The linear structure of tensor decompositions was also discovered separately in the numerical analysis community, known as the *tensor train* (TT) decomposition (Oseledets, 2011). In terms of efficiency and stability, the TT format may be positioned between the CP decomposition and the Tucker decomposition, i.e., achieving linear scaling in dimensions while the problem of searching for minimal separation rank is still well defined (Khoromskij, 2012). Its applications in the physical community have motivated many computational algorithms to perform data approximations and other operations in the TT/MPS format. First, a full tensor can be represented in the TT format via the singular value decompositions with prescribed accuracy. Also, the concept of the alternating direction methods developed in quantum physics can be applied to the approximation, linear, or eigenvalue problems, where the standard methods may be applied to the components of the format (Holtz et al., 2012). The TT format is the workhorse tensor format being used in the current work, and will be discussed in details in Chapter 3.

Although the tensor decompositions were originally designed to separate the "physical" dimensions, the same framework may achieve a further cost reduction in ev-

ery single dimensions. A one dimensional tensor (vector) can be unfolded, or quantized, into a few virtual dimensions. Here, quantization suggests the entry index is described using binary variables rather than a single variable, reshaping a one-dimensional object into an nth-order tensor. Here, n equals the logarithm of total number of entries to base 2, and is usually referred as virtual dimensions. When these virtual dimensions are decomposed via TT framework, the storage requirement of the resulting representation scales logarithmically with the original size. Such representation is known as the Quantized TT (QTT) format (Khoromskij, 2009a, 2011; Oseledets, 2010, 2009). It is important that the rank bounds for many functional vectors in the QTT format has been estimated rigorously (Khoromskij, 2009a; Oseledets, 2013), and such theoretical justifications guarantee the logarithmic growth of the storage requirement with respect to the data volume. The QTT decomposition is also applicable to the compression of the matrices (Oseledets, 2010), which is useful in the discretisation of multi-dimensional PDEs over a find grid. It is thus a stable and efficient prototype for our implementations. We refer the readers to two recent survey papers, Khoromskaia & Khoromskij (2014a) and Khoromskij & Repin (2015), on various applications of the QTT format. Recently, the QTT has been applied to decompose the factor matrices in the Tucker decomposition, which results in the two-level QTT-Tucker format (Dolgov & Khoromskij, 2013b).

1.2.2 Tensor-based computational biology

The first step toward tensor business in computational biology was kicked off by a proliferation of high-throughput data. Tensor modelling of large-scale molecular biological data has been used to statistically infer the connectivity structure of the biological networks (Omberg et al., 2007; Alter & Golub, 2005; Acar et al., 2007). Other examples include predicting causal coordination between eukaryotic DNA replication origin activity and mRNA expression (Alter & Golub, 2004), as well as predicting the

cancer's formation and growth (Lee et al., 2012; Sankaranarayanan et al., 2015).

However, the development of tensor-based methodologies to tackle the curse of dimensionality in mechanistic stochastic models is lagging far behind its development in statistical models. Only until very recently, several attempts were made to apply tensor-based algorithms to simulate the chemical master equations. Hegland & Garcke (2011) and Ammar et al. (2012) attempted to represent the CME as a sum of rank one tensors, and the solution is sought in the form of a CP decomposition. The approach required recomputing a lower rank CP representation at each step of the algorithm, but there is lack of robust algorithms for such a purpose. To avoid the problem of approximation in the CP format, Jahnke & Huisinga (2008) proposed to use the Dirac-Frenkel principle and project the density function onto a manifold composed of suitably chosen basis functions with predetermined maximal CP rank. But a priori estimating the tensor rank of the CME solution is not possible. More recently, two separate trails (Kazeev et al., 2014; Dolgov & Khoromskij, 2015) attempted to apply the TT/QTT format to direct solution of the CME. Still, applying the tensor-based computational methods for analysing stochastic models is in its infancy compared to its applications in statistical analysis.

One reason for less advancement in applying tensor-based method to stochastic modelling is that the decompositions need to be sought "implicitly". In statistical modelling, a data cube is usually provided explicitly and tensor decompositions may be directly applied, while in stochastic modelling, we never handle a high-dimensional data cube, but construct approximate equation solvers in a separated representation and compute the tensor decomposition as the solution of the equation. This means the separability of the solution, like probability distribution functions, is "implicitly" embedded in the separability of the model equations. Thus, if the system is destined to be high-dimensional, it is important to model the system in a tensor-friendly manner.

It also requires many algorithms and operations involved in the model simulation to be translated into the tensor framework. Elementary linear algebra operations, such as addition and matrix-vector multiplication, are straightforward and have been well studied (Kolda & Bader, 2009), but tensorising other procedures are not as simple, such as discretisation methods, time integration methods, solving linear systems, and eigenvalue solvers. Although there has been increasing interest in these topics in recent years (Kazeev & Khoromskij, 2012; Kazeev et al., 2012; Holtz et al., 2012; Dolgov et al., 2014), the methods are still far from being robust comparing to the traditional matrix-based algorithms. Especially, lack of a robust tensor linear solver may have a detrimental effect on the computation of the time-dependent and stationary solutions, since iteratively solving linear systems is usually required (Liao et al., 2015). This issue is investigated in details in Chapter 3.

Beside developing tensor-based numerical simulation algorithms, this dissertation also follows a tensor-based computational thinking in stochastic modelling. The tensor-based computational thinking requires i) an ability to reason at the block matrix level to expose tensor product-like operations, and ii) an ability to reason at the multi-index level about the evaluation of the functional evaluation and the order of its operations. With hindsight it may seem very natural, but this tensor perspective is difficult to achieve, and it helps to addresses critical questions in tensor-based analysis of stochastic modelling. The questions include: what class of stochastic models can be represented efficiently in the separated tensor format? which types of extra features can be added to an existing model while keeping its separability? does the solution admit a tensor approximation with a low separation rank? how can parametric analysis of the stochastic models be performed in the tensor framework? what information can be efficiently derived from the tensor formatted solutions? We explore these topics in chapters 4 and 5.

1.3 Organization of the thesis

The thesis has three content chapters: i) analysis of the modelling error of the CFPE; ii) numerical solution of the CFPE in the tensor format; and iii) application of the tensor method to parametric analysis of the stochastic models. The contents in each chapter are relatively distinct and can be read in isolation. A continuous thread runs through the thesis from modelling analysis to simulation algorithm and then to biological applications.

In Chapter 2, we first give a brief review of the existing non-spatial models of the stochastic reaction networks. Then, we study the convergence of the stationary distributions of the CFPE and the CME in the limit of large volume. Given a system of N chemical species, we show that the accuracy of the CFPE is of order $\mathcal{O}\left(V^{-(N+1)/2}\right)$ for general mass-action reaction systems, and is of order $\mathcal{O}\left(V^{-(N+2)/2}\right)$ for systems satisfying the detailed balance. We will also study of the failure of the CFPE in capturing the noise-induced multi-stability, and address it by using the multi-scale methods, including the quasi-stationary approximation and the hybrid CME-CFPE methods.

Chapter 3 studies the numerical solution of the high-dimensional CFPE in the tensor formats. For numerical discretisation, we first apply the theorem of generalised principle eigenvalues to study the effect of artificial domain truncation. Given a truncated domain, we propose a second-order monotone finite difference scheme. Under certain conditions, the assembled stiffness matrix preserves the non-positivity of the CFPE operator. Particularly, the matrix can be constructed exactly in a low-parametric, separated tensor matrix representation, i.e., the sum of rank-one tensor matrices, which can be subsequently converted into the QTT format. We will also show the conditions when the principle eigenvector admits a low-rank structure. Under those conditions, we present the (adaptive) inverse power scheme and

the multi-grid acceleration to solve the principle eigenvalue problem in the tensor structure. Finally, we show how to combine the parameter space with the state space in the tensor formulation such that the steady state distributions of the CFPE with ranges of parameter values can be solved in a single computation.

Chapter 4 exemplifies the tensor framework to perform parametric analysis of the stochastic networks, including parameter estimation, stochastic bifurcation analysis, and robustness analysis. We adopt the classic methodologies for parametric analysis, while the underlying algebraic operations are all performed in the QTT format. In this way, parametric analysis of the stochastic models could benefit from the efficient high-dimensional tensor operations.

Chapter 5 summarises the author's key contributions, and discusses some possible directions for future research.

"Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better."

Cohen (2004)

2

Stochastic modelling of biological processes

This chapter aims at analysing the modelling error of the Fokker-Planck approximation of the chemical master equation. Starting out from the background introduction of the modelling of mass-action reaction systems, we studies the convergence of the CFPE distribution towards the CME distribution in the large volume limit. We also propose the quasi-stationary approximation (QSA) and the hybrid CME-CFPE equation to address the failure of the CFPE in capturing the noise-induced multi-stability.

2.1 Modelling the chemical reaction network

A well-mixed chemical reaction system, confined to a constant volume V, consists of $N \geq 1$ chemical species $\{S_1, \ldots, S_V\}$ that undergoes a finite number, M, of chemical

reactions. For the j-th reaction, we denote the vectors $\boldsymbol{\nu}_j^-, \boldsymbol{\nu}_j^+ \in \mathbb{N}^N$ by the amounts of molecules of each chemical species (substance) that are consumed and created in one instance of the reaction, respectively. We write the short form of the j-th reaction as $\boldsymbol{\nu}_j^- \to \boldsymbol{\nu}_j^+$, or in an elementary form as

$$\nu_{j,1}^- S_1 + \cdots + \nu_{j,N}^- S_N \overset{k_j}{\to} \nu_{j,1}^+ S_1 + \cdots + \nu_{j,N}^+ S_N, \tag{2.1}$$

where k_j denotes the j-th reaction rates. The net effect, $\boldsymbol{\nu}_j = \boldsymbol{\nu}_j^+ - \boldsymbol{\nu}_j^-$, in (2.1) is termed as the j-th *state changing vectors*, and matrix $\boldsymbol{\nu} = (\boldsymbol{\nu}_1, \ldots, \boldsymbol{\nu}_M)$ is referred as the *stoichiometry matrix*.

Definition 2.1. *Let $\mathscr{S}_S = \{S_i\}_{i=1}^N$, $\mathscr{S}_{\boldsymbol{\nu}} = \{\boldsymbol{\nu}_j\}_{j=1}^M$ and $\mathscr{S}_R = \{\boldsymbol{\nu}_j^- \to \boldsymbol{\nu}_j^+\}_{j=1}^M$ denotes the sets of the species, stoichiometric vectors, and reactions, respectively. The triple $\{\mathscr{S}_S, \mathscr{S}_{\boldsymbol{\nu}}, \mathscr{S}_R\}$ is called a chemical reaction network.*

The notion of a reaction network is independent of the model descriptions, and a choice between the stochastic and deterministic models is made based upon the features of the problem at hand. The stochastic model is used when the molecular number is low on average, while the deterministic model is used when the system is close to the thermodynamic limit.

Since we are interested in the chemical species whose populations are changed by one or more reactions in (2.1), we always make the following assumption on the stoichiometric coefficients throughout the rest of the thesis.

Assumption 2.2. *The sets of stoichiometric vectors, $\mathscr{S}_{\boldsymbol{\nu}}$, of a chemical reaction network satisfies $\sum_{j=1}^M |\nu_{j,i}| > 0$ for $i = 1, \ldots, N$.*

In what follows we will present the stochastic formulations of the chemical reaction network.

2.1.1 The chemical master equation

A fundamental premise of stochastic chemical kinetics is the so-called *propensity function*, that specifies the probabilistic feature of the chemical reactions (Gillespie, 1977). For the j-th reaction, we write the propensity function as $\alpha_j(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathbb{N}^N$ is a *state vector* representing the number of molecules of each species. Suppose that the j-th reaction is of mass-action form like in (2.1), the propensity function takes the form

$$\alpha_j(\boldsymbol{x}) = k_j \prod_{i=1}^{N} V^{(-\nu_{j,i}^- + 1)} \prod_{l=1}^{\nu_{j,i}^-} (x_i - (l-1)). \tag{2.2}$$

The above formula is derived out of molecular physics (Gillespie, 1977), and the underlying physical rationale can be briefly summarised as: a randomly chosen combination of $\{S_i\}_{i=1}^N$ will react according to reaction $(\boldsymbol{\nu}_j^- \to \boldsymbol{\nu}_j^+)$ in (2.1) with probability $\alpha_j(t)\mathrm{d}t$ in the next infinitesimal time interval $[t, t+\mathrm{d}t)$. If $(\boldsymbol{\nu}_j^- \to \boldsymbol{\nu}_j^+)$ occurs, the state vector is modified by the j-th stoichiometric vector as

$$\boldsymbol{x}(t+\mathrm{d}t) = \boldsymbol{x}(t) + \boldsymbol{\nu}_j.$$

More generally, let $r_j(t)$ denote the number of firings of the j-th reaction occurs within $[0, t)$, given the initial molecular population $\boldsymbol{x}(0)$, the evolution of the state vector is given by

$$\boldsymbol{x}(t) = \boldsymbol{x}(0) + \sum_{j=1}^{M} r_j(t)\boldsymbol{\nu}_j.$$

The process $r_j(t)$ is a Poisson counting process with α_j being the intensity, i.e.,

$$r_j(t) = \mathscr{P}_j\left(\int_0^t \alpha_j(\boldsymbol{x}(s))\mathrm{d}s\right),$$

where $\mathscr{P}_j(\cdot)$ are independent, unit-rate Poisson processes. Hence, the state vector

$\boldsymbol{x}(t)$ follows a continuous-time, discrete-space Markov chain given by

$$\boldsymbol{x}(t) = \boldsymbol{x}(0) + \sum_{j=1}^{M} \boldsymbol{\nu}_j \mathscr{P}_j \left(\int_0^t \alpha_j(\boldsymbol{x}(s)) \mathrm{d}s \right). \tag{2.3}$$

The process (2.3) may be simulated by the Gillespie stochastic simulation algorithm (SSA) (Gillespie, 1976), as well as its equivalent formulations (Gibson & Bruck, 2000; Gillespie, 2001), and we refer the reader to a recent review (Gillespie, 2007) for details.

The chemical master equation exactly describes the evolution of probability for such a Markov chain process. Let $P_{\mathrm{m}}(\boldsymbol{n}, t|\boldsymbol{n}_0, 0)$, or in short $P_{\mathrm{m}}(\boldsymbol{n}, t)$, be the probability that the system, initiated from the initial state $\boldsymbol{x}(0) = \boldsymbol{n}_0$, reaches the state $\boldsymbol{x}(t) = \boldsymbol{n}$ at time t. The CME can be written as

$$\begin{cases} \frac{\partial P_{\mathrm{m}}(\boldsymbol{n},t)}{\partial t} = \mathscr{L}_{\mathrm{m}} P_{\mathrm{m}}(\boldsymbol{n}, t), \quad \boldsymbol{x} \in \mathbb{N}^N, \ \ t > 0, \\ P_{\mathrm{m}}(\boldsymbol{n}, t) \geq 0, \quad \sum_{\boldsymbol{n} \in \mathbb{N}^N} P_{\mathrm{m}}(\boldsymbol{n}, t) = 1, \end{cases} \tag{2.4}$$

with initial condition $P_{\mathrm{m}}(\boldsymbol{x}, 0)$, where $\mathscr{L}_{\mathrm{m}}$ is the CME operator defined as

$$\mathscr{L}_{\mathrm{m}} P_{\mathrm{m}}(\boldsymbol{x}, t) = \sum_{j=1}^{M} \left\{ \mathscr{E}^{-\boldsymbol{\nu}_j} \left[\alpha_j(\boldsymbol{x}) P_{\mathrm{m}}(\boldsymbol{x}, t) \right] - \alpha_j(\boldsymbol{x}) P_{\mathrm{m}}(\boldsymbol{x}, t) \right\}.$$

The step operator $\mathscr{E}^{-\boldsymbol{\nu}_j}$, when applied to any function $f(\boldsymbol{n})$, modifies the argument by $-\boldsymbol{\nu}_j$ and yields $f(\boldsymbol{n} - \boldsymbol{\nu}_j)$.

The CME (2.4) was derived from the probability laws for independent and mutually exclusive events (Gillespie, 1992). The meaning of the equation can be interpreted as: the instantaneous rate of change of the probability of the system being in a particular state is the difference between the inflow and outflow probabilities induced by each chemical reactions. After evolving sufficiently long time, the proba-

bility density can be approximated by the stationary CME, written as

$$\begin{cases} \mathscr{L}_{\mathrm{m}}\pi_{\mathrm{m}}(\boldsymbol{n}) = 0, \quad \boldsymbol{n} \in \mathbb{N}^N \\ \pi_{\mathrm{m}}(\boldsymbol{n}) \geq 0, \quad \sum_{\boldsymbol{n}\in\mathbb{N}^N} \pi_{\mathrm{m}}(\boldsymbol{n}) = 1, \end{cases} \tag{2.5}$$

where $\pi_{\mathrm{m}}(\boldsymbol{x}) = \lim_{t\to\infty} P_{\mathrm{m}}(\boldsymbol{x}, t)$.

The CME (2.4) treats each possible state of the system as a separate variable, thus it is an infinite-size ODE system. Although the analytical solution can be obtained under certain scenarios (Jahnke & Huisinga, 2007), solving the CME exactly is intractable in general. But, for numerical approximations, it is possible to truncate the positive orthant by imposing a sufficiently large maximum copy number of each chemical species based on the reachability of states (Munsky & Khammash, 2006).

2.1.2 The reaction rate equation

It can be shown that an appropriate volume scaling of the Markov process (2.3) in the thermodynamic limit would yield an continuous evolution process for molecular concentrations (Kurtz, 1972)

$$\boldsymbol{\phi}(t) = \boldsymbol{\phi}(0) + \sum_{j=1}^{M} \boldsymbol{\nu}_j \int_0^t f_j(\boldsymbol{\phi}(s))\mathrm{d}s, \tag{2.6}$$

where $\boldsymbol{\phi} = (\phi_1, \phi_2, \ldots, \phi_N)^T$ is a state vector of molecular concentrations, and the macroscopic rate function f_j for a mass-action reaction system (2.1) is defined by

$$f_j(\boldsymbol{\phi}) = k_j \prod_{i=1}^{N} \phi_i^{\nu_{j,i}^-}. \tag{2.7}$$

Then, the deterministic *reaction rate* (RE) equation can be obtained by taking the time derivative to both sides of (2.6), which gives a system of ODE

$$\frac{\mathrm{d}\phi_i}{\mathrm{d}t} = \sum_{j=1}^{M} \nu_{j,i} f_j(\boldsymbol{\phi}), \tag{2.8}$$

for $i = 1, 2, \ldots, N$.

2.1.3 The Fokker-Planck approximation of the CME

In reality, many biological processes happen in an intermediate regime between the CME description and the deterministic description. In this regime, the intrinsic noise is still playing a crucial role, while the discreteness is less important in the noise contributions. Therefore, a continuous model would be more effective in the noise analysis of the stochastic networks, and we introduce a continuous Fokker-Planck approximation of the CME (2.4), as well as its equivalent formulation the chemical Langevin equation.

2.1.3.1 The Chemical Fokker-Planck Equation

The original derivation of the chemical Fokker-Planck equation (CFPE) was proposed by Kramers (1940) and Moyal (1949b,a) through a Taylor series expansion of the CME (2.4) as

$$\begin{aligned}\frac{\mathrm{d}P_{\mathrm{km}}(\boldsymbol{x},t)}{\mathrm{d}t} = \sum_{i=1}^{\infty}(-1)^i \sum_{\substack{c_1,c_2,\ldots,c_N=0 \\ c_1+c_2+\cdots+c_N=i}}^{i} \frac{1}{c_1!c_2!\cdots c_N!} \frac{\partial^i}{\partial x_1^{c_1}\partial x_2^{c_2}\cdots\partial x_N^{c_N}} \\ \times \left\{ \left[\sum_{j=1}^{M} (\nu_{j,1}^{c_1}\nu_{j,2}^{c_2}\cdots\nu_{j,N}^{c_N})\alpha_j(\boldsymbol{x}) \right] P_{\mathrm{km}}(\boldsymbol{x};t) \right\},\end{aligned} \tag{2.9}$$

where $P_{\mathrm{km}}(\boldsymbol{x},t)$, $\boldsymbol{x} \in \mathbb{R}^N$, can be regarded as a smooth reconstruction of the discrete CME solution $P_{\mathrm{m}}(\boldsymbol{n},t)$, $\boldsymbol{n} \in \mathbb{N}^N$, in the positive orthant but with an extension to cover the negative orthant. Equation (2.9) is the so-called Kramers-Moyal equation, and a

truncation after its second order derivative terms gives the CFPE (2.10).

Formally, we write the CFPE as

$$\begin{cases} \frac{\partial P_{\mathrm{f}}(\boldsymbol{x},t)}{\partial t} = \mathscr{L}_{\mathrm{f}} P_{\mathrm{f}}(\boldsymbol{x},t), \quad \boldsymbol{x} \in \Omega_{\mathrm{f}}, \ \ t > 0 \\ P_{\mathrm{f}}(\boldsymbol{x},t) \geq 0, \quad \int_{\Omega_{\mathrm{f}}} P_{\mathrm{f}}(\boldsymbol{x},t) \mathrm{d}\boldsymbol{x} = 1, \end{cases} \tag{2.10}$$

with initial condition $P_{\mathrm{f}}(\boldsymbol{x},0)$, where $\mathscr{L}_{\mathrm{f}}$ is the Fokker-Planck operator defined as

$$\mathscr{L}_{\mathrm{f}} P_{\mathrm{f}}(\boldsymbol{x},t) = \nabla \cdot \left[\nabla (\boldsymbol{D}(\boldsymbol{x}) P_{\mathrm{f}}(\boldsymbol{x},t)) - \boldsymbol{\mu}(\boldsymbol{x}) P_{\mathrm{f}}(\boldsymbol{x},t) \right],$$

where the diffusion matrix $\boldsymbol{D} \equiv \left(D_{i,j}\right)_{N\times N}$ and the drift vector $\boldsymbol{\mu} \equiv \left(\mu_1, \mu_2, \ldots, \mu_N\right)^T$ are defined by

$$D_{i,j}(\boldsymbol{x}) = \frac{1}{2} \sum_{\ell=1}^{M} \nu_{\ell,i} \nu_{\ell,j} \alpha_\ell(\boldsymbol{x}) \quad \text{and} \quad \mu_i(\boldsymbol{x}) = \sum_{\ell=1}^{M} \nu_{\ell,i} \alpha_\ell(\boldsymbol{x}), \quad i,j = 1,2,\ldots,N. \tag{2.11}$$

We assume that within the domain of definition $\Omega_{\mathrm{f}} \subset \mathbb{R}^N$ the following ellipticity condition is satisfied,

$$\sum_{i,j=1}^{N} D_{i,j}(\boldsymbol{x}) \xi_i \xi_j \geq 0, \quad \forall \boldsymbol{x} \in \Omega_{\mathrm{f}}, \ \text{and} \ \ \boldsymbol{\xi} = (\xi_1, \ldots, \xi_N)^T \neq \boldsymbol{0}, \tag{2.12}$$

that is the diffusion matrix $\boldsymbol{D}$ is positive semi-definite.

The CFPE (2.10) describes how likely the system will be in a certain portion of the state space, rather than a particular integer state in the CME (2.4). So a major and important difference between the CME and the CFPE is that x_i is a non-negative integer for the CME and a real number for the CFPE. Often the stationary probability distribution is of particular interests, it can be obtained by solving the stationary

CFPE as

$$\begin{cases} \mathscr{L}_{\mathrm{f}}\pi_{\mathrm{f}}(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \Omega_{\mathrm{f}}, \\ \pi_{\mathrm{f}}(\boldsymbol{x}) \geq 0, \quad \int_{\Omega_{\mathrm{f}}} \pi_{\mathrm{f}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1, \end{cases} \tag{2.13}$$

where $\pi_{\mathrm{f}}(\boldsymbol{x}) = \lim_{t\to\infty} P_{\mathrm{f}}(\boldsymbol{x}, t)$.

It is worth mentioning that the perfunctory truncation of the Taylor expansion (2.9) does not lend us any information about the domain of definition of the CFPE, Ω_{f}, as well as the associated boundary conditions. It is usually assumed that the system lies in the thermodynamic limit for the validity of the CFPE, and in such case the effect of the boundary conditions at zero are negligible. And for numerical approximation, we will apply the Dirichlet boundary conditions to any computational domain (Grima et al., 2011).

2.1.3.2 The chemical Langevin equation

The *chemical Langevin equation* describes the underlying Markov process whose time evolution of probability distribution satisfies the CFPE (2.10). It can be written in the form of a system of N Itö stochastic differential equations (SDEs)

$$\begin{cases} \mathrm{d}\boldsymbol{x}(t) = \boldsymbol{\mu}(\boldsymbol{x}(t))\mathrm{d}t + \boldsymbol{\sigma}(\boldsymbol{x}(t))\mathrm{d}\boldsymbol{w}, \quad t > 0 \\ \boldsymbol{x}(0) = \boldsymbol{x}_0, \end{cases} \tag{2.14}$$

where $\boldsymbol{x} \in \mathbb{R}^N$ is the state vector of molecular populations, $\boldsymbol{\mu} \in \mathbb{R}^N$ is the drift vector defined in (2.11), $\boldsymbol{\sigma} \in \mathbb{R}^{N\times M_{\mathrm{cle}}}$ is a diffusion matrix, and $\boldsymbol{w}$ is an M_{cle}-dimensional Wiener process. The matrix $\boldsymbol{\sigma}$ should satisfy (Wilkinson, 2011)

$$\boldsymbol{\sigma}(\boldsymbol{x})\boldsymbol{\sigma}^T(\boldsymbol{x}) = \boldsymbol{D}, \tag{2.15}$$

where $\boldsymbol{D}$ is the diffusion matrix defined in (2.11).

2.2 How accurate is the Fokker-Planck approximation of the chemical master equation?

In this section, we investigate the order of convergence of the CFPE to the CME in the limit of large volumes. Specifically, our main interest here is to derive the leading order error in the difference of the distribution,

$$\sup_{\boldsymbol{n}\in\mathbb{N}^N} |\pi_{\mathrm{m}}(\boldsymbol{n}) - \pi_{\mathrm{f}}(\boldsymbol{n})|, \tag{2.16}$$

where π_{m} is the stationary solution of the CME (2.5) and π_{f} is the stationary solution of the CFPE (2.13).

Some related references have attempted to approach the problem (2.16) from difference angles. For example, Kurtz (1978) proved that the difference between the jump and continuous Markov processes is of order $\log(V)/V$. Grima et al. (2011) used system-size expansion to prove that the CFPE predictions of the mean and the variance are accurate to order $V^{-3/2}$. However, the error between the steady state distributions of the two equations remains not clear. Since our analysis of the stochastic reaction networks later on will be mainly based on the probability distribution, approximated by solving the underlying CFPE, understanding the behaviours of the error in (2.16) is more relevant and important.

Our approach is motivated by the reference, Thomas & Grima (2015). The paper proposed a perturbative analysis of the CME of the reaction networks with single chemical species, i.e., $N = 1$, and the CME distribution is sought in a series expansion in the inverse powers of the square root of volume V. We generalise the approach to the reaction networks with multiple dimensional cases, in which case additional technicalities are involved.

It is worth mentioning that, although the CFPE was originally derived from the

Kramers-Moyal expansion (2.9), the error analysis is heavily based on the system size expansion. We discuss in Section § 2.2.1 that why the Kramers-Moyal expansion does not lend us necessary information about the convergence between the two equations.

The derivation of the leading order errors in (2.16) is divided into three steps. The first step will be presented in Appendix A, where we solve a perturbative solution of the stationary CME. This is based on the system size expansion of the CME and transform the distribution of molecular populations into the distribution of the fluctuations around the deterministic mean. We perform an asymptotic analysis to the distribution of fluctuations, and the expanded coefficients are analytically solved using the Hermite expansions and the eigenfunction expansion. In the second step in Appendix B, we apply the same perturbation analysis to the CFPE, and obtain a perturbative solution of the stationary CFPE. Then, given the perturbative solutions of the both stationary CME and the stationary CFPE, the last step in Section § 2.2.2 derives the leading order error by comparing the expanded coefficients in both equations.

2.2.1 The problem with the Kramers-Moyal expansion

As mentioned in Section § 2.1.3.1, the CFPE approximation is obtained by a second-order truncation of the Taylor expansion of the CME. Thus, a natural approach to show the convergence of the CFPE distribution to the CME distribution is to introduce the volume scaling in the expansion in the Kramers-Moyal equation (2.9), and prove the validity of the second-order truncation in the large volume limit.

Following Gillespie (2000), a "concentration" form of the Kramers-Moyal expansion can be obtained via inserting the molecular concentrations $\boldsymbol{\phi} = \boldsymbol{n}/V$ into (2.9), and the

result is

$$\frac{\partial \tilde{P}_{\mathrm{km}}(\boldsymbol{\phi};t)}{\partial t} = \sum_{i=1}^{\infty}\sum_{j=1}^{M}(-1)^i\left(\frac{1}{V}\right)^{i-1}\sum_{\substack{c_1,c_2,\ldots,c_N=0\\ c_1+c_2\cdots+c_N=i}}^{i}\frac{\boldsymbol{\nu}_j^{\boldsymbol{c}}}{\boldsymbol{c}!}\left(\frac{\partial}{\partial\boldsymbol{\phi}}\right)^{\boldsymbol{c}} f_j(\boldsymbol{\phi})\tilde{P}_{\mathrm{km}}(\boldsymbol{\phi};t), \tag{2.17}$$

where f_j is the deterministic rate function of the j-th reaction defined in (2.7), and for notational simplicity we have employed the abbreviation

$$\boldsymbol{c}! = c_1!\cdots c_N! \quad \text{and} \quad (\partial/\partial\boldsymbol{\phi})^{\boldsymbol{c}} = (\partial/\partial\phi_1)^{c_1}\cdots(\partial/\partial\phi_N)^{c_N}.$$

Due to the change of variables, the distribution $\tilde{P}_{\mathrm{km}}$ of the continuous molecular concentrations $\boldsymbol{\phi}$ in (2.17) is related to the distribution P_{m} of the discrete molecular population $\boldsymbol{n}$ in (2.4) by

$$\tilde{P}_{\mathrm{km}}(\boldsymbol{\phi},t) = V^N P_{\mathrm{m}}(\boldsymbol{n},t).$$

Since the right-hand-side of (2.17) is an expansion in the inverse powers of system volume V, Gillespie (2000) argued that the factors of V^{-1} tends to kill off the higher order terms in the thermodynamic limit, and leaves only the order V^0 terms which constitutes a variational description of the deterministic reaction rate equation. Then, he argued that the CFPE corresponds to a less drastic truncation after order V^{-1} terms, i.e.,

$$\frac{\partial \tilde{P}_{\mathrm{fp}}(\boldsymbol{\phi};t)}{\partial t} = -\sum_{i=1}^{N}\sum_{j=1}^{M}\nu_{j,i}\frac{\partial}{\partial\phi_i}f_j(\boldsymbol{\phi})\tilde{P}_{\mathrm{fp}}(\boldsymbol{\phi};t) + \frac{1}{2V}\sum_{i,i'=1}^{N}\sum_{j=1}^{M}\nu_{j,i}\nu_{j,i'}\frac{\partial^2}{\partial\phi_i\partial\phi_{i'}}f_j(\boldsymbol{\phi})\tilde{P}_{\mathrm{fp}}(\boldsymbol{\phi};t), \tag{2.18}$$

where $\tilde{P}_{\mathrm{fp}}(\boldsymbol{\phi},t) = V^N P_{\mathrm{f}}(\boldsymbol{x},t)$.

One conclusion may be inferred by comparing (2.18) with (2.17) which appears that the leading order term of the time derivatives of the difference $\tilde{P}_{\mathrm{km}} - \tilde{P}_{\mathrm{fp}}$ is the order V^{-2} term. However, a fundamental issue with truncation on (2.17) is that the higher

order derivatives on $\tilde{P}_{\mathrm{km}}$ also grows as volume V goes to infinity. To see this, let us consider the order V^0 terms in (2.17), written as

$$\frac{\partial \tilde{P}^0_{\mathrm{fp}}(\boldsymbol{\phi};t)}{\partial t} = -\sum_{i=1}^{N}\sum_{j=1}^{M} \nu_{j,i} \frac{\partial}{\partial \phi_i} f_j(\boldsymbol{\phi}) \tilde{P}^0_{\mathrm{fp}}(\boldsymbol{\phi};t),$$

and it can be shown that the equation in fact corresponds to the deterministic reaction rate equation. Assuming the expansion in (2.17) is valid, the distribution $\tilde{P}_{\mathrm{km}}$ converges to the order V^0 solution $\tilde{P}^0_{\mathrm{fp}}$, i.e., a Dirac delta distribution centred at the mean concentration value. Since the derivatives of the delta distribution blows up to infinity at the mean concentration value, and this fact invalidates the assumption of the expansion (2.17) that the coefficients in the higher order terms are bounded. Such problem is bypassed if instead using the system size expansion of both the CME and the CFPE and comparing the expanded solutions on different orders.

2.2.2 Comparison of the distributions of the CFPE and the CME

We have first performed the perturbation analysis of the CME and the CFPE, respectively, in Appendices A and B. We now use these results to derive the modelling error between the stationary distribution of the CME, π_{m}, and the stationary distribution of the CFPE, π_{f}.

The starting point of the comparison is applying the transformation of variables in (A.1), and we have

$$\pi_{\mathrm{m}}(\boldsymbol{n}) - \pi_{\mathrm{f}}(\boldsymbol{n}) = V^{-N/2}\left[\Pi_{\mathrm{m}}(\boldsymbol{\epsilon}) - \Pi_{\mathrm{f}}(\boldsymbol{\epsilon})\right], \tag{2.19}$$

and it transforms the error between the probability distributions of the population $\boldsymbol{n} \in \mathbb{N}^N$ to the error between the probability distributions of the fluctuations $\boldsymbol{\epsilon} \in \mathbb{R}^N$. Then, we substitute the perturbation expansion of Π_{m} in (A.20) and the expansion of

Π_{f} in (B.5) into (2.19), which yields

$$\pi_{\mathrm{m}}(\boldsymbol{n}) - \pi_{\mathrm{f}}(\boldsymbol{n}) = V^{-N/2}\left[\sum_{j=0}^{\infty} \Delta\Pi_j(\boldsymbol{\epsilon}) V^{-j/2}\right], \tag{2.20}$$

where

$$\Delta\Pi_j(\boldsymbol{\epsilon}) = \Pi_{\mathrm{m},j}(\boldsymbol{\epsilon}) - \Pi_{\mathrm{f},j}(\boldsymbol{\epsilon}), \quad j = 0, 1, 2, \ldots, \tag{2.21}$$

represents the difference between the expanded coefficients on the j-th order. Thus, to show that the distributions π_{m} and π_{f} agree to order $V^{-k/2}$ for some $k \geq 0$, one needs to prove that $\Delta\Pi_j = 0$ for all $j \leq k - N$.

Next, we substitute the Hermite expansions, (A.24) and (B.9), of the expanded coefficients into (2.21), and obtain

$$\Delta\Pi_j(\boldsymbol{\epsilon}) = \sum_{s=1}^{3j} \sum_{\substack{i_1, i_2, \ldots, i_N = 0 \\ i_1 + i_2 + \cdots + i_N = s}}^{s} \left(c_{\boldsymbol{i}}^{(j)} - c_{\mathrm{f},\boldsymbol{i}}^{(j)}\right) H_{\boldsymbol{i}}(\boldsymbol{\epsilon}, \boldsymbol{\Sigma}) \Pi_0(\boldsymbol{\epsilon}, \boldsymbol{\Sigma}), \tag{2.22}$$

where $H_{\boldsymbol{i}}(\boldsymbol{\epsilon}, \boldsymbol{\Sigma})$ represents the Hermite polynomial with respect to covariance $\boldsymbol{\Sigma}$, and $\Pi_0(\boldsymbol{\epsilon}, \boldsymbol{\Sigma})$ represents the centred Gaussian distribution with covariance $\boldsymbol{\Sigma}$. Note that, in the above expansion, we have used the fact that the maximum summation indices equal for both Hermite expansions in (A.24) and (B.9), and this is inferred from Propositions A.8 and B.1.

Since the Hermite polynomials are orthogonal, (2.22) suggests that $\Delta\Pi_j(\boldsymbol{\epsilon}) = 0$ if and only if

$$c_{\boldsymbol{i}}^{(j)} = c_{\mathrm{f},\boldsymbol{i}}^{(j)}, \tag{2.23}$$

for all $|\boldsymbol{i}| \leq N_j$. We have derived that the coefficients $\left\{c_{\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}| \leq 3j}$ can be obtain using (A.57), and the coefficients $\left\{c_{\mathrm{f},\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}| \leq 3j}$ satisfy (B.20). Thus, the condition (2.23) is

fulfilled if and only if we have

$$b_{\boldsymbol{i}}^{(j)} = b_{\mathrm{f},\boldsymbol{i}}^{(j)}, \tag{2.24}$$

for all $|\boldsymbol{i}| \leq 3j$. Summarising, we have the following lemma.

Lemma 2.3. *The error* $\Delta\Pi_j$ *in* (2.21) *is zero if and only if* $b_{\boldsymbol{i}}^{(s)} = b_{\mathrm{f},\boldsymbol{i}}^{(s)}$ *for all* $|\boldsymbol{i}| \leq 3s$, $s \leq j$.

Lemma 2.3 suggests that the convergence order of $\pi_{\mathrm{m}} - \pi_{\mathrm{f}}$ can be investigated by successive validating the equality (2.24) from $j = 1$ until the equality does not hold any more. In what follows, we prove the order of convergence for systems with and without detailed balance conditions.

2.2.2.1 Systems without the detailed balance condition

The leading order error term in (2.20) can be determined via comparing the coefficients of (A.52) and (B.17) on different orders. For mass-action reaction systems (2.1) without detailed balance condition, the SSE of the CME follows exactly the expression given in (A.8), and the values of $\left\{b_{\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\leq 3j}$ corresponds to (A.52). The result is summarised in the following Theorem.

Theorem 2.4. *Consider a general mass-action reaction system* (2.1)*. Assume that the deterministic reaction rate equation has a single asymptotic stable fixed point, and the covariance matrix* $\boldsymbol{\Sigma}$ *in* (A.19) *is strictly positive definite. For sufficiently large volume* V*, the error between the solution* π_{m} *of the stationary CME* (2.5) *and the solution* π_{f} *of the stationary CFPE* (2.13) *converges as*

$$\sup_{\boldsymbol{n}\in\mathbb{N}^N} |\pi_{\mathrm{m}}(\boldsymbol{n}) - \pi_{\mathrm{f}}(\boldsymbol{n})| \leq CV^{-(N+1)/2}, \tag{2.25}$$

where N *is the number of chemical species, and* C *is a constant independent of* V.

Proof. We have shown in (A.22) and (B.10) that both coefficients $\Pi_{\mathrm{m},0}$ and $\Pi_{\mathrm{f},0}$ are the steady state distribution of the LNA (A.15). Given the assumption that the covariance matrix $\boldsymbol{\Sigma}$ is strictly positive definite, coefficients $\Pi_{\mathrm{m},0}$ and $\Pi_{\mathrm{f},0}$ are finite and well-defined. Then, according to the series expansion of the error in (2.20), the convergence in (2.25) is obtained.

We next show that $\Pi_{\mathrm{m},1} = \Pi_{\mathrm{f},1}$ is not satisfied without further assumptions. According to Lemma 2.3, it suffices to show that $b^{(1)}_{\boldsymbol{n}} \neq b^{(1)}_{\mathrm{f},\boldsymbol{n}}$ for some $|\boldsymbol{n}| \leq 3$. Subtracting (A.52) from (B.17) yields

$$b^{(1)}_{\boldsymbol{n}} - b^{(1)}_{\mathrm{f},\boldsymbol{n}} = \sum_{\substack{i_1,\ldots,i_N=0 \\ i_1+\cdots+i_N=3}}^{3} \mathscr{F}^{\boldsymbol{i},\boldsymbol{0}}_{3,0,0} \mathscr{G}^{\boldsymbol{i},\boldsymbol{0}}_{\boldsymbol{n},\boldsymbol{0}}, = \sum_{\substack{i_1,\ldots,i_N=0 \\ i_1+\cdots+i_N=3}}^{3} \mathscr{G}^{\boldsymbol{i},\boldsymbol{0}}_{\boldsymbol{n},\boldsymbol{0}} \sum_{\substack{c_1,\ldots,c_N=0 \\ c_1+\cdots+c_N=3}}^{3} \mathscr{C}_{\boldsymbol{c}} \mathscr{D}^{\boldsymbol{c},\boldsymbol{0}}_{3,0,0}. \tag{2.26}$$

In the first equality of (2.26) we have used the fact that $\tilde{c}^{(0)}_{\boldsymbol{0}} = 1$ and $\tilde{c}^{(0)}_{\boldsymbol{m}} = 0$ for $|\boldsymbol{m}| \geq 1$; and in the second equality, the explicit expression for $\mathscr{F}^{\boldsymbol{i},\boldsymbol{0}}_{3,0,0}$ in (A.34) is inserted. It is easy to see that (2.26) does not equal to 0 in general, and thus the leading error term $\Delta\Pi_j$ does not vanish, meaning that the convergence order in (2.25) should not be improved. □

Remark 2.5. The error bound in (2.25) of Theorem 2.4 was derived in the large volume limit, such that the higher order terms in (2.19) is bounded by a constant on the leading order. That is the higher order errors all contribute to the constant C on the right-hand side of (2.25), and one can expanded as

$$C = \Delta\Pi_1 + \mathscr{O}\left(V^{-1/2}\right).$$

Then,

$$\hat{C} = \sup_{\boldsymbol{\epsilon}\in\mathbb{R}^N} \Delta\Pi_1(\boldsymbol{\epsilon}) V^{-1/2} = V^{-1/2} \sum_{s=1}^{3} \sum_{\substack{i_1,i_2,\ldots,i_N=0 \\ i_1+i_2+\cdots+i_N=s}}^{s} \left(c^{(1)}_{\boldsymbol{i}} - c^{(1)}_{\mathrm{f},\boldsymbol{i}}\right) H_{\boldsymbol{i}}(\boldsymbol{\epsilon},\boldsymbol{\Sigma}) \Pi_0(\boldsymbol{\epsilon},\boldsymbol{\Sigma}),$$

can be used as an error indicator for the modelling error $\sup_{\boldsymbol{n}\in\mathbb{N}^N} |\pi_{\mathrm{m}}(\boldsymbol{n}) - \pi_{\mathrm{f}}(\boldsymbol{n})|$.

2.2.2.2 Reversible systems with the detailed balanced condition

Let us reconsider the formula given in (2.26). A sufficient condition for $b^{(1)}_{\boldsymbol{n}} - b^{(1)}_{\mathrm{f},\boldsymbol{n}} = 0$ is $\mathscr{D}^{\boldsymbol{c},\boldsymbol{0}}_{3,0,0} = 0$ for all $|\boldsymbol{c}| = 3$, which would subsequently lead to that the order $V^{-1/2}$ terms $\Delta\Pi_1$ in (2.20) also vanishes. We will show in this section that such condition is fulfilled by all reversible reaction systems obeying the detailed balanced condition.

Definition 2.6. *A chemical reaction network is called reversible, if for every reaction of the form* (2.1) *with a positive reaction rate $k_{r+} > 0$, there exists a backward reaction with a positive reaction rate $k_{r-} > 0$.*

Definition 2.7. *A reversible reaction network is detailed balanced, if for each pair of reversible reactions, we have*

$$f_{r+}(\boldsymbol{\phi}) = f_{r-}(\boldsymbol{\phi}), \quad 1 \le r \le M/2, \tag{2.27}$$

where $f_{r\pm}$ are the deterministic rate functions in (A.6)*, $\boldsymbol{\phi} > 0$ is a positive equilibrium of the deterministic reaction rate equations, and M is the total number of reactions.*

Theorem 2.8. *Consider a reversible reaction system obeying the detailed balance condition. Assume that the covariance matrix $\boldsymbol{\Sigma}$ in* (A.19) *is strictly positive definite. The error between the CME solution π_{m} and the CFPE solution π_{f} satisfies*

$$\sup_{\boldsymbol{n}\in\mathbb{N}^N} |\pi_{\mathrm{m}}(\boldsymbol{n}) - \pi_{\mathrm{f}}(\boldsymbol{n})| \le CV^{-(N+2)/2}, \tag{2.28}$$

where N is the number of chemical species, and C is a constant independent of volume V.

Proof. We first apply (A.10) and write the explicit formula of $\mathscr{D}_{3,0,0}^{\boldsymbol{c},\mathbf{0}}$ as

$$\mathscr{D}_{3,0,0}^{\boldsymbol{c},\mathbf{0}} = \frac{1}{6}\sum_{r=1}^{M}\prod_{n=1}^{N}\nu_{r,n}^{i_n} f_r(\boldsymbol{\phi}).$$

According to Definition 2.6 on reversible reaction systems, we can split the above summation over all reactions into a summation over the forward reactions and a summation over the backward reactions, which yields

$$\mathscr{D}_{3,0,0}^{\boldsymbol{c},\mathbf{0}} = \frac{1}{6}\sum_{r=1}^{M/2}\prod_{n=1}^{N}\left(\nu_{r+,n}^{i_n} f_{r+}(\boldsymbol{\phi}) + \nu_{r-,n}^{i_n} f_{r-}(\boldsymbol{\phi})\right) = \frac{1}{6}\sum_{r=1}^{M/2}\left(f_{r+}(\boldsymbol{\phi}) - f_{r-}(\boldsymbol{\phi})\right)\prod_{n=1}^{N}\nu_{r+,n}^{i_n},$$

where we have used the fact that $\prod_{n=1}^{N}\nu_{r+,n}^{i_n} = -\prod_{n=1}^{N}\nu_{r-,n}^{i_n}$ for $|\boldsymbol{i}| = 3$. Inserting (2.27) of Definition 2.7 into the above formula, we obtain $\mathscr{D}_{3,0,0}^{\boldsymbol{c},\mathbf{0}} = 0$ for all $|\boldsymbol{c}| = 3$. Then, we derive from (2.26) that $b_{\boldsymbol{n}}^{(1)} = b_{\mathrm{f},\boldsymbol{n}}^{(1)}$, and applying Lemma 2.3 gives that the order $V^{-1/2}$ terms $\Delta\Pi_1$ in (2.20) vanishes. Hence, the leading order term in the series expansion (2.20) is $\Delta\Pi_2$, and the argument in (2.28) follows. □

Remark 2.9. The error indicator for $\sup_{\boldsymbol{n}\in\mathbb{N}^N}|\pi_{\mathrm{m}}(\boldsymbol{n}) - \pi_{\mathrm{f}}(\boldsymbol{n})|$ for the systems satisfying the detailed balance condition can be computed by evaluating the leading order error term $\Delta\Pi_2$, and we have

$$\hat{C} = V^{-1}\sup_{\boldsymbol{\epsilon}\in\mathbb{R}^N}|\Delta\Pi_2(\boldsymbol{\epsilon})| = V^{-1}\sum_{s=1}^{6}\sum_{\substack{i_1,i_2,\ldots,i_N=0\\ i_1+i_2+\cdots+i_N=s}}^{s}\left(c_{\boldsymbol{i}}^{(2)} - c_{\mathrm{f},\boldsymbol{i}}^{(2)}\right)H_{\boldsymbol{i}}(\boldsymbol{\epsilon},\boldsymbol{\Sigma})\Pi_0(\boldsymbol{\epsilon},\boldsymbol{\Sigma}).$$

And the indicator $\hat{C}$ converges to the constant C in (2.28) with order $V^{-1/2}$.

2.2.3 A numerical experiment: The birth-death process

We here consider a simple birth-death process,

$$\emptyset \underset{k_2}{\overset{k_1}{\rightleftarrows}} S, \tag{2.29}$$

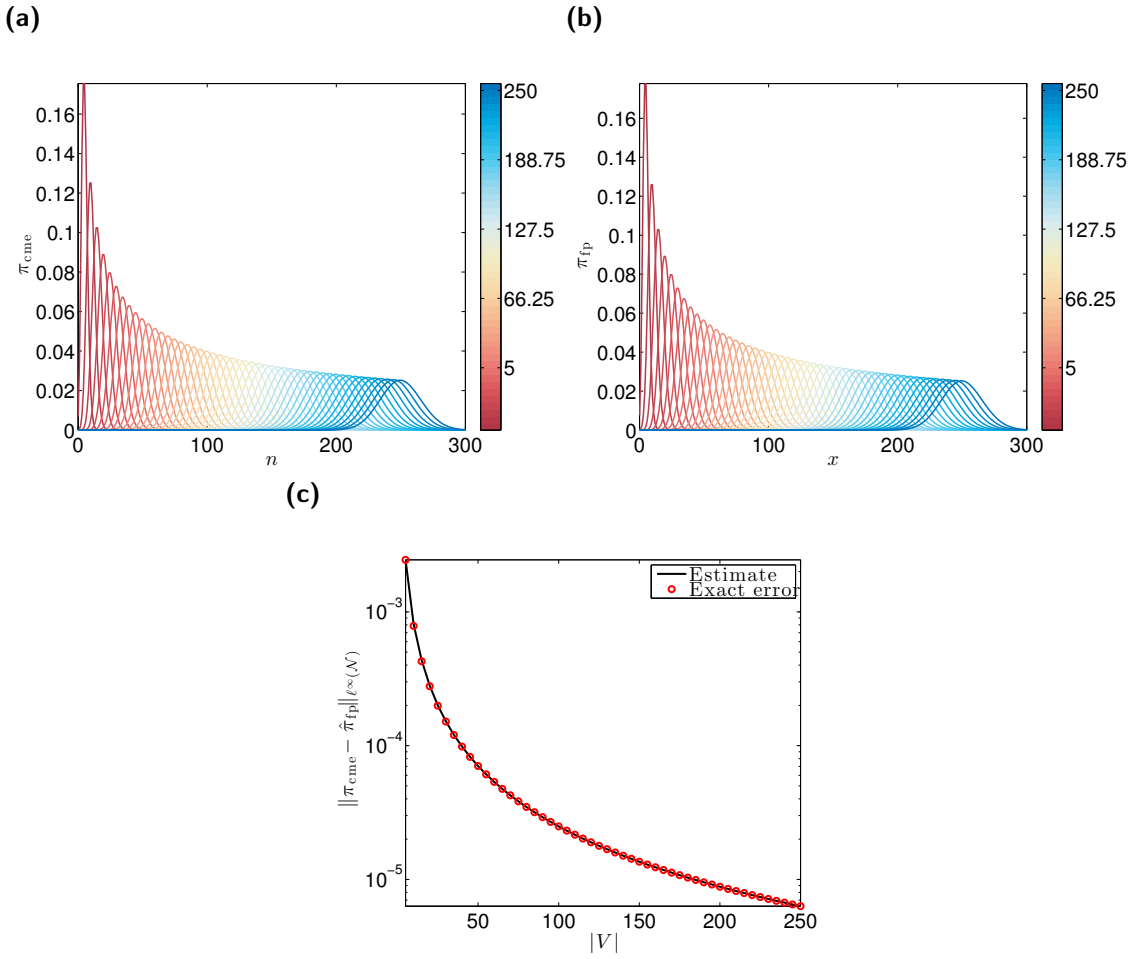

Figure 2.1: *The CME and CFPE predictions of the stationary distribution of the birth-death process* (2.29) *against increasing volume sizes. The reaction constants are set to be* $k_1 = k_2 = 1$. *For different volume sizes noted in the colourbar, the stationary CME distribution* (2.30) *is plotted in* **(a)** *and the stationary CFPE distribution* (2.33) *in* **(b)**. **(c)** *The difference between the CME solution and the CFPE solution. The red circles refers to the exact error measured in* ℓ^∞*-norm, and the black curve represent the right-hand side of* (2.28) *with the fitted values of the constant* $C = 0.07$.

where a single chemical species, denoted by S, is produced by some substrates within certain container of volume V at a constant rate k_1, and degrades with rate constant k_2. The underlying propensity functions are given by

$$\alpha_1 = k_1 V \quad \text{and} \quad \alpha_2 = k_2 x.$$

It can be shown that the stationary distribution of the death-birth process is Poisson distributed, i.e.,

$$\pi_{\mathrm{m}}(n) = \frac{1}{n!}\left(\frac{k_1 V}{k_1}\right)^n \exp\left(-\frac{k_1 V}{k_2}\right), \quad n \in \mathbb{N}. \tag{2.30}$$

The stationary CFPE (2.13) for system (2.29) is given by

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[D(x)\frac{\mathrm{d}\pi_{\mathrm{f}}(x)}{\mathrm{d}x}\right] - \frac{\mathrm{d}}{\mathrm{d}x}[v(x)\pi_{\mathrm{f}}(x)] = 0, \quad x \in \Omega_{\mathrm{f}}, \tag{2.31}$$

with positivity condition $\pi_{\mathrm{f}}(x) \geq 0$ for $x \in \Omega_{\mathrm{f}}$ and normalisation condition $\int_{\Omega_{\mathrm{f}}} \pi_{\mathrm{f}}(x)\mathrm{d}x = 1$. The diffusion coefficient and effective drift coefficient are given by $D(x) = (k_1 V + k_2 x)/2$ and $v(x) = k_1 V - k_2 x - k_2/2$, respectively. It is easily seen that D loses its positivity within the region $(-\infty, x_D]$, where $x_D = -k_1 V/k_2$, and thus we choose $\Omega_{\mathrm{f}} = (x_D, \infty)$.

For $V > V_0$, we integrating (2.31) over x gives

$$\frac{\mathrm{d}\pi_{\mathrm{f}}(x)}{\mathrm{d}x} - \frac{v(x)}{D(x)}\pi_{\mathrm{f}}(x) = \frac{C_1}{D(x)},$$

where C_1 is a constant. Multiplying by the integrating factor $\phi(x) = \exp\left[-\int^x \frac{v(y)}{D(y)}\mathrm{d}y\right]$ we can write the above equation in the compact form $\mathrm{d}(\phi(x)P_{\mathrm{f}}(x)) = C_1\phi(x)/D(x)$. Then another integration yields

$$\pi_{\mathrm{f}}(x) = \phi^{-1}(x)\left(C_1 \int^x \frac{\phi(y)}{D(y)}\mathrm{d}y + C_2\right). \tag{2.32}$$

The constants, C_1 and C_2, are determined to guarantee the positivity constraints $\pi_{\mathrm{f}}(x) \geq 0$ and the normalisation condition $\int_{\Omega_{\mathrm{f}}} \pi_{\mathrm{f}}(x)\mathrm{d}x = 1$. If this condition is satisfied, then a stationary density exists, or otherwise not.

To have a non-trivial solution π_{f} in (2.33), at least one of constants C_1 and C_2 is non-zero. We note that $\phi(x)/D(x) = 2\exp[2x - (4k_1 V/k_2)\log(k_1 V + k_2 x)] \to +\infty$ as

$x \to x_D$ or $x \to +\infty$. Thus we have

$$\lim_{x \to x_0+} \int^x \phi(y)/D(y)\mathrm{d}y = -\infty \quad \text{and} \quad \lim_{x \to +\infty} \int^x \phi(y)/D(y)\mathrm{d}y = +\infty,$$

meaning that to maintain $\pi_{\mathrm{f}}(x)$ positive throughout Ω_{f}, we must have $C_1 = 0$ and solution can only take the form

$$\pi_{\mathrm{f}}(x) = C_2 \phi^{-1}(x) = C_2 \exp\left[-2x + \left(\frac{4k_1 V}{k_2} - 1\right) \log(k_1 V + k_2 x)\right], \quad x \in \Omega_{\mathrm{f}}. \tag{2.33}$$

For $V > V_0$, we have $\lim_{x \to +\infty} \phi(x) = \lim_{x \to x_D+} \phi(x) = +\infty$ and the solution (2.33) is bounded. In this case, C_2 can be chosen to ensure that $\int_{\Omega_{\mathrm{f}}} \pi_{\mathrm{f}}(x)\mathrm{d}x = 1$. Moreover, it can be easily verified that the solution π_{f} in (2.33) naturally satisfies both the homogeneous Dirichlet boundary condition, $\lim_{x \to \partial\Omega_{\mathrm{f}}} \pi_{\mathrm{f}}(x) = 0$, and the no-flux boundary condition, $\lim_{x \to \partial\Omega_{\mathrm{f}}} J_{\mathrm{flux}}(x) = 0$, where the flux is defined as $J_{\mathrm{flux}}(x) = D(x)\mathrm{d}\pi_{\mathrm{f}}(x)/\mathrm{d}x - v(x)\pi_{\mathrm{f}}(x)$.

We evaluate the exact modelling error between the two solutions (visualised in Figures 2.1a and 2.1b), and plot the error measured in ℓ^∞-norm in Figure 2.1c against increasing values of system volumes V. It is obvious that larger system volumes leads to better CFPE approximations. As comparison, we use the black curve to refer to our error estimate in (2.28) of Theorem 2.8. By fitting the constants C in (2.28) to the exact errors (red circles), we obtain a good agreement between the exact errors and the estimated ones.

2.3 Noise-induced multistability

The current section studies the qualitative difference between the CME and the CFPE distributions. The majority of this section is composed of the published work in my co-authored paper, Duncan et al. (2015). The paper introduced a simple chem-

ical system exhibiting noise-induced multi-stability. For such system, the CME's marginal distribution becomes multi-modal as system size decreases below a critical value, while the the CFPE distribution is uni-modal for all system sizes. It also showed that one can recover the multi-modality using the quasi-stationary approximation (QSA) of the CFPE. The author's contribution includes discussing the idea of the QSA, performing finite element and Monte-Carlo simulations of the model equations for the example systems, and editing the manuscript. For completeness, most of the published content will be presented in Sections § 2.3.1 to § 2.3.4.

In addition, I will also introduce a hybrid scheme which partially approximates the CME by the CFPE in Section § 2.3.5. A Monte Carlo scheme is developed here to simulate the stochastic process modelled by the hybrid equation. Then, in Section § 2.3.6, we show that both the hybrid equation and the QSAs can recover the multi-modality in the example systems.

2.3.1 Background

The CFPE's ability to reproduce noise-induced oscillations predicted by the CME has been studied by a number of authors (see for example Erban et al. (2009); Thomas et al. (2013)). This section discuss whether the CFPE' s marginal probability distribution has generally the same number of maxima (modes) as the CME's marginal probability distribution.

It is clear by the CFPE's derivation that for cases where the multi-modality exists at large system sizes (the volume for chemical systems), the CFPE and CME predict the same number of modes of the marginal probability distribution. This is termed deterministic multi-stability since each peak in the marginal probability distribution is associated with a stable solution of the deterministic rate equations. It follows that, for these cases, the multi-modality in the CFPE solution stems from the drift term rather than from the diffusion term of the CFPE. It is also known that

the CFPE overestimates the transition rates between the deterministic stable states corresponding to each mode (Hanggi et al., 1984) however in this section we are only concerned with the ability of the CFPE to predict the correct number of maxima of the CME's probability distribution.

Besides deterministic multi-stability there is also a phenomenon referred to as stochastic multi-stability, noise-induced multi-modality or noise-induced multi-stability. This describes the case where the marginal probability distribution of the CME switches from a uni-modal to a multi-modal distribution as the system-size is decreased below a critical value. The existence of such a phenomenon has been long known (Horsthemke & Lefever, 1984) yet its practical relevance to biology and ecology has only started to be appreciated in the past decade (Kepler & Elston, 2001; Qian et al., 2009; Thomas et al., 2014; Biancalani et al., 2014). Given that noise-induced multi-modality occurs at intermediate or small system sizes, its unclear whether such a phenomenon is captured by the CFPE since the derivation of the latter does implicitly assume large system sizes. Curiously, recently it has been shown that the CFPE can capture the onset of noise-induced bi-modality (Biancalani et al., 2014) for a particular system of reactions which models a colony of foraging ants collecting food from two sources. The remaining question is whether the CFPE can generically capture this phenomenon. In this section we show, by means of examples, that this is not the case. In particular our findings clarify that there are two distinct types of noise-induced multi-modality and that only one of these types is captured by the CFPE.

2.3.2 A gene regulatory system with no feedback

We start by considering the following simple model of transcription regulation without feedback (Kepler & Elston, 2001):

$$S_1 \underset{k_{off}}{\overset{k_{on}}{\rightleftarrows}} S_1^*, \quad S_1 \xrightarrow{k_1} S_1 + S_2, \quad S_1^* \xrightarrow{k_2} S_1^* + S_2, \quad S_2 \xrightarrow{c} \emptyset. \tag{2.34}$$

A gene can be in one of two states S_1 and S_1^*; switching between these two states is random and each state is associated with a different rate of protein (S_2) formation. Once formed, the protein can also decay. This model is a simplification of more realistic gene models where the mRNA is explicitly modelled (Shahrezaei & Swain, 2008). We consider the case where there are N_{gene} genes such that the total number of S_1 and S_1^* equals N_{gene} at all times; although genes typically exist in one or two copies per cell, plasmids are nowadays commonly used to genetically engineer cells with a large number of copies of a given gene (Alberts et al., 1997) and hence our model is of biochemical relevance.

Denoting by t the time and by $\tau = ct$ the dimensionless time, the CME for the reaction system (2.34) is given by:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}P_{\mathrm{m}}(n_1,n_2;\tau) = {} & k_{off}(\mathscr{E}_{n_1}^{-1}-1)(N_{\text{gene}}-n_1)P_{\mathrm{m}}(n_1,n_2;\tau) + k_{on}(\mathscr{E}_{n_1}^{+1}-1)n_1P_{\mathrm{m}}(n_1,n_2,\tau) \\ & + k_1n_1(\mathscr{E}_{n_2}^{-1}-1)P_{\mathrm{m}}(n_1,n_2;\tau) + k_2(N_{\text{gene}}-n_1)(\mathscr{E}_{n_2}^{-1}-1)P_{\mathrm{m}}(n_1,n_2;\tau) \\ & + (\mathscr{E}_{n_2}^{+1}-1)n_2P_{\mathrm{m}}(n_1,n_2;\tau),\end{aligned} \tag{2.35}$$

where $P_{\mathrm{m}}(n_1,n_2;\tau)$ is the probability that at time τ there are n_1 genes in state S_1 and n_2 protein molecules according to the CME, and $\mathscr{E}_{n_1}^m$ and $\mathscr{E}_{n_2}^m$ are step operators such that when they act on a function $f \equiv f(n_1,n_2)$, their action is $\mathscr{E}_{n_1}^m f(n_1,n_2) = f(n_1+m,n_2)$ and $\mathscr{E}_{n_2}^m f(n_1,n_2) = f(n_1,n_2+m)$. The non-dimensional reaction rates

are given by:

$$q_{off} = \frac{k_{off}}{c}, \quad q_{on} = \frac{k_{on}}{c}, \quad q_1 = \frac{k_1}{c}, \quad q_2 = \frac{k_2}{c}.$$

The CFPE for the reaction system (2.34) (with dimensionless time units, $\tau = ct$) is given by:

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\tau}P_{\mathrm{f}}(x_1,x_2;\tau) = &-\frac{\partial}{\partial x_1}\left(\left(q_{off}(N_{\mathrm{gene}}-x_1)-q_{on}x_1\right)P_{\mathrm{f}}(x_1,x_2;\tau)\right)\\ &-\frac{\partial}{\partial x_2}\left(\left(q_1x_1+q_2(N_{\mathrm{gene}}-x_1)-x_2\right)P_{\mathrm{f}}(x_1,x_2;\tau)\right)\\ &+\frac{1}{2}\frac{\partial^2}{\partial x_1^2}\left(\left(q_{off}(N_{\mathrm{gene}}-x_1)+q_{on}x_1\right)P_{\mathrm{f}}(x_1,x_2;\tau)\right)\\ &+\frac{1}{2}\frac{\partial^2}{\partial x_2^2}\left(\left(q_1x_1+q_2(N_{\mathrm{gene}}-x_1)+x_2\right)P_{\mathrm{f}}(x_1,x_2;\tau)\right),\end{aligned} \tag{2.36}$$

where $P_{\mathrm{f}}(x_1,x_2;\tau)$ is the probability that at time τ there are x_1 genes in state S_1 and x_2 protein molecules according to the CFPE.

2.3.3 The Quasi-stationary approximation

Next we solve the CME (2.35) under the condition that the timescales governing the decay of small fluctuations about the steady-state mean number of molecules of S_1 and S_2 are well separated, i.e, the quasi-stationary approximation (QSA) of the CME. Since the system (2.34) consists of purely first-order reactions, the equations for the means $\langle n_1\rangle$ and $\langle n_2\rangle$ as obtained from the CME (2.35) are precisely given by the conventional rate equations:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\langle n_1\rangle = -q_{on}\langle n_1\rangle + q_{off}(N_{\mathrm{gene}}-\langle n_1\rangle)$$

and

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\langle n_2\rangle = q_1\langle n_1\rangle + q_2\left(N_{\mathrm{gene}}-\langle n_1\rangle\right)-\langle n_2\rangle.$$

The two characteristic non-dimensional timescales are given by the absolute value of the inverse of the eigenvalues of the Jacobian of the above system of rate equations and are: $\tau_1 = (q_{on} + q_{off})^{-1}$ and $\tau_2 = 1$, where the former governs the decay of small fluctuations in S_1 and the latter the same but for S_2. Clearly gene switching between the two states S_1 and S_1^* is much slower than the rest of the processes in the system whenever $\tau_1 \gg 1$, i.e., the protein reaches steady-state in a time much shorter than the time it takes for a gene to switch from one state to another.

Now we approximately solve the CME (2.35) in the quasi-stationary limit given by $\varepsilon = \tau_1^{-1} \ll 1$. Rescaling time by $\tau \to \varepsilon\tau$ we obtain the following master equation:

$$\frac{\mathrm{d}}{\mathrm{d}\tau} P_{\mathrm{m}}(n_1, n_2; \tau) = \frac{1}{\varepsilon} \mathscr{L}^0 P_{\mathrm{m}}(n_1, n_2; \tau) + \mathscr{L}^1 P_{\mathrm{m}}(n_1, n_2; \tau), \tag{2.37}$$

where the two operators $\mathscr{L}^0$ and $\mathscr{L}^1$ are given by:

$$\mathscr{L}^0 = (q_1 n_1 + q_2(N_{\mathrm{gene}} - n_1))(E_{n_2}^{-1} - 1) + (E_{n_2}^{+1} - 1)n_2, \tag{2.38}$$

$$\mathscr{L}^1 = \alpha(E_{n_1}^{+1} - 1)n_1 + \beta(E_{n_1}^{-1} - 1)(N_{\mathrm{gene}} - n_1), \tag{2.39}$$

and $\alpha = q_{on}/(q_{on} + q_{off})$ and $\beta = 1 - \alpha$. Note that $\mathscr{L}^0$ acts only on the protein numbers n_2 while $\mathscr{L}^1$ acts only on the gene numbers n_1.

In order to solve Equation (2.37) in the limit of small ε, we consider the perturbation ansatz of the steady state solution $\pi_{\mathrm{m}}(n_1, n_2) = \lim_{\tau\to\infty} P_{\mathrm{m}}(n_1, n_2; \tau)$:

$$\pi_{\mathrm{m}}(n_1, n_2) = \pi_{\mathrm{m}}^0(n_1, n_2) + \varepsilon\pi_{\mathrm{m}}^1(n_1, n_2) + \ldots + \varepsilon^i \pi_{\mathrm{m}}^i(n_1, n_2) + \ldots$$

Substituting the latter in Equation (2.37) and comparing coefficients of powers of ε

we obtain the following equations:

$$O\left(\frac{1}{\varepsilon}\right): \mathscr{L}^0 \pi_{\mathrm{m}}^0 = 0, \tag{2.40}$$

$$O(1): \mathscr{L}^0 \pi_{\mathrm{m}}^1 + \mathscr{L}^1 \pi_{\mathrm{m}}^0 = \frac{\partial}{\partial \tau} \pi_{\mathrm{m}}^0 = 0. \tag{2.41}$$

Note that $\partial_\tau \pi_{\mathrm{m}}^0 = 0$ by the assumption of steady-state. We can write

$$\pi_{\mathrm{m}}^0(n_1, n_2) = \pi_{\mathrm{m}}^0(n_2 | n_1) \pi_{\mathrm{m}}(n_1),$$

where $\pi_{\mathrm{m}}^0(n_2|n_1)$ is the stationary density for S_2 given the number of S_1 molecules is n_1, and $\pi_{\mathrm{m}}(n_1)$ is the marginal stationary distribution for S_1. Summing the $O(1)$ equation over n_2, we obtain:

$$\sum_{n_2 \in \mathbb{N}} \left[\mathscr{L}^0 \pi_{\mathrm{m}}^1(n_1, n_2) \right] + \mathscr{L}^1 \left[\sum_{n_2 \in \mathbb{N}} \pi_{\mathrm{m}}^0(n_2 | n_1) \pi_{\mathrm{m}}(n_1) \right] = 0. \tag{2.42}$$

The first term on the left hand side is zero by the definition of $\mathscr{L}^0$ in Equation (2.38). The second term simplifies by the normalisation condition $\sum_{n_2 \in \mathbb{N}} \pi_{\mathrm{m}}^0(n_2 | n_1) = 1$. Thus Equation (2.42) reduces to:

$$\mathscr{L}^1 \pi_{\mathrm{m}}(n_1) = 0,$$

which possesses a unique normalised solution given by:

$$\pi_{\mathrm{m}}(n_1) = \frac{\binom{N_{\mathrm{gene}}}{n_1}}{(1+\lambda)^{N_{\mathrm{gene}}}} \lambda^{n_1}, \quad n_1 \in \{0, \ldots, N_{\mathrm{gene}}\} \tag{2.43}$$

where $\lambda = \beta/\alpha = (1-\alpha)/\alpha = q_{off}/q_{on}$. From the $O(\varepsilon^{-1})$ equation we obtain:

$$\pi_{\mathrm{m}}(n_1) \mathscr{L}^0 \pi_{\mathrm{m}}^0(n_2 | n_1) = 0.$$

This equation can be easily solved yielding the Poissonian:

$$\pi_{\mathrm{m}}^{0}(n_2\,|\,n_1) = \frac{1}{n_2!}\left(q_1 n_1 + q_2(N_{\mathrm{gene}} - n_1)\right)^{n_2} e^{-(q_1 n_1 + q_2(N_{\mathrm{gene}} - n_1))}. \tag{2.44}$$

Finally multiplying Equation (2.43) and Equation (2.44) and summing over n_1, we obtain by the law of total probability the stationary distribution of the protein numbers in the limit of small ε:

$$\pi_{\mathrm{m}}(n_2) \approx \sum_{n_1=0}^{N_{\mathrm{gene}}} \binom{N_{\mathrm{gene}}}{n_1} \frac{\lambda^{n_1}}{n_2!} \frac{(q_1 n_1 + q_2(N_1 - n_1))^{n_2}}{(1+\lambda)^{N_{\mathrm{gene}}}} e^{-(q_1 n_1 + q_2(N_{\mathrm{gene}} - n_1))}. \tag{2.45}$$

In principle a similar quasi-stationary limit can be applied to the CFPE Equation (2.36). In practice, however, the equation thus obtained cannot be reduced beyond an integral which has no closed form solution; thus in what follows we obtain the stationary distribution of the protein numbers of the CFPE numerically (see later for details).

2.3.4 CME vs CFPE solutions

Let V be the volume of the compartment in which the chemical reaction network (2.34) is confined. Furthermore let $\phi_2 = n_2/V$ be the protein concentration and $\phi_1 = N_{\mathrm{gene}}/V$ be the (constant) total gene concentration. We now study the number of modes of the quasi-stationary distribution of ϕ_2 as a function of V; in this thought experiment, the volume V is varied at constant ϕ_1 such that an increase in volume, necessarily translates into a proportionate increase in the total number of genes (this also implies that the number of proteins increases with the volume). Now the CME distribution of ϕ_2, is given by $\pi(\phi_2) = V\pi_{\mathrm{m}}(\phi_2 V)$. This probability distribution solution consists of a superposition of $N_{\mathrm{gene}} + 1$ Poisson distributions; this is since there are $N_{\mathrm{gene}} + 1$ combinations of N_{gene} molecules which can be in one of two states. This implies that $\pi(\phi_2)$ is generally multi-modal, with at most $N_{\mathrm{gene}} + 1$ modes. Since

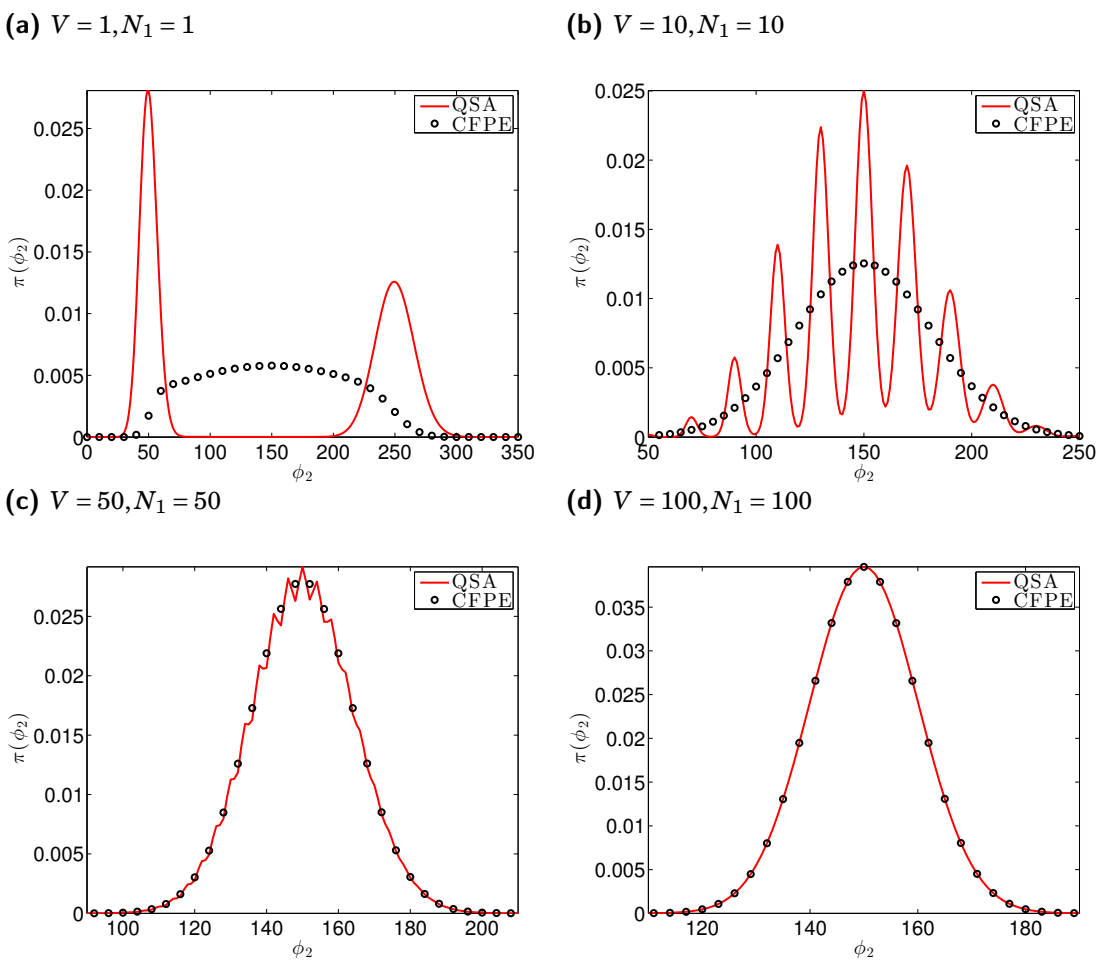

Figure 2.2: *Comparison of the stationary distribution of protein concentration obtained from SSA simulations with the QSA solution $\pi_{\mathrm{m}}(\phi_2)$ obtained by Equation* (2.45) *and the numerical solution of the CFPE given by Equation* (2.36) *for different values of volume. The volumes used are $V = 1, 10, 50, 100$ for* **(a)** *–* **(d)** *respectively; since the gene concentration ϕ_1 is fixed to 1, this implies that the gene copy numbers N_1 are the same as the volumes. Parameters (see text) are such that timescale separation is enforced. The QSA and SSA predict multi-modality below a certain volume whereas the CFPE predicts a uni-modal distribution for all volumes.*

N_{gene} increases with V, we would expect the modality of $\pi(\phi_2)$ to increase with the volume, if the Poissonians are well separated. On the other hand, given that the deterministic rate equations of the chemical system are monostable, we also know, by the system-size expansion Van Kampen (2007), that in the thermodynamic limit of large volumes, the probability distribution $\pi(\phi_2)$ tends to a Gaussian and thus uni-modal. Hence the overall picture is that the number of modes of $\pi(\phi_2)$ should increase with V for V less than some critical volume V^* and decrease with increasing

volume for $V > V^*$. The smallest volume possible is that for which there is one gene $N_{\text{gene}} = 1$; hence the maximum number of modes in the limit of small volumes is 2.

In Figure 2.2, we plot the QSA solution $\pi(\phi_2)$ for four different volumes with parameters $\phi_1 = 1$, $q_{on} = q_{off} = 10^{-3}$, $q_1 = 50$, $q_2 = 250$; we compare this with the solution from the CFPE (2.36). Note that for the chosen parameters, $\varepsilon = 2 \times 10^{-3} \ll 1$, which implies timescale separation; this is reflected in the excellent agreement between the analytical approximation (QSA) and the CME. As predicted above, the modality of the probability distribution of the CME/QSA goes through a maximum as the volume is progressively increased, with the number of modes for very low and large volumes being 2 and 1, respectively. Thus the CME predicts noise-induced multi-modality as the volume is decreased beyond some critical value; we call this "noise-induced" since the particle numbers becomes smaller with the volume and the size of intrinsic noise correspondingly increases.

The CFPE solution is obtained by discretising the PDE (2.36) using the continuous Galerkin finite element method (Larsson & Thomée, 2008) over a triangulation of the domain $[0, N_1] \times [0, N_2^*]$ where $N_2^* > 0$ is some artificial maximum protein number which is chosen sufficiently large such that its value makes no significant difference to the solution; no-flux boundary conditions are imposed along the domain boundaries. Note that the CFPE, unlike the CME, does not predict noise-induced multi-modality as the volume is decreased from 100 to 1 – rather it is uni-modal for all four volumes in Figure 2.2. Note also that for some volumes below $V = 1$, the CFPE does predict two modes of the probability distribution (e.g. a mode at zero and one at a non-zero concentration for $V = 1/2$), however for such volumes the total number of genes is less than 1 (since $\phi_1 = 1$); hence we can more precisely state that over the whole range of physically meaningful volumes, the CFPE is uni-modal. Note that the no-flux (reflective) boundary conditions on the CFPE are artificial in the sense that unlike the CME, the CFPE does not naturally lead to a restriction of gene numbers

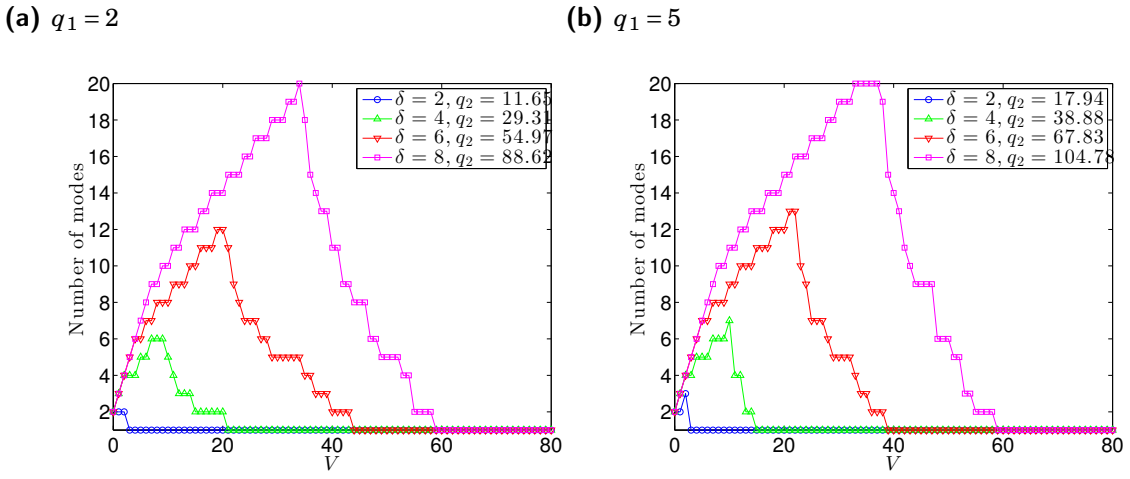

Figure 2.3: *Plot of the number of modes of $\pi_{\mathrm{m}}(\phi_2)$ as a function of the volume V, for different values of $\delta = \sqrt{q_2} - \sqrt{q_1}$. The larger is the difference between the production rates q_1, q_2, the larger is the volume range over which the stationary distribution of the CME is multi-modal; in contrast, the solution of the CFPE is unimodal for all volumes.*

between 0 and N_{gene} – the imposition of such artificial reflective conditions can lead to additional errors in the CFPE predictions (see for example Schnoerr et al. (2014)).

The behaviour elucidated in Figure 2.2 is not particular to the parameter set used; in Figure 2.3 we show the number of modes of the QSA distribution of protein $\pi_{\mathrm{m}}(\phi_2)$ as a function of the volume V for 8 different parameter sets (with fixed $\phi_1 = 1$, $\lambda = 1$). In all cases, the number of modes is 1 for large volumes, increases as the volume decreases and reaches a maximum at some critical volume; as the volume is decreased further, the number of modes steadily decreases until it reaches a value of 2 at the lowest volume of $V = 1$ (below this volume the total number of genes is less than one and hence unrealistic). The CFPE probability distribution solution is uni-modal for all volumes.

A comparison of Figures 2.3a and 2.3b shows that the dimensionless parameter $\delta = \sqrt{q_2} - \sqrt{q_1}$ (and not the individual values of q_1 and q_2) appears to be the principle factor determining the maximum number of modes of the probability distribution, as well as the critical volume at which this occurs. In particular the larger is δ, the larger is the volume (and the associated number of genes) above which the

probability distribution of the CME becomes uni-modal and agrees with the CFPE. A heuristic explanation of this is as follows. For the case of one gene ($N_{\text{gene}} = 1$), the QSA solution (2.45) predicts two modes, a Poissonian with mean and variance equal to q_1 and a second Poissonian with mean and variance equal to q_2. Now say that $q_1 < q_2$; then the condition of well separated Poissonians can be formulated as $q_1 + \sqrt{q_1} \ll q_2 - \sqrt{q_2}$ which is equivalent to $\delta = \sqrt{q_2} - \sqrt{q_1} \gg 1$. The larger δ is, the more pronounced the bimodal character of the distribution at $N_{\text{gene}} = 1$; hence one would expect a larger total number of genes is needed for the multi-modality to be washed away – this is consistent with the role played by δ in Figure 2.3.

2.3.5 Partial approximation of the CME by the CFPE

We have shown in the previous section, by means of a gene circuit (2.34), that the CFPE does not always capture the noise-induced multi-modality predicted by the CME. This implies that there are two different types of systems: those for which the CFPE can capture the inherent noise-induced multi-modality such as the foraging ant model studied in Biancalani et al. (2014) and those for which the CFPE fails to predict the phenomenon. The main difference between these two classes of systems is that in the former the switching between different states is described by the CFPE's diffusion term (see discussion in Biancalani et al. (2014)) whereas the multi-modality studied in this article cannot be so described.

We note that the probability distribution of the CME Eq. (12) is a superposition of Poissonians, each one associated with an element of the set $\{(0, N_{\text{gene}}), (1, N_{\text{gene}} - 1), \ldots, (N_{\text{gene}}, 0)\}$ which describes the possible combinations of genes in states S_1 and S_1^* – it is this direct association between the distribution and the inherent discreteness of the system that leads to the CFPE's inability to capture the multi-modality of the CME. Hence in some sense the phenomenon described here maybe more aptly described as discreteness-induced multi-modality.

Thus we generally expect the CFPE to fail to reproduce noise-induced multi-modality whenever the multi-modality can be explicitly related to the switching of a molecule between a discrete number of conformational states; this is since the CFPE is a continuum approximation and hence cannot capture discreteness. This is the case of a broad class of gene regulatory networks under timescale separation conditions (Thomas et al., 2014). Since enzymes also switch between a number of discrete conformational states (English et al., 2006) we also expect the CFPE to fail to reproduce multi-modality in the enzyme product distributions under certain conditions.

In the current section, we introduce a hybrid model that couples the CME and the CFPE, in a way that the microscopic species are described by the CME and the macroscopic species are incorporated into the CFPE approximation.

2.3.5.1 A short derivation.

Now we assume that, among the N reacting species, there are $N_{\boldsymbol{x}}$ of them which have small molecular number and $N_{\boldsymbol{y}} = N - N_{\boldsymbol{x}}$ of them have at least one magnitude more molecules. We keep using $\boldsymbol{x} = (x_1, x_2, \ldots, x_N)^T$ to denote the molecular population of microscopic species, and use $\boldsymbol{y} = (y_1, y_2, \ldots, y_{N_{\boldsymbol{y}}})^T$ to denote the molecular concentration of the macroscopic species. Then the propensity functions of the hybrid system is defined by

$$\alpha_j(\boldsymbol{x}, \boldsymbol{y}) = k_j \exp\left[\left(1 - \sum_{i=1}^{N} \nu_{j,i}^-\right) \log V\right] \prod_{i=1}^{N} (\nu_{j,i}^-)! \left[\binom{x_i}{\nu_{j,i}^-} \mathbb{1}_{i \le N_{\boldsymbol{x}}} + \binom{y_i}{\nu_{j,i}^-} \mathbb{1}_{i > N_{\boldsymbol{x}}}\right].$$

Let $P_{\mathrm{m}}(\boldsymbol{x}, \boldsymbol{y}; t)$ be the probability of the system being in a particular state $\boldsymbol{X}(t) = (\boldsymbol{x}, \boldsymbol{y})^T$ at time t. In order to apply the continuous Fokker-Planck approximation to

the macroscopic species, we organise the CME (2.4) as below:

$$\frac{\mathrm{d}P_{\mathrm{m}}(\boldsymbol{x},\boldsymbol{y};t)}{\mathrm{d}t} = \sum_{j=1}^{M}\left[\alpha_j(\boldsymbol{x}-\boldsymbol{\xi}_j,\boldsymbol{y}-\boldsymbol{\eta}_j)P_{\mathrm{m}}(\boldsymbol{x}-\boldsymbol{\xi}_j,\boldsymbol{y}-\boldsymbol{\eta}_j;t) - \alpha_j(\boldsymbol{x}-\boldsymbol{\xi}_j,\boldsymbol{y})P_{\mathrm{m}}(\boldsymbol{x}-\boldsymbol{\xi}_j,\boldsymbol{y};t)\right] + \sum_{j=1}^{M}\left[\alpha_j(\boldsymbol{x}-\boldsymbol{\xi}_j,\boldsymbol{y})P_{\mathrm{m}}(\boldsymbol{x}-\boldsymbol{\xi}_j,\boldsymbol{y};t) - \alpha_j(\boldsymbol{x},\boldsymbol{y})P_{\mathrm{m}}(\boldsymbol{x},\boldsymbol{y};t)\right], \tag{2.46}$$

where $\boldsymbol{\xi}_j \equiv \boldsymbol{\nu}_{j,1:N_{\boldsymbol{x}}}$ represents the a vector contains the first $N_{\boldsymbol{x}}$ entries of j-th row of the stoichiometric matrix, and $\boldsymbol{\eta}_j \equiv \boldsymbol{\nu}_{j,N_{\boldsymbol{x}}+1:N}$ refers to the last $N_{\boldsymbol{y}}$ entries. The first sum in the above equation could be viewed as the probability flux in the coordinates of the macroscopic species, and the second sum is the flux in the coordinates of the microscopic species. Thus, the original equation (2.4) is separately projected onto the spaces of the macroscopic and microscopic variables.

We may need two additional assumptions before proceeding to the continuous approximations. First, we require that all the components of $\boldsymbol{y}$ are very large compared to 1, and the macroscopic species are in the realm of "large number" approximation. Second, the function $\omega_j(\boldsymbol{x},\boldsymbol{y};t) = \alpha_j(\boldsymbol{x},\boldsymbol{y})P_{\mathrm{m}}(\boldsymbol{x},\boldsymbol{y};t)$ must be infinitely differentiable in the real variable $\boldsymbol{y}$ at any time t. Given the two assumptions, we can safely use Taylor's theorem to expand the first sum in (2.46) around $(\boldsymbol{x}-\boldsymbol{\xi}_j,\boldsymbol{y})$, which gives

$$\frac{\mathrm{d}P_{\mathrm{m}}(\boldsymbol{x},\boldsymbol{y};t)}{\mathrm{d}t} = \sum_{i=1}^{\infty}(-1)^i \sum_{\substack{c_1,c_2,\ldots,c_{N_{\boldsymbol{y}}}=0 \\ c_1+c_2+\cdots+c_{N_{\boldsymbol{y}}}=i}}^{i} \frac{1}{c_1!c_2!\cdots c_{N_{\boldsymbol{y}}}!}\frac{\partial^i}{\partial y_1^{c_1}\partial y_2^{c_2}\cdots\partial y_{N_{\boldsymbol{y}}}^{c_{N_{\boldsymbol{y}}}}} \times\left\{\left[\sum_{j=1}^{M}(\eta_{j,1}^{c_1}\eta_{j,2}^{c_2}\cdots\eta_{j,N_{\boldsymbol{y}}}^{c_{N_{\boldsymbol{y}}}})\alpha_j(\boldsymbol{x},\boldsymbol{y})\right]P_{\mathrm{m}}(\boldsymbol{x},\boldsymbol{y};t)\right\} + \sum_{j=1}^{M}\left[\alpha_j(\boldsymbol{x}-\boldsymbol{\xi}_j,\boldsymbol{y})P_{\mathrm{m}}(\boldsymbol{x}-\boldsymbol{\xi}_j,\boldsymbol{y};t) - \alpha_j(\boldsymbol{x},\boldsymbol{y})P_{\mathrm{m}}(\boldsymbol{x},\boldsymbol{y};t)\right]. \tag{2.47}$$

The above equation could be viewed as the coupling of the chemical Kramers-Moyal equation (2.9) and the CME (2.4). It requires us to regard the molecular population of the macroscopic species, $\boldsymbol{y}$, as real numbers rather than the natural integers. A hybrid equation that couples the CFPE and the CME can be obtained by truncating

the CKME part at the second order terms. We write it as (Sjöberg, 2007)

$$\begin{cases} \frac{\partial P_{\mathrm{cf}}(\boldsymbol{x},\boldsymbol{y};t)}{\partial t} = \left(\mathscr{L}_{\mathrm{m}}^{\boldsymbol{x}} + \mathscr{L}_{\mathrm{f}}^{\boldsymbol{y}}\right) P_{\mathrm{cf}}(\boldsymbol{x},\boldsymbol{y};t), \quad \boldsymbol{x} \in \mathbb{N}^{N_{\boldsymbol{x}}}, \quad \boldsymbol{y} \in \Omega_{\mathrm{f}}^{\boldsymbol{y}}, \quad t > 0, \\ P_{\mathrm{cf}}(\boldsymbol{x},\boldsymbol{y};t) \geq 0, \quad \sum_{\boldsymbol{x}\in\mathbb{N}^{N_{\boldsymbol{x}}}} \int_{\Omega_{\mathrm{f}}^{\boldsymbol{y}}} P_{\mathrm{cf}}(\boldsymbol{x},\boldsymbol{y};t)\mathrm{d}\boldsymbol{y} = 1, \end{cases} \tag{2.48}$$

where the two operators $\mathscr{L}_{\mathrm{m}}^{\boldsymbol{x}}$ and $\mathscr{L}_{\mathrm{f}}^{\boldsymbol{y}}$ are given by

$$\begin{aligned} \mathscr{L}_{\mathrm{m}}^{\boldsymbol{x}} P_{\mathrm{cf}}(\boldsymbol{x},\boldsymbol{y};t) &= \sum_{j=1}^{M} \left[\alpha_j(\boldsymbol{x}-\boldsymbol{\xi}_j,\boldsymbol{y}) P_{\mathrm{cf}}(\boldsymbol{x}-\boldsymbol{\xi}_j,\boldsymbol{y};t) - \alpha_j(\boldsymbol{x},\boldsymbol{y}) P_{\mathrm{cf}}(\boldsymbol{x},\boldsymbol{y};t)\right] \\ \mathscr{L}_{\mathrm{f}}^{\boldsymbol{y}} P_{\mathrm{cf}}(\boldsymbol{x},\boldsymbol{y};t) &= -\sum_{i=1}^{N_{\boldsymbol{y}}} \frac{\partial}{\partial y_i} \left(\sum_{j=1}^{M} \eta_{j,i} \alpha_j(\boldsymbol{x},\boldsymbol{y}) P_{\mathrm{cf}}(\boldsymbol{x},\boldsymbol{y};t) \right) \\ &\quad + \frac{1}{2} \sum_{i,i'=1}^{N_{\boldsymbol{y}}} \frac{\partial^2}{\partial y_i \partial y_{i'}} \left(\sum_{j=1}^{M} \eta_{j,i}\eta_{j,i'} \alpha_j(\boldsymbol{x},\boldsymbol{y}) P_{\mathrm{cf}}(\boldsymbol{x},\boldsymbol{y};t) \right). \end{aligned}$$

Then the domain $\Omega_{\mathrm{f}}^{\boldsymbol{y}}$ is chosen such that the diffusion coefficients of $\mathscr{L}_{\mathrm{f}}^{\boldsymbol{y}}$ is positive definite.

Note that the truncation of the continuous part in (2.47) at the second order term is the source of modelling error in the hybrid equation (2.48). Therefore, the convergence of the modelling error in the limit of large volumes can be found in a similar manner as in Section § 2.2.

2.3.5.2 Monte Carlo simulation method

We present a kinetic Monte Carlo strategy to generate statistical accurate trajectories satisfying the hybrid equation (2.48).

The hybrid model (2.48) has $\boldsymbol{x}$ as discrete stochastic variables and $\boldsymbol{y}$ as continuous variables. The discrete variables $\boldsymbol{x}$ follows the Markov jump process of the CME, while the continuous variables obeys a partial form of the canonical CLE (2.14):

$$\mathrm{d}y_i(t) = \sum_{j=1}^{M} \eta_{j,i} \alpha_j(\boldsymbol{x}-\boldsymbol{\xi}_j,\boldsymbol{y})\mathrm{d}t + \sum_{j=1}^{M} \eta_{j,i} \sqrt{\alpha_j(\boldsymbol{x}-\boldsymbol{\xi}_j,\boldsymbol{y})}\mathrm{d}w_j, \quad i = 1,\ldots,N_{\boldsymbol{y}}, \tag{2.49}$$

where $\boldsymbol{w} = (w_1, w_2, \ldots, w_M)$ is an M-dimensional Wiener process. The SDE (2.49) may be numerically integrated using a variety of numerical methods, such as Euler-Maruyama method, Milstein method and Runge-Kutta method. These stochastic numerical integrators usually require partitioning a time interval $[0, T]$ into subintervals with a time increment Δt. It is possible that, within Δt, the number of the discrete variables $\boldsymbol{x}$ changes, and the parameters of (2.49) would be different. Therefore, it is important to dynamically monitor the occurrence of any reaction associated with discrete variables within the time increment Δt. If more than one reaction occurs within Δt, we must rewind the state of the system to the previous one, decrease Δt of the SDE integrator until only zero or one reaction occurs. When a single reaction, assuming the j-th reaction, fires within Δt, the corresponding reaction time τ_j must be computed, and the changes in the number of discrete species should be updated accordingly, and then re-integrate the SDEs from t to $t + \tau_j$ rather to $t + \Delta t$.

We use the system of integral algebraic constraints, or jump equations (Salis & Kaznessis, 2005), to monitor occurrence of the reactions in the system that alternate the number of the discrete variables. The jump equation, governing the reaction time of the j-th reaction, is given by

$$\int_{t_0}^{t_0+t} \alpha_j(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{k}) \mathrm{d}s + \ln(URN_j) = R_j|^t, \tag{2.50}$$

where t_0 is the time which the reaction last occurred, URN_j is the j-th uniform random number within $(0, 1)$, and $R_j|^{\tau_j}$ denotes the residue value. There is one jump equation (2.50) for each reaction. One can determine whether the j-th reaction has occurred by simply monitoring the sign of the residual, according to

$$R_j|^t < 0 \rightarrow t < \tau_j, \quad R_j|^t > 0 \rightarrow t > \tau_j, \quad R_j|^t = 0 \rightarrow t = \tau_j,$$

where t is the current simulation time. This suggests that whether the j-th reac-

tion has occurred is determined by a zero crossing of the residue $R_j|^t$. Note that the reaction propensities $\alpha_j(\boldsymbol{x}, \boldsymbol{y})$, $j = 1, 2, \ldots, N$, are time-dependent, because the continuous variables, $\boldsymbol{y}(t)$, vary with time. To compute the j-th reaction time τ_j, we use a Riemann sum to evaluate the integral in (2.50) with time increment Δt from t_0 to t, where t is the time just prior to the j-th residue's zero crossing. We should have $t - t_0 \leq \Delta t$, or otherwise we keep numerical integration and let $t = t + \Delta t$ until it is satisfied. Since the time increment Δt is typically small, we use a first-order approximation to the next reaction time, τ_j, as

$$\tau_j = \frac{-R_j|^t}{\alpha_j(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{k})} + t.$$

This expression is used to compute the reaction time after having monitored the occurrence of the reactions that alternate the number of discrete variables. Once τ_j is computed, we re-integrate the SDEs numerically from $t_0 + t'$ to $t_0 + \tau$, and assign a new random number to URN_j.

The whole procedure, as a variant of the Next Reaction method (Gibson & Bruck, 2000), is illustrated in Algorithm 1. The proposed Next Reaction Hybrid method (NRHM) might be regarded as an extension of the hybrid algorithm previously proposed in (Salis & Kaznessis, 2005). But the difference here is that we have utilised the CLE (2.49) whose the populations of the microscopic species are shifted by a stoichiometric vector, and this correction could be significant for low copy number of microscopic species.

Algorithm 1 Next reaction hybrid algorithm.

Require: $\boldsymbol{x} = \boldsymbol{x}_0$, $\boldsymbol{x}_{\text{last}} = \boldsymbol{x}_0$, $\boldsymbol{y} = \boldsymbol{y}_0$, $\boldsymbol{y}_{\text{last}} = \boldsymbol{y}_0$, $\boldsymbol{k} = \boldsymbol{k}_0$, $\boldsymbol{k}_{\text{last}} = \boldsymbol{k}_0$, $t = t_{\text{start}}$, $t_{\text{last}} = t_{\text{start}}$, $\Delta t = \Delta t_0$, $R = \ln(URN)$, $R_{\text{last}} = \ln(URN)$.

while $t < t_{\text{end}}$ **do**
  Compute propensities, α_j, $j = 1, \ldots, M$;
  Numerically integrate the CLE (2.14) from t to $t + \Delta t$;
  Update the residuals in (2.50) for all reactions that changes discrete variables by $R_j^{\text{new}} = R_j + \alpha_j \Delta t$;
  Count the number of zero crossing so the residuals, $ZC = \text{Count}(R^{\text{new}} \geq 0)$;
  if $ZC = 0$ **then**
    $\boldsymbol{y} \leftarrow \boldsymbol{y}(t + \Delta t)$, $\boldsymbol{k} \leftarrow \boldsymbol{k}(t + \Delta t)$, $t \leftarrow t + \Delta t$;
  else if $ZC = 1$, the μ-th reaction has occurred, **then**
    Compute $\tau_\mu = -R_\mu / \alpha_\mu$;
    Numerically integrate the CLE (2.14) from t to $t + \tau_\mu$;
    Save the state of the system for possible rewind: $\boldsymbol{x}_{\text{last}} \leftarrow \boldsymbol{x}$, $\boldsymbol{y}_{\text{last}} \leftarrow \boldsymbol{y}$, $\boldsymbol{k}_{\text{last}} \leftarrow \boldsymbol{k}$, $t_{\text{last}} \leftarrow t$;
    Update: $\boldsymbol{x} \leftarrow \boldsymbol{x} + \boldsymbol{\xi}_\mu$, $\boldsymbol{y} \leftarrow \boldsymbol{y}(t + \tau_\mu)$, $\boldsymbol{k} \leftarrow \boldsymbol{k}(t + \tau_\mu)$, $t \leftarrow t + \tau_\mu$, $\Delta t \leftarrow \Delta t_0$;
    Reset the μ-th residue: $R_\mu = \ln(URN_\mu)$;
    Update the residue for the rest of the reactions $R_j = R_j + \alpha_j \tau_\mu$, $j \neq \mu$;
  else
    Rewind the state of the system and decrease Δt: $\boldsymbol{x} \leftarrow \boldsymbol{x}_{\text{last}}$, $\boldsymbol{y} \leftarrow \boldsymbol{y}_{\text{last}}$, $\boldsymbol{k} \leftarrow \boldsymbol{k}_{\text{last}}$, $\Delta t = \Delta t/2$;
  end if
end while

2.3.6 Gene regulatory system with feedback

We now consider a variation of scheme (2.34) whereby the switching between the two gene states is mediated by protein binding:

$$S_1 + S_2 \underset{k'_{off}}{\overset{k'_{on}}{\rightleftarrows}} S_1^*, \quad S_1 \xrightarrow{k'_1} S_1 + S_2, \quad S_1^* \xrightarrow{k'_2} S_1^* + S_2, \quad S_2 \xrightarrow{c'} \emptyset. \tag{2.51}$$

This system constitutes a genetic negative feedback loop if $k'_2 < k'_1$ and a positive feedback loop if $k'_2 > k'_1$. An exact solution of the CME for the case of one gene has been reported in (Grima et al., 2012). We study this example numerically using the CME, the CFPE and CCLE for a positive feedback system with N_{gene} genes and the parameter set: $c' = 1$, $k'_{on} = 6.67 \times 10^{-6}$, $k'_{off} = 0.001$, $k'_1 = 50$, $k'_2 = 250$, $\phi_N = 1$ for 3 different volumes $V = 1, 10, 50$. Note that as before the number of genes equals $N_{\text{gene}} = \phi_1 V$ and hence the three volumes chosen correspond to a system with 1, 10 and 50 gene copy numbers, respectively. Note also that the parameters quoted here are the same as the rate constants that would appear in a rate equation formulation,

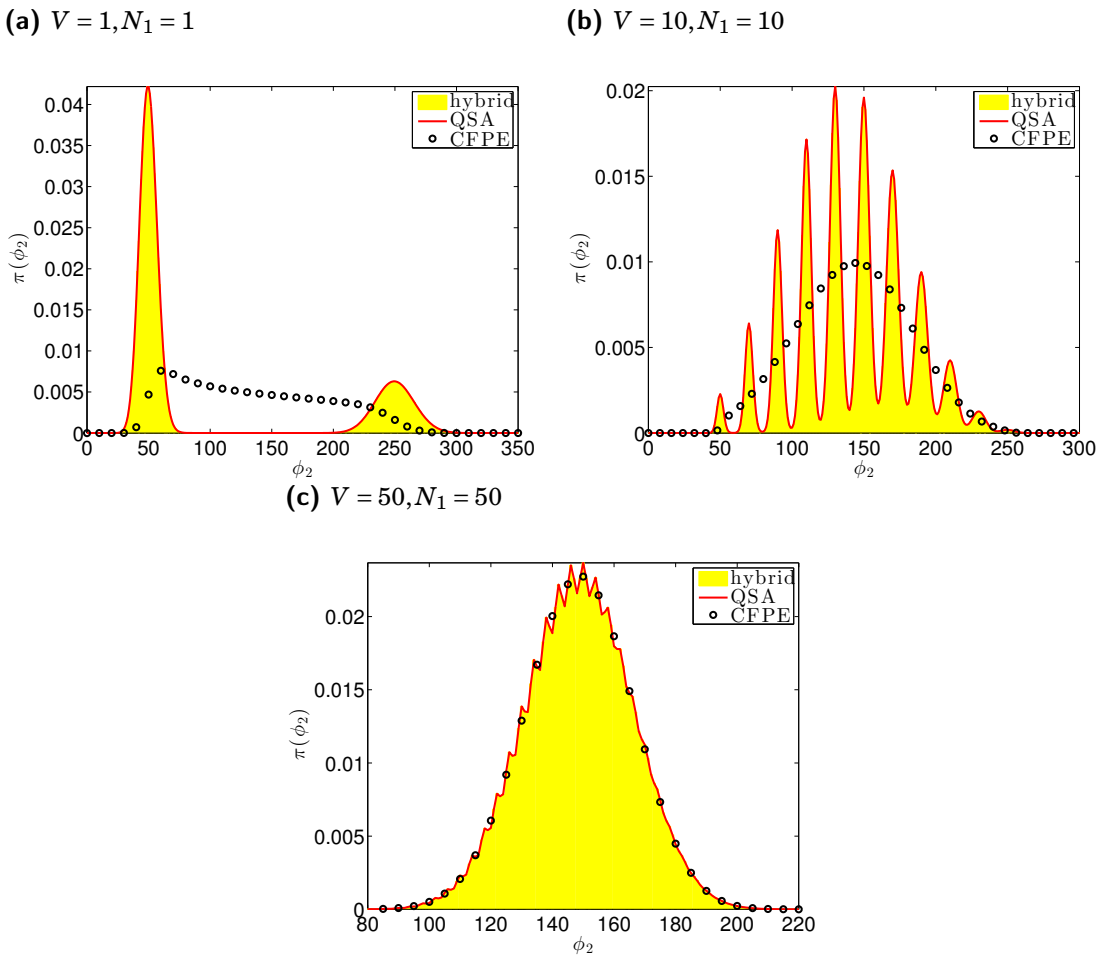

Figure 2.4: *Comparison of the stationary distribution of protein concentration obtained from SSA simulations, the numerical solution of the CFPE for different values of volume, V, and simulations of the hybrid model by NRHM for the genetic feedback loop* (2.51)*. See text for parameter values. Note that as for the gene feedback loop with no feedback (Figure 2.2), the CFPE does not capture the onset of multi-modality as the volume (and the number of gene molecules N_{gene}) is decreased below a critical value.*

and are chosen so that timescale separation conditions are met.

The results of simulations using the SSA, the Galerkin finite element method for the CFPE and simulations of the NRHM are shown and contrasted in Figure 2.4. As for the scheme (2.34) studied in the previous section, the CFPE does not reproduce the noise-induced multi-modality predicted by the CME, and the resulting distribution remains uni-modal. To generate samples of the stationary distribution of the hybrid model (2.48) we used an Euler-Maruyama discretisation with time-step size

$\Delta t = 10^{-2}$ to generate a single realisation of $T = 10^9$ time-units long. Samples of the chain were obtained every 10^2 time-units, from which a histogram of $\phi_2(t)$ was generated. In Figure 2.4, we see very good agreement between the hybrid and corresponding CME distributions for all volume sizes.

"Simplification of modes of proof is not merely an indication of advance in our knowledge of a subject, but is also the surest guarantee of readiness for further progress."

Kelvin & Tait (1873)

3

Tensor-based Computations

This chapter studies the numerical approximation of the steady state distribution of the CFPE,

$$\mathscr{L}_{\mathrm{f}}\pi_{\mathrm{f}}(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \Omega_{\mathrm{f}}, \quad \text{and} \quad \int_{\Omega_{\mathrm{f}}} \pi_{\mathrm{f}}(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = 1, \tag{3.1}$$

using the tensor methods. Here, the CFPE operator $\mathscr{L}_{\mathrm{f}}$ is defined in (2.10), and π_{f} is its eigenfunction (kernel) corresponding to the zero eigenvalue.

3.1 Numerical approximations of the stationary CFPE

In the current section, we study two essential steps in the numerical considerations of (3.1): the domain truncation in Section § 3.1.1 and the finite difference discretisation in Section § 3.1.2.

3.1.1 Artificial domain truncation

The (unbounded) domain of definition Ω_f needs to be truncated into some bounded computational domain $\Omega \subset \Omega_f$ such that numerical discretisations can be imposed. Usually, domain Ω covers the region in which the vast majority of the probability density sits, and the truncated version of (3.2) becomes

$$\begin{cases} \mathscr{L}_f P_\Omega(\boldsymbol{x}) = \lambda(\mathscr{L}_f, \Omega) P_\Omega(\boldsymbol{x}) & \boldsymbol{x} \in \Omega, \\ P_\Omega(\boldsymbol{x}) = 0 & \boldsymbol{x} \in \partial\Omega, \\ \int_\Omega P_\Omega(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} = 1, \end{cases} \tag{3.2}$$

where $\lambda(\mathscr{L}_f, \Omega)$ and P_Ω represent the principal eigenvalue and eigenfunction of $\mathscr{L}_f$. In this section, we study the convergence of P_Ω in (3.2) towards π_f in (3.1) as $\Omega \to \Omega_f$.

We start with a notion of a generalised principal eigenvalue (Berestycki et al., 1994):

$$\lambda(\mathscr{L}_f, \Omega) := \sup \left\{ \lambda : \exists P_\Omega \in W^{2,N}_{\mathrm{loc}}(\Omega),\ P_\Omega > 0,\ (\mathscr{L}_f - \lambda) P_\Omega \leq 0 \text{ in } \Omega \right\}. \tag{3.3}$$

For Ω bounded and smooth, the above definition of $\lambda(\mathscr{L}_f, \Omega)$ coincides with the classic Dirichlet principle eigenvalue in (3.2) (Berestycki & Rossi, 2015). From the Krein-Rutman theory (Krein & Rutman, 1948), there exists a unique real principal eigenvalue $\lambda(\mathscr{L}_f, \Omega)$, and it admits a positive eigenfunction $P_\Omega \in W^{2,p}(\Omega)$ which is unique up to a constant multiple.

For unbounded Ω, it is possible that $\lambda(\mathscr{L}_f, \Omega)$ does not admit any bounded principal eigenfunction, or $\lambda(\mathscr{L}_f, \Omega)$ is not simple and the associated eigenfunction may not be unique (Berestycki & Rossi, 2015). We assume the principal eigenfunction π_f is bounded and unique, and first present some previously known properties of $\lambda(\mathscr{L}_f, \Omega_f)$.

Proposition 3.1. *The principal eigenvalue $\lambda(\mathscr{L}_f, \Omega)$ defined in* (3.3) *satisfies the fol-*

lowing properties:

(*i*) *If* $\Omega' \subset \Omega \subset \Omega_{\mathrm{f}}$*, then* $\lambda(\mathscr{L}_{\mathrm{f}}, \Omega') \leq \lambda(\mathscr{L}_{\mathrm{f}}, \Omega)$*, with strictly inequality if* Ω' *is bounded and* $\left|\Omega \backslash \Omega'\right| > 0$.

(*ii*) *If* $(\Omega_n)_{n\in\mathbb{N}}$ *is a family of nonempty domains such that*

$$\Omega_n \subset \Omega_{n+1}, \quad \bigcup_{n\in\mathbb{N}} \Omega_n = \Omega, \quad \Omega \subset \Omega_{\mathrm{f}},$$

then $\lambda(\mathscr{L}_{\mathrm{f}}, \Omega_n) \uparrow \lambda(\mathscr{L}_{\mathrm{f}}, \Omega)$ *as* $n \to \infty$*. Reversely, if a family of the domains* $(\Omega'_n)_{n\in\mathbb{N}}$ *satisfies* $\Omega'_1 \backslash \Omega'$ *being bounded, and*

$$\Omega'_n \supset \Omega'_{n+1}, \quad \bigcap_{n\in\mathbb{N}} \Omega'_n = \Omega', \quad \Omega'_1 \subset \Omega_{\mathrm{f}},$$

then $\lambda(\mathscr{L}_{\mathrm{f}}, \Omega'_n) \downarrow \lambda(\mathscr{L}_{\mathrm{f}}, \Omega')$ *as* $n \to \infty$.

(*iii*) *If* $\lambda(\mathscr{L}_{\mathrm{f}}, \Omega) < \infty$*, then there exists a positive function* $P_\Omega \in W^{2,p}_{\mathrm{loc}}$*,* $p < \infty$ *satisfying*

$$\mathscr{L}_{\mathrm{f}} P_\Omega = \lambda(\mathscr{L}_{\mathrm{f}}, \Omega) P_\Omega \quad \text{a.e. in } \Omega. \tag{3.4}$$

and if Ω *is bounded, then* P_Ω *is unique up to a multiplicative constant.*

(*iv*) *If the domain* Ω *and a family of bounded domain* $\{\Omega_n\}_{n\in\mathbb{N}}$ *as in (iii), then the corresponding principal eigenfunctions* P_{Ω_n} *of* $\mathscr{L}_{\mathrm{f}}$ *in* Ω_n *by* (3.4)*, normalised by* $P_{\Omega_n}(\boldsymbol{x}_0) = 1$ *for a given* $\boldsymbol{x}_0 \in \Omega$*, converge (up to subsequences) in* $\mathscr{C}^1_{\mathrm{loc}}(\bar{\Omega})$ *to an eigenfunction* P_Ω *with eigenvalue* $\lambda(\mathscr{L}_{\mathrm{f}}, \Omega)$.

The arguments in Proposition 3.1 should all be understood on the fact that the ellipticity condition (2.12) is satisfied. The proof of Proposition 3.1 can be found in Berestycki & Rossi (2015), more specifically, the second part of property (*ii*) is Theorem 1.10, property (*iv*) is given in the proof of Theorem 1.4, the uniqueness of property (*iii*) is provided by Property A.1, and the other properties can be found in Proposition 2.3.

Based on the instruments given in Proposition 3.1, we can proceed to show the convergence of P_Ω to π_f in the following theorem.

Theorem 3.2. *Consider the Dirichlet eigenvalue problem* (3.2) *defined on a bounded domain* $\Omega \subset \Omega_\mathrm{f}$. *If the stationary CFPE* (2.13) *admits an unique bounded solution up to a constant multiplication with vanishing values along* Ω_f, *then we have*

$$\lambda(\mathscr{L}_\mathrm{f}, \Omega) < 0, \qquad \text{and} \qquad \lambda(\mathscr{L}_\mathrm{f}, \Omega) \uparrow 0 \ \text{ as } \ \Omega \to \Omega_\mathrm{f}.$$

Further, if both the eigenfunctions P_Ω *and* π_f *are normalised so that* $P_\Omega(\boldsymbol{x}) = \pi_\mathrm{f}(\boldsymbol{x}) = 1$ *for a given* $\boldsymbol{x} \in \Omega$, *we have the convergence* $P_\Omega \to \pi_\mathrm{f}$ *in* Ω *as* $\Omega \to \Omega_\mathrm{f}$.

Proof. By assuming the existence and uniqueness of the solution π_f in the stationary CFPE (2.13), we directly have the generalised principal eigenvalue $\lambda(\mathscr{L}_\mathrm{f}, \Omega_\mathrm{f}) = 0$, and the existence and uniqueness of a bounded and positive eigenfunction $P_{\Omega_\mathrm{f}} = \pi_\mathrm{f}$ corresponding to $\lambda(\mathscr{L}_\mathrm{f}, \Omega_\mathrm{f})$ in (3.3). Then, Proposition 3.1(*ii*) implies a negative principal eigenvalues, i.e., $\lambda(\mathscr{L}_\mathrm{f}, \Omega_\mathrm{f}) < 0$. From Proposition 3.1(*ii*), we have the eigenvalue convergence $\lambda(\mathscr{L}_\mathrm{f}, \Omega) \uparrow 0$ as $\Omega \to \Omega_\mathrm{f}$. Using Proposition 3.1(*iv*), we obtain the eigenfunction convergence, and this completes the proof. □

Theorem 3.2 suggests that the principal eigenvalue should be shifted to some negative value if the original domain Ω_f is truncated with the artificial vanishing boundary condition. This implies that there exists no non-trivial stationary distribution for the time-dependent CFPE (2.10) in the truncated domain Ω. But the shape of the stationary distribution π_f in Ω can still be approximated by the principal eigenfunction P_Ω. The accuracy of the approximation can be measured by the magnitude of $\lambda(\mathscr{L}_\mathrm{f}, \Omega)$, and from 3.1(*i*), increasing the domain size Ω would yield a more accurate approximation.

3.1.2 Numerical discretisation

Now we apply the finite difference method (FDM) to discretise the Dirichlet eigenvalue problem (3.2) into a matrix principal eigenvalue problem,

$$\boldsymbol{A_h p_h} = \lambda_{\boldsymbol{h}} \boldsymbol{p_h}, \tag{3.5}$$

where $(\lambda_{\boldsymbol{h}}, \boldsymbol{p_h})$ is the principal eigen-pair of the stiffness matrix $\boldsymbol{A_h}$. For self-adjoint operators, the convergence of the FDM has been extensively studied (Keller, 1965; Gary, 1965; Kuttler, 1970). But for non-self-adjoint operators, it is not possible to build a monotone and converging difference scheme using a narrow stencil (Carasso, 1969; Feng et al., 2013).

Recently, monotone difference schemes for elliptic operators with mixed derivatives have been proposed (Samarskii et al., 2002; Rybak, 2004; Matus & Rybak, 2004). These schemes can be directly applied to discretise the CFPEs in lower dimensions. But the resulting stiffness matrix can not be constructed from the formatted tensors, and thus can not be used in higher dimensions. In the current section, we extend the existing scheme to a tensor-friendly version, and the resulting stiffness matrix can be exactly constructed by the separated tensors. We also show the scheme yields an irreducible M-matrix, and is second-order accurate.

Note that we will follow the traditional matrix-vector-based notation in the current section that the nodal points are enumerated by one index and the corresponding discretised functions are stacked into a "long" vector, but we will consider storage compression by tensor representation in the later sections.

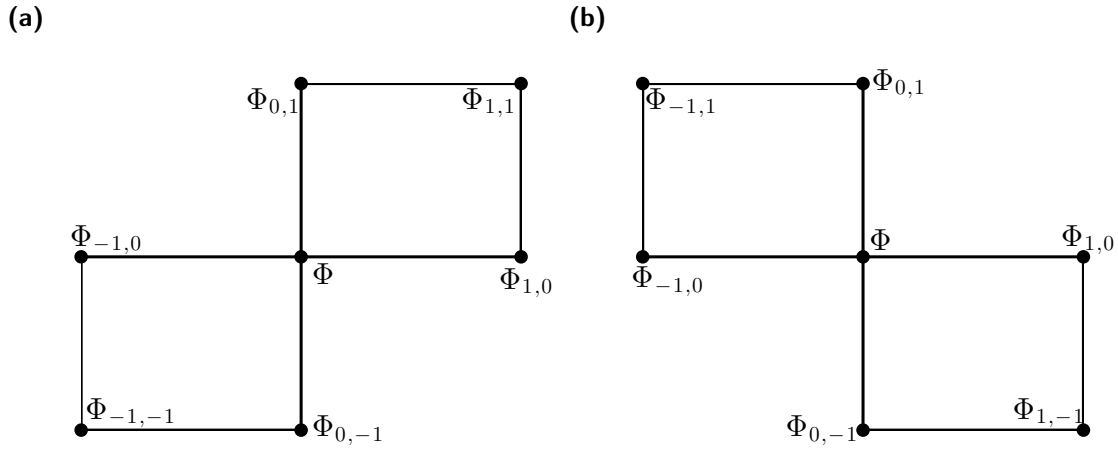

Figure 3.1: *2D visual illustration of the seven-point stencil of the difference scheme* (3.9)*, with* **(a)** *for* Λ^+ *and* **(b)** *for* Λ^-.

3.1.2.1 A finite difference scheme

In an N-dimensional hypercube $\Omega = \mathscr{I}_1 \times \cdots \times \mathscr{I}_N$, $\mathscr{I}_d = [a_d, b_d]$, $d = 1,2,\ldots,N$, we consider the uniform grid $\omega_{\boldsymbol{h}}$:

$$\omega_{\boldsymbol{h}} = \left\{ \boldsymbol{x}_{i_1,\ldots,i_N} = (x_{1,i_1},\ldots,x_{N,i_N}) : x_{d,i_d} = a_d + i_d h_d, i_d = 1,\ldots,n_d, d = 1,\ldots,N \right\}, \quad (3.6)$$

with constant grid step $h_d = (b_d - a_d)/(n_d + 1)$. The set of boundary grid points is given by

$$\tilde{\omega}_{\boldsymbol{h}} = \{ \boldsymbol{x}_{i_1,\ldots,i_N} = (x_{1,i_1},\ldots,x_{N,i_N}) : i_d = 0 \text{ or } n_d + 1, d = 1,\ldots N \}.$$

We choose the computational domain $\Omega \subset \mathbb{R}^N_+$ such that all the propensity functions, α_j, $j = 1,\ldots,M$, in (2.2) are non-negative.

Let us construct a difference scheme taking account of of the sign of the diffusion coefficients. We will use the following notation conventions of difference operator. For any function $w : \Omega \mapsto \mathbb{R}$ and any interior base-point $(x_{1,i_1},\ldots,x_{N,i_N}) \in \omega_{\boldsymbol{h}}$, we use

w to denote the nodal value, $w(x_{1,i_1},\ldots,x_{N,i_N})$, and

$$w^{(\pm 1d)} = w(x_{1,i_1},\ldots,x_{d,i_d}\pm h_d,\ldots,x_{N,i_N}),$$

is the value of w at the adjacent nodal point in the positive/negative direction of x_d. One-sided differences, as a replacement of the first derivative $\partial w/\partial x_d$, are denoted by

$$w_{x_d} = \frac{w^{(+1d)}-w}{h_d}, \quad \text{and} \quad w_{\bar{x}_d} = \frac{w-w^{(-1d)}}{h_d}, \tag{3.7}$$

for forward difference and backward difference schemes, respectively. We can approximate the second order differential operators with mixed derivatives in the CFPE (2.13),

$$\mathscr{L}_{d_1,d_2}(w) = \frac{\partial^2 D_{d_1,d_2}(\boldsymbol{x})w}{\partial x_{d_1}\partial x_{d_2}}, \quad d_1\neq d_2, \tag{3.8}$$

by the following difference operators

$$\begin{aligned}\Lambda^+_{d_1,d_2}(w) &= \frac{((D^+_{d_1,d_2}w)_{x_{d_1}})_{x_{d_2}} + ((D^+_{d_1,d_2}w)_{\bar{x}_{d_1}})_{\bar{x}_{d_2}}}{2},\\ \Lambda^-_{d_1,d_2}(w) &= \frac{((D^-_{d_1,d_2}w)_{x_{d_1}})_{\bar{x}_{d_2}} + ((D^-_{d_1,d_2}w)_{\bar{x}_{d_1}})_{x_{d_2}}}{2},\end{aligned} \tag{3.9}$$

for $d_1\neq d_2$, $d_1,d_2=1,\ldots,N$, where $D^{\pm}_{d_1,d_2}$ represent the partial sums of the positive and negative terms in (2.11), respectively, i.e.,

$$D^+_{d_1,d_2} = \frac{1}{2}\sum_{r=1}^{M} \nu_{r,d_1}\nu_{r,d_2}\alpha_{d_1,d_2}\cdot \mathbb{1}_{\nu_{r,d_1}\nu_{r,d_2}>0}, \quad D^-_{d_1,d_2} = D_{d_1,d_2} - D^+_{d_1,d_2}. \tag{3.10}$$

The difference operators in (3.9) is defined on a seven-point stencil illustrated in Figure 3.1. If w is a $\mathscr{C}^4$ continuous function, it can be shown that the difference operator $\Lambda^{\pm}_{d_1,d_2}$ is a second order approximation of the differential operator $\mathscr{L}_{d_1,d_2}$,

i.e.,

$$\Lambda^{\pm}_{d_1,d_2} w - \mathscr{L}_{d_1,d_2} w = \mathscr{O}(|\boldsymbol{h}|^2), \qquad \boldsymbol{h} = (h_1, h_2, \ldots, h_N)^T.$$

In case of the second order derivative operator in a single variable, $\mathscr{L}_{d,d}$, defined similarly as in (3.8), we replace it by the standard three-point scheme,

$$\Lambda_{d,d}(w) = ((D_{d,d} w)_{\bar{x}_d})_{x_d}, \quad d = 1, \ldots, N. \tag{3.11}$$

The first order derivative operators in the drift terms of the CFPE (2.10) are approximated by the central difference scheme

$$\Lambda_d(w) = \frac{(\mu_d w)_{x_d} + (\mu_d w)_{\bar{x}_d}}{2}, \quad d = 1, \ldots, N, \tag{3.12}$$

which is known to be of second order accuracy.

Combining the above difference operators, we approximate the CFPE eigenvalue problem (3.2) by the difference scheme

$$\boldsymbol{A}_{\boldsymbol{h}} \boldsymbol{p}_{\boldsymbol{h}} = \sum_{d=1}^{N} \left(\Lambda_d \left(\boldsymbol{p}_{\boldsymbol{h}} \right) + \Lambda_{d,d} \left(\boldsymbol{p}_{\boldsymbol{h}} \right) \right) + \sum_{\substack{d_1, d_2 = 1 \\ d_1 \neq d_2}}^{N} \left(\Lambda^{+}_{d_1,d_2} \left(\boldsymbol{p}_{\boldsymbol{h}} \right) + \Lambda^{-}_{d_1,d_2} \left(\boldsymbol{p}_{\boldsymbol{h}} \right) \right) = \lambda_{\boldsymbol{h}} \boldsymbol{p}_{\boldsymbol{h}}, \tag{3.13}$$

where $\boldsymbol{A}_{\boldsymbol{h}} \in \mathbb{R}^{(n_1 \cdots n_N) \times (n_1 \cdots n_N)}$ denotes a large (multi-dimensional) matrix, and $(\lambda_{\boldsymbol{h}}, \boldsymbol{p}_{\boldsymbol{h}})$ is the associate principal eigenpair. The stencils of difference schemes in (3.9), (3.11) and (3.12), are compact, using only $(3^N - 1)$ surrounding nodes for their discretisations in N dimensions. Due to the homogeneous Dirichlet boundary condition in (3.2), the boundary nodes $\hat{\omega}_{\boldsymbol{h}}$ are not included in $\boldsymbol{A}_{\boldsymbol{h}}$.

3.1.2.2 Monotonicity

Next we show the stiffness matrix $\boldsymbol{A_h}$ possesses the similar properties of the original operator $\mathscr{L}_\mathrm{f}$, such as the simplicity of the principal eigenvalue and the positivity of the corresponding eigenvector. We first define the notion of the L-matrix and M-matrix (Ostrowski, 1937).

Definition 3.3 (Layton & Morley (1987))**.** *An $n \times n$ matrix $\boldsymbol{A}$ is called*

(*i*) *a monotone matrix if $\boldsymbol{A}^{-1}$ exists and $\boldsymbol{Ax} \le \boldsymbol{Ay}$ implies $\boldsymbol{x} \le \boldsymbol{y}$; equivalently, if $\boldsymbol{A}^{-1} \ge 0$;*

(*ii*) *an L-matrix if $\boldsymbol{A} = (A_{i,j})_{n\times n}$ with $A_{i,i} \ge 0$ and $A_{i,j} \le 0$ for $i \ne j$;*

(*iii*) *an M-matrix if $\boldsymbol{A}$ is both an L-matrix and a monotone matrix.*

Here all matrix and vector inequalities should be interpreted element-wise. Next we show the stiffness matrix is an irreducible L-matrix.

Lemma 3.4. *For sufficiently small grid size h_d, $d = 1,2,\ldots,N$, if the following conditions are satisfied:*

$$\frac{D_{d,d}}{h_d} > \sum_{\substack{d'=1\\ d'\ne d}}^{N} \frac{D^+_{d,d'} + D^+_{d',d} - (D^-_{d,d'} + D^-_{d',d})}{2h_{d'}} \quad \text{for all } \boldsymbol{x} \in \omega_{\boldsymbol{h}}, \tag{3.14}$$

then the negative stiffness matrix $(-\boldsymbol{A_h})$ is an irreducible L-matrix.

Proof. Let Φ_0 be the coefficient at the central node $\boldsymbol{x} \in \omega_h$, $\Phi_{\pm\ell}$ be the coefficient at the surface node $(x_{1,i_1},\ldots,x_{\ell,n_\ell} \pm h_\ell,\ldots,x_{N,i_N})$. Let $\Phi_{\pm\ell_1,\pm\ell_2}$ be the coefficient at the corner node $(x_{1,i_1},\ldots,x_{\ell_1,n_{\ell_1}} \pm h_{\ell_1},\ldots,x_{\ell_2,n_{\ell_2}} \pm h_{\ell_2},\ldots,x_{N,i_N})$. Then entries of $\boldsymbol{A_h}$ are

given by

$$\begin{aligned}
\Phi_0 &= -\sum_{d=1}^{N} \frac{2D_{d,d}}{h_d^2} + \sum_{\substack{d_1,d_2=1 \\ d_1\neq d_2}}^{N} \frac{D^+_{d_1,d_2} + D^+_{d_2,d_1} - (D^-_{d_1,d_2} + D^-_{d_2,d_1})}{h_{d_1} h_{d_2}}, \\
\Phi_{\pm\ell} &= \frac{D_{\ell,\ell}}{h_\ell^2} - \sum_{\substack{\ell'=1 \\ \ell'\neq\ell}}^{N} \frac{D^+_{\ell,\ell'} + D^+_{\ell',\ell} - (D^-_{\ell,\ell'} + D^-_{\ell',\ell})}{2h_\ell h_{\ell'}} \pm \frac{\mu_\ell}{2h_\ell}, \\
\Phi_{\pm\ell_1,\pm\ell_2} &= \frac{D^+_{\ell_1,\ell_2} + D^+_{\ell_2,\ell_1}}{2h_{\ell_1} h_{\ell_2}}, \quad \text{and} \quad \Phi_{\pm\ell_1,\mp\ell_2} = -\frac{D^-_{\ell_1,\ell_2} + D^-_{\ell_2,\ell_1}}{2h_{\ell_1} h_{\ell_2}}.
\end{aligned} \tag{3.15}$$

Here, Φ_0 refers to the diagonal values of $\boldsymbol{A_h}$, and the other coefficients corresponds to the off-diagonals.

According to Definition 3.3, matrix $(-\boldsymbol{A_h})$ is an L-matrix if $\Phi_0 \le 0$ and $\Phi_{\pm\ell}$, $\Phi_{\pm\ell_1,\pm\ell_2}$, $\Phi_{\pm\ell_1,\mp\ell_2} \ge 0$. By the splitting (3.10), we have the corner nodes, $\Phi_{\pm\ell_1,\pm\ell_2}$ and $\Phi_{\pm\ell_1,\mp\ell_2}$, to be non-negative. Then, for sufficiently small grid size h_d, $d = 1,\ldots,N$, where the diffusion part dominates the drift part, the positivity of $\Phi_{\pm\ell}$, as well as the negativity of Φ_0, is directly satisfied by the condition (3.14). □

Remark 3.5. Let $h_1 = h_2 = \ldots = h_N = h$ in (3.14). If stoichiometric coefficients of (2.1) satisfy the following the following two conditions: 1) for all $i = 1,2,\ldots,N$ and $j = 1,2,\ldots,M$,

$$2\nu_{j,i}^2 \ge \sum_{\substack{i'=1 \\ i'\neq i}}^{N} |\nu_{j,i}\nu_{j,i'}|, \tag{3.16}$$

and 2) there exists at least one combination of $\hat{i} \in [1,N]$ and $\hat{j} \in [1,M]$ that the above inequality is strictly satisfied, i.e.,

$$2\nu_{\hat{j},\hat{i}}^2 > \sum_{\substack{i'=1 \\ i'\neq \hat{i}}}^{N} |\nu_{\hat{j},\hat{i}}\nu_{\hat{j},i'}|, \tag{3.17}$$

then the condition (3.14) can always be satisfied by choosing a sufficiently small h.

Proof. Remark 3.5 can be proved directly by first multiplying the propensity function α_j to the both sides of (3.16), summing over j, and applying (3.17), which yields

$$2\sum_{j=1}^{M} \nu_{j,i}^2 \alpha_j > \sum_{j=1}^{M} \sum_{\substack{i'=1 \\ i' \neq i}}^{N} |\nu_{j,i}\nu_{j,i'}| \alpha_j.$$

Then, assuming the grid sizes h_i, $i = 1,2,\ldots,N$, are equal for all dimensions, the above formula satisfies the condition (3.14). □

For $N = 1$, the conditions are satisfied directly. For $N > 1$, Remark 3.5 suggests the following combined necessary conditions for the negative stiffness matrix as an irreducible L-matrix:

1) each chemical reaction leads to the population changes in at most two chemical species,
2) if two chemical species changes population numbers, the absolute changes should be equal,
3) at least one chemical reaction only changes a single species' population.

Conditions 1) and 2) follow (3.16), and condition 3) follows (3.17). One basic example is the isometrization chain system, which will be presented in (3.99).

Proposition 3.6 (Layton & Morley (1987))**.** *Suppose that* $\boldsymbol{A}$ *is an irreducible L-matrix. If* $\boldsymbol{A}$ *is diagonally semi-dominant with at least one row of* $\boldsymbol{A}$ *strictly dominant – that is, for some* i*,*

$$A_{i,i} - \sum_{\substack{j=1 \\ j \neq i}} \left|A_{i,j}\right| > 0, \tag{3.18}$$

then $\boldsymbol{A}$ *is an M-matrix and* $\boldsymbol{A}^{-1} > 0$*.*

Theorem 3.7. *If the conditions in Lemma 3.4 are satisfied, then the negative stiffness matrix* $(-\boldsymbol{A_h})$ *in* (3.13) *is an irreducible M-matrix, and the underlying difference*

scheme is a monotone scheme. The principal eigen-pair $(\lambda_{\boldsymbol{h}}, \boldsymbol{p}_{\boldsymbol{h}})$ *of* $\boldsymbol{A}_{\boldsymbol{h}}$ *has the following properties:*

(*i*) $\lambda_{\boldsymbol{h}}$ *is real and algebraically (and geometrically) simple;*

(*ii*) $Re(\lambda_{\boldsymbol{h}}) < 0$ *and* $\left|\lambda_{\boldsymbol{h}}^{\ell}\right| \geq |\lambda_{\boldsymbol{h}}|$ *for every eigenvalue* $\lambda_{\boldsymbol{h}}^{\ell}$ *of* $\boldsymbol{A}_{\boldsymbol{h}}$, $\ell = 1, 2, \ldots, n_1 \cdots n_N$;

(*iii*) $\boldsymbol{p}_{\boldsymbol{h}}(\boldsymbol{x}) > 0$ *for all* $\boldsymbol{x} \in \omega_{\boldsymbol{h}}$.

Proof. If $(-\boldsymbol{A}_{\boldsymbol{h}})$ is an irreducible M-matrix, we have the monotonicity feature by Definition 3.3, and the properties (*i*)–(*iii*) follow from Person's theorem for positive matrices (Horn & Johnson, 2012).

An intuitive way to prove $(-\boldsymbol{A}_{\boldsymbol{h}})$ to be an M-matrix is to show $\boldsymbol{A}_{\boldsymbol{h}}$ is diagonally dominant (3.18). But this is not obvious at all, because the row entries given in (3.15) are all state-dependent, and coefficients on different base-points do not cancel each other exactly. Instead, we here use the fact the transpose of an M-matrix is also an M-matrix (Soto, 2001; Yip, 1996), and try to prove the transpose $(-\boldsymbol{A}_{\boldsymbol{h}}^T)$ is an M-matrix.

It is important to note that the transposed matrix $\boldsymbol{A}_{\boldsymbol{h}}^T$ is the stiffness matrix of the finite difference approximations of the adjoint Fokker-Planck operator $\mathscr{L}_{\mathrm{f}}^*$ defined as

$$\mathscr{L}_{\mathrm{f}}^* P = \sum_{i,j=1}^{N} D_{i,j} \frac{\partial^2 P}{\partial x_i \partial x_j} + \sum_{i=1}^{N} \mu_i \frac{\partial P}{\partial x_i},$$

for function $P \in \mathscr{C}^2(\Omega)$. The difference scheme of approximating $\mathscr{L}_{\mathrm{f}}^*$ should be analogous to the scheme for $\mathscr{L}_{\mathrm{f}}$. We write it as

$$\begin{aligned} \tilde{\Lambda}_{d_1,d_2}^{\pm}(w) &= \frac{\tilde{D}_{d_1,d_2}^{\pm}(w_{x_{d_1}})_{x_{d_2}} + \tilde{D}_{d_1,d_2}^{\pm}(w_{\bar{x}_{d_1}})_{\bar{x}_{d_2}}}{2}, \quad d_1 \neq d_2, \;\; d_1, d_2 = 1, 2, \ldots, N, \\ \tilde{\Lambda}_{d,d}(w) &= \tilde{D}_{d,d}(w_{\bar{x}_d})_{x_d}, \quad \tilde{\Lambda}_d(w) = \frac{\tilde{\mu}_d w_{x_d} + \tilde{\mu}_d w_{\bar{x}_d}}{2}, \quad d = 1, \ldots, N, \end{aligned} \tag{3.19}$$

where the diffusion and drift coefficients, $\tilde{D}_{d_1,d_2}^{\pm}$ and $\tilde{\mu}_d$, are defined similarly as in (3.9), but these functions are all evaluated at the central base-point. We then write

the stiffness matrix $\boldsymbol{A}_{\boldsymbol{h}}^T$ as a sum of diffusion part and drift part, i.e.,

$$\boldsymbol{A}_{\boldsymbol{h}}^T = \boldsymbol{B}_{\boldsymbol{h}}^T + \boldsymbol{C}_{\boldsymbol{h}}^T,$$

where the diffusion and drift parts are respectively given by

$$\boldsymbol{B}_{\boldsymbol{h}}^T w = \sum_{d=1}^{N} \tilde{\Lambda}_{d,d}(w) + \sum_{\substack{d_1,d_2=1 \\ d_1 \neq d_2}}^{N} (\Lambda_{d_1,d_2}^{+}(w) + \Lambda_{d_1,d_2}^{-}(w)), \quad \text{and} \quad \boldsymbol{C}_{\boldsymbol{h}}^T w = \sum_{d=1}^{N} \tilde{\Lambda}_d(w).$$

It is easily seen that matrix $\boldsymbol{B}_{\boldsymbol{h}}^T$ is of order h_d^{-2}, while matrix $\boldsymbol{C}_{\boldsymbol{h}}^T$ is of order h_d^{-1}. Thus, we only need to prove $(-\boldsymbol{B}_{\boldsymbol{h}}^T)$ is an M-matrix, and then, for sufficiently small mesh size h_d, $d = 1,2,\ldots,N$, matrix $(-\boldsymbol{A}_{\boldsymbol{h}}^T)$ is an M-matrix (Alfa et al., 2002).

We next study the entries of $\boldsymbol{B}_{\boldsymbol{h}}^T$. Similarly as in (3.15), we define $\tilde{\Phi}_0$ to be the coefficients at the central node, $\tilde{\Phi}_{\pm d}$ coefficients at the edge nodes, and $\tilde{\Phi}_{\pm d_1,\pm d_2}$ and $\tilde{\Phi}_{\pm d_1,\mp d_2}$ to be the coefficients at the corner nodes. The entries of matrix $\boldsymbol{B}_{\boldsymbol{h}}^T$ are given by

$$\begin{aligned}
\tilde{\Phi}_0 &= -\sum_{d=1}^{N} \frac{2\tilde{D}_{d,d}}{h_d^2} + \sum_{\substack{d_1,d_2=1 \\ d_1 \neq d_2}}^{N} \frac{\tilde{D}_{d_1,d_2}^{+} + \tilde{D}_{d_2,d_1}^{+} - (\tilde{D}_{d_1,d_2}^{-} + \tilde{D}_{d_2,d_1}^{-})}{h_{d_1} h_{d_2}}, \\
\tilde{\Phi}_{\pm d_1} &= \frac{\tilde{D}_{d_1,d_1}}{h_{d_1}^2} - \sum_{\substack{d_2=1 \\ d_2 \neq d_1}}^{N} \frac{\tilde{D}_{d_1,d_2}^{+} + \tilde{D}_{d_2,d_1}^{+} - (\tilde{D}_{d_1,d_2}^{-} + \tilde{D}_{d_2,d_1}^{-})}{2h_{d_1} h_{d_2}} \\
\tilde{\Phi}_{\pm d_1,\pm d_2} &= \frac{\tilde{D}_{d_1,d_2}^{+} + \tilde{D}_{d_2,d_1}^{+}}{2h_{d_1} h_{d_2}}, \quad \text{and} \quad \tilde{\Phi}_{\pm d_1,\mp d_2} = -\frac{\tilde{D}_{d_1,d_2}^{-} + \tilde{D}_{d_2,d_1}^{-}}{2h_{d_1} h_{d_2}}.
\end{aligned} \tag{3.20}$$

Similarly as the proof of Lemma (3.4), under condition (3.14), matrix $\boldsymbol{B}_{\boldsymbol{h}}^T$ is an irreducible L-matrix.

It remains to prove $(-\boldsymbol{B}_{\boldsymbol{h}}^T)$ satisfies the condition in Proposition 3.6. Let the i-th row of matrix $\boldsymbol{B}_{\boldsymbol{h}}^T$ refer to the central base-point $\boldsymbol{x} \in \omega_{\boldsymbol{h}}$. We define $\mathscr{S}(\boldsymbol{x})$ to be the stencil of the difference scheme (3.19) around the nodal point $\boldsymbol{x}$, and define $\mathscr{S}'(\boldsymbol{x}) = \mathscr{S}(\boldsymbol{x}) \backslash \{\boldsymbol{x}\}$. If $\mathscr{S}'(\boldsymbol{x}) \subset \omega_{\boldsymbol{h}}$, all the surrounding nodes are interior nodes. Let $B_{i,j}$ be the

entry of matrix $(-\boldsymbol{B}_{\boldsymbol{h}}^T)$, we have

$$
\begin{aligned}
B_{i,i} &= -\tilde{\phi}_0, \\
\sum_{\substack{j=1 \\ j\neq i}}^{n} |B_{i,j}| &= -\sum_{d=1}^{N} \left(\tilde{\Phi}_{-d} + \tilde{\Phi}_{d}\right) - \sum_{\substack{d_1,d_2=1 \\ d_1\neq d_2}}^{N} \left(\tilde{\Phi}_{+d_1,+d_2} + \tilde{\Phi}_{-d_1,+d_2} + \tilde{\Phi}_{+d_1,-d_2} + \tilde{\Phi}_{-d_1,-d_2}\right).
\end{aligned}
$$

Using the formulas in (3.20), we can easily verify that the i-th row is diagonally semi-dominant, i.e.,

$$
B_{i,i} - \sum_{\substack{j=1 \\ j\neq i}}^{n} |B_{i,j}| = 0, \quad \mathscr{S}'(\boldsymbol{x}) \subset \omega_h.
$$

In the case $\mathscr{S}'(\boldsymbol{x}) \cap \tilde{\omega}_{\boldsymbol{h}} \neq \emptyset$ where some of the nodes of the stencil locate at the boundary $\tilde{\omega}_{\boldsymbol{h}}$ (or beyond), the corresponding coefficients become zero. Then we have

$$
\sum_{\substack{j=1 \\ j\neq i}}^{n} |B_{i,j}| < -\sum_{d=1}^{N} \left(\tilde{\Phi}_{-d} + \tilde{\Phi}_{d}\right) - \sum_{\substack{d_1,d_2=1 \\ d_1\neq d_2}}^{N} \left(\tilde{\Phi}_{+d_1,+d_2} + \tilde{\Phi}_{-d_1,+d_2} + \tilde{\Phi}_{+d_1,-d_2} + \tilde{\Phi}_{-d_1,-d_2}\right),
$$

meaning that the i-th row becomes strictly diagonally dominant. This completes the proof. □

Remark 3.8. According to Theorem 3.7, the proposed difference scheme is monotonic only for sufficiently small mesh sizes, due to the use of the two-sided central difference scheme (3.12) in approximating the first order derivatives. This condition can be relaxed by using a one-sided difference scheme, and taking account of the sign of the drift coefficients. More precisely, one may approximate $\partial(\mu_d w)/\partial x_d$ by

$$
\Lambda_d(w) = \Lambda_d^+(w) + \Lambda_d^-(w), \quad \Lambda_d^+(w) = \left(\mu_d^+ w\right)_{x_d}, \quad \Lambda_d^-(w) = \left(\mu_d^- w\right)_{\bar{x}_d},
$$

where $\mu^{\pm}$ are defined by

$$\mu_d^+(\boldsymbol{x}) = \sum_{j=1}^{M} \nu_{j,d}\alpha_j(\boldsymbol{x}) \cdot \mathbb{1}_{\nu_{j,d}>0} \quad \mu_d^- = \mu_d - \mu_d^+, \quad i,j = 1,2,\ldots,N.$$

However, it has only the first order accuracy, while the previous proposed scheme is of second order, as studied in the following section.

3.1.2.3 Discretisation error

The current section studies the discretisation error in the proposed difference scheme (3.13). The eigen-pairs of matrix $\boldsymbol{A_h}$ are denoted by $\left(\lambda_{\boldsymbol{h}}^j, \boldsymbol{p_h}\right)$, $j = 1,2,\ldots,n_1\cdots n_N$. The eigenvalues $\left\{\lambda_{\boldsymbol{h}}^j\right\}$ are assumed to be in a decreasing order, $0 > \lambda_{\boldsymbol{h}} > \lambda_{\boldsymbol{h}}^2 \geq \lambda_{\boldsymbol{h}}^3 \geq \cdots$, and they approximate the first $(n_1 \times \cdots \times n_N)$ eigenvalues of operator $\mathscr{L}_{\mathrm{f}}$ in (3.2). The eigenvectors are normalised by $\left\|\boldsymbol{p}_{\boldsymbol{h}}^j\right\|_1 / \prod_i h_i = 1$.

Theorem 3.9. *Let $(\lambda_{\boldsymbol{h}}, \boldsymbol{p_h})$ be principal eigenpairs of $\boldsymbol{A_h}$ with normalisation $\left\|\boldsymbol{p_h}\right\|_1 = 1$. Let $\lambda_\Omega \in \mathbb{R}$, $P_\Omega \in \mathscr{C}^4(\Omega)$ be the principle eigenpairs of $\mathscr{L}_{\mathrm{f}}$ in domain Ω with homogeneous boundary conditions, and let $\boldsymbol{p}_\Omega$ be the vector obtained from P_Ω by grid point evaluation in $\omega_{\boldsymbol{h}}$. Assume $\boldsymbol{p}_\Omega$ normalised so that $\boldsymbol{p}_{\boldsymbol{h}}^T\boldsymbol{p}_\Omega = 1$, and that all the eigenvectors, $\boldsymbol{p}_{\boldsymbol{h}}^j$, $j = 1,2,\ldots,n$, are bounded. Let $h = \max_{1\leq d\leq N} h_d$ satisfy the conditions of Theorem 3.7, then the following inequalities hold,*

$$|\lambda_\Omega - \lambda_{\boldsymbol{h}}| \leq C_1 h^2 \quad \text{and} \quad \left\|\boldsymbol{p}_\Omega - \boldsymbol{p_h}\right\|_\infty \leq C_2 h^2. \tag{3.21}$$

Here, constants C_1 and C_2 depend on the diffusion and drift coefficients, $D_{i,j}$ and μ_i, $i,j = 1,\ldots N$, and the third- and the forth-order derivatives of $\boldsymbol{p_h}$, $\inf_{j\neq 1}|\lambda_\Omega - \lambda_{\boldsymbol{h}}^j|$, and $\left\|(\boldsymbol{p}_{\boldsymbol{h}}^1,\ldots,\boldsymbol{p}_{\boldsymbol{h}}^n)\right\|_\infty$.

Proof. For sufficiently smooth $P_\Omega(\boldsymbol{x})$, the finite difference approximation (3.13) can

be written as

$$\boldsymbol{A}_{\boldsymbol{h}}\boldsymbol{p}_{\Omega} + \tau_{\boldsymbol{h}} = \lambda_{\Omega}\boldsymbol{p}_{\Omega}, \tag{3.22}$$

where $\tau_{\boldsymbol{h}}$ is the truncation error of the difference scheme. One can show, by using Taylor expansion in space (Samarskii et al., 2002), that

$$\|\tau_{\boldsymbol{h}}\|_{\infty} \leq \bar{C}_0 h^2, \tag{3.23}$$

where $\bar{C}_0$ is a constant depending on $D_{i,j}$, μ_i, $i,j = 1,\ldots N$, and the thrid-order and the forth order derivatives of $\boldsymbol{p}_{\Omega}$. Now we write $\boldsymbol{p}_{\Omega}$ as a linear combination of the eigenvectors of $\boldsymbol{A}_{\boldsymbol{h}}$, i.e.,

$$\boldsymbol{p}_{\Omega} = \sum_j c_j \boldsymbol{p}_{\boldsymbol{h}}^j. \tag{3.24}$$

Substituting into (3.22) yields

$$\tau_{\boldsymbol{h}} = (\lambda_{\Omega} - \boldsymbol{A}_{\boldsymbol{h}})\boldsymbol{p}_{\Omega} = \sum_j c_j(\lambda_{\Omega} - \lambda_{\boldsymbol{h}}^j)\boldsymbol{p}_{\boldsymbol{h}}^j. \tag{3.25}$$

Let $\boldsymbol{q}_{\boldsymbol{h}}$ be the normalised principal left eigenvector of $\boldsymbol{A}_{\boldsymbol{h}}$ corresponding to $\lambda_{\boldsymbol{h}}$, then multiply $\boldsymbol{q}_{\boldsymbol{h}}^T$ to both sides of the equation (3.25) and use the relation in (3.23), we have

$$|c_1(\lambda_{\Omega} - \lambda_{\boldsymbol{h}})| \leq \left|\boldsymbol{q}_{\boldsymbol{h}}^T \tau_{\boldsymbol{h}}\right| \leq \left\|\boldsymbol{q}_{\boldsymbol{h}}\right\|_1 \|\tau_{\boldsymbol{h}}\|_{\infty} \leq \bar{C}_1 h^2, \tag{3.26}$$

where $\bar{C}_1$ is a constant depending only on constant $\bar{C}_0$ in (3.23). For $c_1 \neq 0$, we divide both sides of the inequality by $|c_1|$. This gives the convergence rate of the principal eigenvalue in (3.21).

Now we derive the convergence of the eigenvector $\boldsymbol{p_h}$ towards $\boldsymbol{p}_\Omega$. Let

$$\boldsymbol{Q} = \left(\boldsymbol{p_h^1}, \boldsymbol{p_h^2}, \ldots, \boldsymbol{p_h^n}\right),$$

be a matrix whose columns are the right eigenvectors of $\boldsymbol{A_h}$, then the matrix form of (3.24) reads

$$\boldsymbol{p}_\Omega = \boldsymbol{Q}\boldsymbol{c},$$

where $\boldsymbol{c} = (c_1, c_2, \ldots, c_n)^T$ denotes a column vector with entries c_j. Similarly, the matrix form of (3.25) is given by

$$\tau_{\boldsymbol{h}} = (\lambda_\Omega - \boldsymbol{A_h})\boldsymbol{Q}\boldsymbol{c} = \boldsymbol{Q}\Phi_{\boldsymbol{h}}\boldsymbol{c}, \tag{3.27}$$

where $\Phi_{\boldsymbol{h}}$ is a diagonal matrix defined as

$$\Phi_{\boldsymbol{h}} = \begin{pmatrix} \lambda_{\boldsymbol{h}} - \lambda_\Omega & 0 & \cdots & 0 \\ 0 & \lambda_{\boldsymbol{h}}^2 - \lambda_\Omega & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \lambda_{\boldsymbol{h}}^n - \lambda_\Omega \end{pmatrix}_{n\times n}.$$

From Theorem 3.7, $\lambda_{\boldsymbol{h}}$ is a simple eigenvalue of $\boldsymbol{A_h}$, and $\boldsymbol{p_h}$ is linearly independent with $\boldsymbol{p_h^j}$, $j = 2, 3, \ldots$. Thus we can define a transformation that makes $\boldsymbol{p_h^j}$, $j = 2, 3, \ldots$, orthogonal to $\boldsymbol{p_h}$. We define matrix $\boldsymbol{Z} = (Z_{i,j})_{n\times n}$ as

$$\boldsymbol{Z} = \begin{pmatrix} 1 & -\frac{\langle \boldsymbol{p_h}, \boldsymbol{p_h^2}\rangle}{\langle \boldsymbol{p_h}, \boldsymbol{p_h}\rangle} & \cdots & -\frac{\langle \boldsymbol{p_h}, \boldsymbol{p_h^n}\rangle}{\langle \boldsymbol{p_h}, \boldsymbol{p_h}\rangle} \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}_{n\times n},$$

where $\langle \boldsymbol{p_h}, \boldsymbol{p_h^j}\rangle$ denotes the inner product of vectors $\boldsymbol{p_h}$ and $\boldsymbol{p_h^j}$. Let $\tilde{\boldsymbol{Q}} = \boldsymbol{Q}\boldsymbol{Z}$, the

first column of $\tilde{\boldsymbol{Q}}$ is orthogonal to all other columns. Then, multiplying $\tilde{\boldsymbol{Q}}^T$ to both sides of (3.27) gives

$$\tilde{\boldsymbol{Q}}^T \tau_{\boldsymbol{h}} = \tilde{\boldsymbol{Q}}^T \tilde{\boldsymbol{Q}} \boldsymbol{Z}^{-1} \Phi_{\boldsymbol{h}} \boldsymbol{c}. \tag{3.28}$$

From the definition of $\tilde{\boldsymbol{Q}}$, the structure of $\tilde{\boldsymbol{Q}}^T \tilde{\boldsymbol{Q}}$ is of the form

$$\tilde{\boldsymbol{Q}}^T \tilde{\boldsymbol{Q}} = \begin{pmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{0} & \bar{\boldsymbol{Q}}^T \bar{\boldsymbol{Q}} \end{pmatrix}_{n \times n},$$

where $\bar{\boldsymbol{Q}}$ is an sub-matrix of $\tilde{\boldsymbol{Q}}$ with the first column vector removed. It can also be verified that

$$\tilde{\boldsymbol{Q}}^T \tilde{\boldsymbol{Q}} \boldsymbol{Z}^{-1} = \begin{pmatrix} 1 & -\boldsymbol{Z}_{1,2:n} \\ \boldsymbol{0} & \bar{\boldsymbol{Q}}^T \bar{\boldsymbol{Q}} \end{pmatrix}_{n \times n},$$

where $\boldsymbol{Z}_{1,2:n}$ refers to a vector formed by the first row of matrix $\boldsymbol{Z}$ except the first element. Further, since $\Phi_{\boldsymbol{h}}$ is diagonal matrix, we can reduce (3.28) to

$$\bar{\boldsymbol{Q}}^T \tau_{\boldsymbol{h}} = \bar{\boldsymbol{Q}}^T \bar{\boldsymbol{Q}} \bar{\Phi}_{\boldsymbol{h}} \bar{\boldsymbol{c}},$$

where $\bar{\Phi}_{\boldsymbol{h}}$ is a sub-matrix of $\Phi_{\boldsymbol{h}}$ with the first row and the first column removed, and $\bar{\boldsymbol{c}}$ is a sub-vector of $\boldsymbol{c}$ with entries $(c_2, \ldots, c_n)^T$. Then, we have

$$\|\bar{\boldsymbol{c}}\|_\infty \le \left\|\bar{\Phi}_{\boldsymbol{h}}^{-1}\right\|_\infty \left\|(\bar{\boldsymbol{Q}}^T \bar{\boldsymbol{Q}})^{-1} \bar{\boldsymbol{Q}}^T\right\|_\infty \|\tau_{\boldsymbol{h}}\|_\infty \le \bar{C}_2 \left\|\bar{\Phi}_{\boldsymbol{h}}^{-1}\right\|_\infty h^2, \tag{3.29}$$

with constant $\bar{C}_2 > 0$, depending on the ∞-norm of matrices $\bar{\boldsymbol{Q}}$ and $\bar{\boldsymbol{Q}}^{-1}$, and constant $\bar{C}_0$ in (3.23). Here we have assumed that eigenvectors $\boldsymbol{p}_{\boldsymbol{h}}^j$ have bounded entries, i.e.,

$\left\| \boldsymbol{p}_{\boldsymbol{h}}^{j} \right\|_{\infty} < \infty$. In above inequality, we notice that

$$\left\| \bar{\Phi}_{\boldsymbol{h}}^{-1} \right\|_{\infty} = \max_{j=2,\ldots,n} \left| \lambda_{\Omega} - \lambda_{\boldsymbol{h}}^{j} \right|^{-1}.$$

Since we have proved the convergence of the eigenvalue $\lambda_{\boldsymbol{h}}$ towards λ_{Ω}, for sufficiently small h, there exists a positive constant $\tilde{C}_2$, such that

$$\inf_{j \neq 1} |\lambda_{\Omega} - \lambda_{\boldsymbol{h}}^{j}| \geq \tilde{C}_2 > 0.$$

We can further bound $\|\bar{\boldsymbol{c}}\|_{\infty}$ in (3.29) by

$$\|\bar{\boldsymbol{c}}\|_{\infty} \leq \frac{\bar{C}_2}{\tilde{C}_2} h^2 = C_2' h^2,$$

where $C_2' = \bar{C}_2 / \tilde{C}_2$ is a positive constant. Since we have assumed the bounded eigenvectors, it then follows that

$$\|\boldsymbol{p}_{\Omega} - \boldsymbol{p}_{\boldsymbol{h}}\|_{\infty} = \| \sum_{j=1}^{n} c_j \boldsymbol{p}_{\boldsymbol{h}}^{j} - \boldsymbol{p}_{\boldsymbol{h}} \|_{\infty} \leq \|\bar{\boldsymbol{c}}\|_{\infty} \|\boldsymbol{Q}\|_{\infty} \leq C_2 h^2,$$

where constant C_2 depends on C_2' and $\|\boldsymbol{Q}\|_{\infty}$. This completes the proof. □

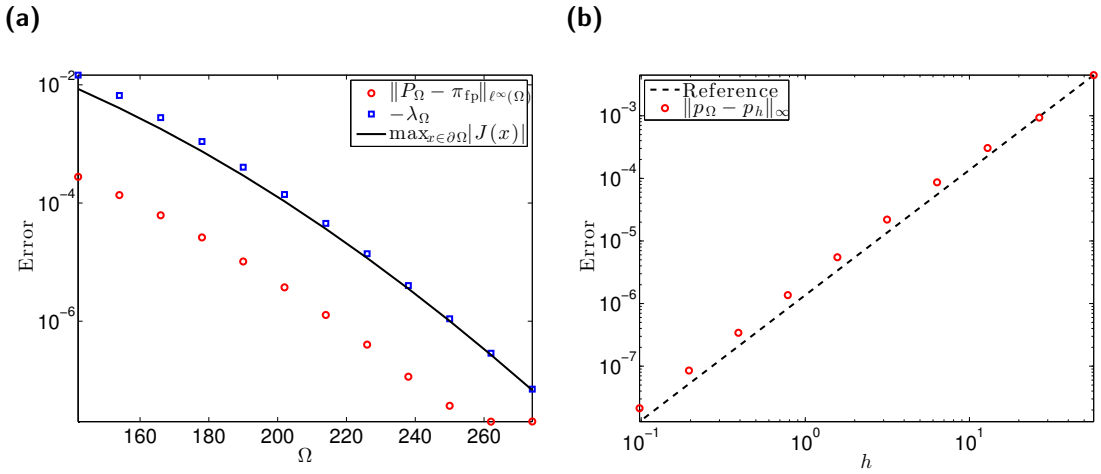

Figure 3.2: *Numerical illustrations of the artificial boundary error and the discretisation error in the CFPE of the birth-death process* (2.29) *with* $N = 1$ *chemical species. The reaction constants are* $k_1 = k_2 = 1$ *and system volume* $V = 500$. **(a)** *The domain size* Ω *dependence of the artificial boundary error of* (3.30) *in approximating* (2.31). *Function* P_Ω *is approximated by a finite difference solution with very small mesh size* $h = 0.1$. **(b)** *Discretisation error of the proposed monotone difference scheme in solving* (2.31). *The exact error is plotted with red circles, and the dashed curve refers to the reference line* Ch^2 *with constant* $C = 1.37 \times 10^{-6}$.

3.1.3 Numerical illustrations of boundary truncation error and the discretisation error

Let us consider the simple birth-death process (2.29). We can write the truncated CFPE (3.2) for the process as

$$\begin{cases} \mathscr{L}_{\mathrm{f}} P_\Omega(x) = \frac{\mathrm{d}}{\mathrm{d}x}\left[D(x)\frac{\mathrm{d}P_\Omega(x)}{\mathrm{d}x}\right] - \frac{\mathrm{d}}{\mathrm{d}x}\left[v(x)P_\Omega(x)\right] = \lambda_\Omega P_\Omega(x), \quad x \in \Omega, \\ P_\Omega(x) = 0, \quad x \in \partial\Omega, \\ \int_{\Omega_{\mathrm{f}}} P_\Omega(x)\mathrm{d}x = 1, \end{cases} \tag{3.30}$$

where λ_Ω is the principal eigenvalue and $\Omega \subset \Omega_{\mathrm{f}}$. For $\Omega = \Omega_{\mathrm{f}}$, we have derived the exact solution π_{f} in (2.33).

We fix the centre of Ω at $x = 500$, and numerically solve (3.30) for different domain sizes. The Ω-dependence of the boundary truncation error is plotted in Figure 3.2a, and the convergence agrees with Theorem 3.2.

In Figure 3.2a, we also plot the maximum boundary flux against domain sizes. The probability flux, or current, $J_{\text{flux}}(x)$ is defined as (Risken, 1989)

$$J_{\text{flux}}(x) = D(x)\frac{\mathrm{d}P_\Omega(x)}{\mathrm{d}x} - v(x)P_\Omega(x).$$

The quantity $J_{\text{flux}}(x)$ is the probability flux, or current (Risken, 1989). A physical explanation of $J_{\text{flux}}(x)$ is the amount of the flow of particles through the point x. From Figure 3.2a, the decay rate of the maximum flux is similar to the decay rate of λ_Ω, both equivalent to the decay rate of $\|P_\Omega - \pi_{\text{f}}\|_{\ell^\infty(\Omega)}$ with a constant multiplication. They may be used as error indicators for the artificial boundary error, but we are not able to provide theoretical proof for such argument at the moment.

The discretisation error is plotted in Figure 3.2b. The underlying matrix eigenvalue problem (3.5) is solved with the MATLAB eigs function. When computing the exact error, vector $\boldsymbol{p}_\Omega$ as a grid evaluation of function P_Ω is approximated by a grid evaluation of π_{f}. This is because the analytical formula for P_Ω in (3.30) is not known for general domain Ω. We choose sufficient large domain size $\Omega = 400$, such that this approximation error, $\|P_\Omega - \pi_{\text{f}}\|_{\ell^\infty(\Omega)}$, is negligible compared to the presented discretisation error. In Figure 3.2b, we could observe that the difference scheme is of second-order accuracy as predicted by Theorem 3.9.

3.2 Separation of dimensions

Solving the CFPE (3.5) using the classic matrix-vector-based approaches can be difficult, because the number of degrees of freedom grows exponentially in dimensions. We will show the problem can be lifted using tensor approaches, and in the current section, we introduce essential knowledge of tensor formats, particularly the tensor train (TT) format and the quantized tensor train (QTT) format.

The class of the tensor methods that lead to linear complexity in the dimensions

are distinctly based on the principle of separation of variables. The techniques, especially the classic canonical and tucker decompositions, have been used as a tool for data processing in chemometrics, psychometrics and in signal processing, see the bibliography in Kolda & Bader (2009). Its application in the numerical solution of PDEs in scientific computing was originated by the use of the rank-structured tensor calculations of multivariate functions and operators in the Hartree-Fock equation (Khoromskij & Khoromskaia, 2009; Khoromskij et al., 2011; Khoromskaia, 2010; Khoromskaia et al., 2013; Khoromskaia & Khoromskij, 2014b). Since then, significant progress has been made in approximating the multidimensional functions and operators in the TT and QTT formats. The TT format is a particular form of the matrix product states (MPS) techniques originally developed in the physics community to address the similar problem (White, 1992; Verstraete et al., 2004; Vidal, 2003). Literature review on the most frequently used tensor formats can be found in (Khoromskij, 2012).

The numerical cost of basic multi-linear algebra operations on formatted tensor, like the TT format, usually scales linearly in dimension N, but could be polynomial in the volume size n. The introduction of the quantization in the TT approximation further enhanced the data compression with a log-volume complexity $\mathcal{O}(N\log(n))$ (Khoromskij, 2009a, 2011). It has been shown that the basic classes of function related tensors, such as exponential, trigonometric and polynomial, can be constructed in the QTT format with constant ranks. The exponential convergence in the QTT ranks has also been proven for a class of analytic function related tensors. And we refer the readers to the details in (Khoromskij & Oseledets, 2010b; Oseledets, 2013; Dolgov et al., 2012).

Here, the main ingredients in the tensor-structured solution of the CFPE include: 1) the FDM discretisation, 2) numerical multi-linear algebra of the low-parametric tensor formats, 3) approximate the multivariate functions and operators in the low-

parametric tensor formats, 4) the quantized tensor approximation, and 5) tensor-structured method for solving the discrete systems of the PDE. Step 1) has been studied in section 3.1, and the numerical details of the tensor representations, particularly the TT format, is introduced in the current section. Steps 4) and 5) will be discussed in sections 3.3 and 3.4.

3.2.1 Notations and conventions

Definition 3.10 (Tensor)**.** *An N-dimensional, or Nth-order, tensor $\boldsymbol{x} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ is an N-way array with entries referenced with N indices denoted as $x_{j_1,\ldots,j_N}$ or $x(j_1,\ldots,j_N)$ for $j_k \in \{1,2,\ldots,n_k\}$ and $k = 1,\ldots,N$.*

A grid evaluation of a function naturally yields a tensor, and the order of the tensor can include spatial, temporal, and parametric variables. A scalar is a zeroth-order tensor, a vector is a first-order tensor, and a matrix is a second-order tensor.

A *vectorisation* of tensor $\boldsymbol{x}$ of N-th order is a process that stack the entries of $\boldsymbol{x}$ into a long column vector $\boldsymbol{y}$, denoted by $\boldsymbol{y} = \mathrm{vec}(\boldsymbol{x})$, with entries given by*

$$y_{j_1+\sum_{k=2}^{N}(j_k-1)\prod_{\ell=1}^{k-1} n_\ell} = x_{j_1,j_2,\ldots,j_N}, \quad j_k = 1,\ldots,n_k, \quad k = 1,\ldots,N.$$

The tensor vectorisation is implicitly used in the vector-based numerical methods for solving PDEs, where the grid evaluation of any function needs to be vectorised to

* The transformation between multi-indexed and single-indexed representation has two difference conventions:

1. The little-endian convention

$$(j_1,j_2,\ldots,j_N) = j_1 + (j_2-1)n_1 + (j_3-1)n_1n_2 + \cdots + (j_N-1)n_1\cdots n_{N-1}.$$

2. The big-endian convention

$$(j_1,j_2,\ldots,j_N) = j_N + (j_{N-1}-1)n_N + (j_{N-2}-1)n_Nn_{N-1} + \cdots + (j_1-1)n_2\cdots n_N.$$

The big-endian notation correspond to the reverse lexicographic order adopted in C language, and the little-endian notation is consistent with Fortran and MATLAB indexing. If not otherwise mentioned, we will always use the little-endian notation in this thesis.

form a linear system, such as Equation (3.5). Reversely, a vector $\boldsymbol{y}$ can be *tensorised* into tensor $\boldsymbol{x}$ using the above index transformation.

The process of reordering the elements of a tensor into a matrix is usually referred as *unfolding* or *flattening*. A mode-k unfolding of tensor $\boldsymbol{x} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ yields a matrix $\boldsymbol{x}_{(n)} \in \mathbb{R}^{n_k \times (n_1 \cdots n_N)}$, in which the kth index j_k of tensor $\boldsymbol{x}$ is arranged as the row index of matrix $\boldsymbol{x}_{(k)}$ and the remaining indices $(i_1, \ldots, i_{k-1}, i_{k+1}, \ldots, i_N)$ as a whole serve as the column index. Similarly, a mode-$(1, \ldots, k)$ unfolding of tensor $\boldsymbol{x}$ generate a matrix, written as $\boldsymbol{x}_{(1,\ldots,k)}$, whose row and column indices are re-arrangements of indices $(j_1, \ldots, j_k)$ and $(j_{k+1}, \ldots, j_N)$ in the reverse lexicographical order, respectively.

Tensors can be multiplied by scalars, vectors, matrices and tensors. Multiplying tensor $\boldsymbol{x} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ by a scalar c is direct, where all entries of $\boldsymbol{x}$ are scaled by c. If $\boldsymbol{x}$ is multiplied by a vector or a matrix, a mode index needs to be specified, and it is known as *mode-k product*, denoted by $\times_k$. The mode-k product of tensor $\boldsymbol{x}$ by vector $\boldsymbol{y} \in \mathbb{R}^{n_k}$ yield tensor $\boldsymbol{z} = \boldsymbol{x} \times_k \boldsymbol{y} \in \mathbb{R}^{n_1 \times \cdots \times n_{k-1} \times n_{k+1} \times \cdots n_N}$, with entries

$$z_{j_1,\ldots,j_{k-1},j_{k+1},\ldots,j_N} = \sum_{j_k=1}^{n_k} x_{j_1,\ldots,j_N} y_{j_k},$$

or equivalently, we can write it in the matrix form as $\boldsymbol{z}_{(k)} = \boldsymbol{x}_{(k)}\boldsymbol{y}$, where $\boldsymbol{z}_{(k)}$ and $\boldsymbol{x}_{(k)}$ are the mode-k unfolding of tensors $\boldsymbol{z}$ and $\boldsymbol{x}$, respectively. Tensor-matrix multiplication is defined in a similar manner. An N-mode tensor $\boldsymbol{x} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ multiplied by a matrix $\boldsymbol{A} \in \mathbb{R}^{m_k \times n_k}$ yields another N-mode tensor $\boldsymbol{y}$ of size $n_1 \times \cdots \times n_{k-1} \times m_k \times n_{k+1} \times \cdots \times n_N$, whose entries are given by

$$y_{j_1,\ldots,j_{k-1},i_k,j_{k+1},\ldots,j_N} = \sum_{j_k=1}^{n_k} x_{j_1,\ldots,j_N} A_{i_k,j_k},$$

whose equivalent matrix formulation is given by $\boldsymbol{y}_{(k)} = \boldsymbol{A}\boldsymbol{x}_{(k)}$.

The *contraction* between two tensors, $\boldsymbol{x} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ and $\boldsymbol{y} \in \mathbb{R}^{m_1 \times \cdots \times m_M}$, where $n_\ell =$

m_k, is performed through the mode-$\binom{k}{\ell}$ product, denoted by $\times_\ell^k$, generating a tensor:

$$\boldsymbol{z} = \boldsymbol{x} \times_\ell^k \boldsymbol{y} \in \mathbb{R}^{n_1 \times \cdots \times n_{\ell-1} \times n_{\ell+1} \times \cdots \times n_N \times m_1 \times m_{k-1} \times m_{k+1} \times \cdots \times m_M},$$

whose entries are defined by

$$z_{i_1,\ldots,i_{\ell-1},i_{\ell+1},\ldots i_N,j_1,\ldots,j_{k-1},j_{k+1},\ldots,j_M} = \sum_{i=1}^{n_\ell} x_{i_1,\ldots,i_{\ell-1},i,i_{\ell+1},\ldots,i_N} y_{j_1,\ldots,j_{k-1},i,j_{k+1},\ldots,j_M},$$

and the common modes have the same size $n_\ell = m_k$. In the case $k = 1$, we usually neglect the super-index k and simply write $\times_\ell^k$ as $\times_\ell$. When the tensor contraction is applied for more than two tensors, the order of the product is backwards, for example, $\boldsymbol{x} \times_{\ell_1}^{k_1} \boldsymbol{y} \times_{\ell_2}^{k_2} \boldsymbol{z} \equiv \boldsymbol{x} \times_{\ell_1}^{k_1} \left(\boldsymbol{y} \times_{\ell_2}^{k_2} \boldsymbol{z} \right)$.

The *tensor product*, denoted by $\otimes$, of $\boldsymbol{x} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ and $\boldsymbol{y} \in \mathbb{R}^{m_1 \times \cdots \times m_M}$ is the tensor $\boldsymbol{z} = \boldsymbol{x} \otimes \boldsymbol{y}$ of size $n_1 \times \cdots \times n_N \times m_1 \times \cdots m_M$, and the entries of $\boldsymbol{z}$ is given by

$$z_{i_1,\ldots,i_N,j_1,\ldots,j_M} = x_{i_1,\ldots,i_N} y_{j_1,\ldots,j_M}.$$

In the context where $\boldsymbol{x}$ and $\boldsymbol{y}$ are first-order tensors, or vectors, the tensor product is referred as *outer product*. And if $\boldsymbol{x}$ and $\boldsymbol{y}$ are second-order tensors, or matrices, the tensor product is similar to the matrix *Kronecker product* with the same notion $\otimes$. The difference is that, if $\boldsymbol{A}$ and $\boldsymbol{B}$ be the matrices of sizes $m \times n$ and $p \times q$ respectively, the Kronecker product is defined as

$$\boldsymbol{C} = \boldsymbol{A} \otimes \boldsymbol{B} = \begin{bmatrix} a_{11}\boldsymbol{B} & \cdots & a_{1,n}\boldsymbol{B} \\ \vdots & \ddots & \vdots \\ a_{n1}\boldsymbol{B} & \cdots & a_{b,n}\boldsymbol{B} \end{bmatrix}.$$

where the product $\boldsymbol{C}$ is a matrix of size $mp \times nq$, whereas the tensor product yields a four-dimensional tensor. A detailed account of properties of tensor product is given

by Kolda & Bader (2009). Some elementary properties include

$$(\boldsymbol{x}+\boldsymbol{y})\otimes\boldsymbol{z}=\boldsymbol{x}\otimes\boldsymbol{z}+\boldsymbol{y}\otimes\boldsymbol{z},$$

$$(\boldsymbol{x}\otimes\boldsymbol{y})(\boldsymbol{z}\otimes\boldsymbol{w})=\boldsymbol{x}\boldsymbol{z}\otimes\boldsymbol{y}\boldsymbol{w},$$

which suggests the summation and the inner product are commutative. Here, the *inner (scalar) product* is defined similarly for vectors, i.e., the inner product of tensors $\boldsymbol{x},\boldsymbol{y}\in\mathbb{R}^{n_1\times\cdots\times n_N}$ yields a constant $z=\boldsymbol{x}\boldsymbol{y}\in\mathbb{R}$ defined by

$$z=\sum_{i_1}^{n_1}\sum_{i_2}^{n_2}\cdots\sum_{i_N}^{n_N}x_{i_1,\ldots,i_N}y_{i_1,\ldots,i_N}.$$

For brevity, we do not indicate explicitly the sizes of the tensors involved; we assume throughout that the sizes of the involved tensors are compatible with the indicated operations.

An N-dimensional tensor $\boldsymbol{x}\in\mathbb{R}^{n_1\times\cdots\times n_N}$ is said to be a rank-1 tensor, if it can be written as a tensor product N factor first-order tensors, $\boldsymbol{x}_i\in\mathbb{R}^{n_i}$, $i=1,\ldots,N$, of the form

$$\boldsymbol{x}=\boldsymbol{x}_1\otimes\boldsymbol{x}_2\otimes\cdots\otimes\boldsymbol{x}_N, \tag{3.31}$$

or equivalently, in a compact form as $\boldsymbol{x}=\bigotimes_{i=1}^{N}\boldsymbol{x}_i$.

The previously introduced tensors can be regarded as a high-dimensional analogue of vectors in the traditional numerical methods. To be consistent, we also introduce a concept of tensor matrices as a multi-dimensional counterpart of matrices.

Definition 3.11. *An N-dimensional, or N-th order, tensor matrix $\boldsymbol{A}\in\mathbb{R}^{(n_1\times\cdots\times n_N)\times(m_1\times\cdots\times m_N)}$ is a $2N$-dimensional tensor, and the entries are referenced with N row indices $(i_1,\ldots,i_N)$ and N column indices $(j_1,\ldots,j_N)$, denoted as $A_{i_1,\ldots,i_N;j_1,\ldots,j_N}$ or $A(i_1,\ldots,i_N;j_1,\ldots,j_N)$ for $i_k\in\{1,2,\ldots,n_k\}$ and $j_k\in\{1,2,\ldots,m_k\}$, $k=1,2,\ldots,N$.*

The process of reshaping a tensor matrix $\boldsymbol{A} \in \mathbb{R}^{(n_1 \times \cdots \times n_N) \times (m_1 \times \cdots \times m_N)}$ into a matrix (second-order tensor) $\boldsymbol{B} \in \mathbb{R}^{(n_1 \cdots n_N) \times (m_1 \cdots m_N)}$ is referred as *matricisation*, denoted by $\boldsymbol{B} = \text{mat}(\boldsymbol{A})$. The entries are ordered as

$$B_{i_1 + \sum_{k=2}^{N} (i_k - 1) \prod_{\ell=1}^{k-1} n_\ell;\ j_1 + \sum_{k=2}^{N} (j_k - 1) \prod_{\ell=1}^{k-1} m_\ell} = A_{i_1, \ldots, i_N; j_1, \ldots, j_N}. \tag{3.32}$$

The product of two tensor matrices are equivalent to the product of their matricisation. That is, given tensor matrices $\boldsymbol{A} \in \mathbb{R}^{(n_1 \times \cdots \times n_N) \times (m_1 \times \cdots \times m_N)}$ and $\boldsymbol{B} \in \mathbb{R}^{(m_1 \times \cdots \times m_N) \times (k_1 \times \cdots \times k_N)}$, the product is $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B} \in \mathbb{R}^{(n_1 \times \cdots \times n_N) \times (k_1 \times \cdots \times k_N)}$, and entries are specified as

$$C_{i_1, \ldots, i_N; j_1, \ldots, j_N} = \sum_{\ell_1 = 1}^{m_1} \cdots \sum_{\ell_N = 1}^{m_N} A_{i_1, \ldots, i_N; \ell_1, \ldots, \ell_N} B_{\ell_1, \ldots, \ell_N; j_1, \ldots, j_N}.$$

which we can also write in the matricised form as $\text{mat}(\boldsymbol{C}) = \text{mat}(\boldsymbol{A})\text{mat}(\boldsymbol{B})$. Also, a tensor matrix $\boldsymbol{A} \in \mathbb{R}^{(n_1 \times \cdots \times n_N) \times (m_1 \times \cdots \times m_N)}$ can be multiplied by a tensor $\boldsymbol{x} \in \mathbb{R}^{m_1 \times \cdots m_N}$ which yields another tensor $\boldsymbol{y} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$, with entries

$$y_{i_1, \ldots, i_N} = \sum_{j_1}^{m_1} \cdots \sum_{j_N}^{m_N} A_{i_1, \ldots, i_N; j_1, \ldots, j_N} x_{j_1, \ldots, j_N}.$$

This is equivalent to the matrix-vector multiplication through $\text{vec}(\boldsymbol{y}) = \text{mat}(\boldsymbol{A})\text{vec}(\boldsymbol{x})$.

In a similar way to the rank-1 tensor, we can define the rank-1 tensor matrix $\boldsymbol{A}$ of size $(n_1 \times \cdots \times n_N) \times (m_1 \times \cdots \times m_N)$ as a tensor product of N factor first-order tensor matrices, $\boldsymbol{A}_i \in \mathbb{R}^{n_i \times m_i}$, $i = 1, 2, \ldots, N$, of the form

$$\boldsymbol{A} = \boldsymbol{A}_1 \otimes \boldsymbol{A}_2 \otimes \cdots \otimes \boldsymbol{A}_N,$$

or equivalently,

$$A(i_1, \ldots, i_N; j_1, \ldots, j_N) = A_1(i_1, j_1) A_2(i_2, j_2) \cdots A_N(i_N, j_N).$$

Here, the tensor product works as the matrix Kronecker product. Also note that the tensor matrices are equivalent to the tensors with permuted and grouped indices, and thus, if a tensor decomposition technique applies to the tensors, it works on tensor matrices with permuted indices as well.

3.2.2 Tensor decompositions

3.2.2.1 Low-rank approximation – a two-dimensional example

The key concept in the tensor decomposition is to separate the dimensions, and the most straightforward approach is to separate using the tensor product as in (3.31). The rank-one representation can be generalised to a summation of several rank-one terms, and the representation can be derived out from the singular value decomposition (SVD).

To illustrate the basic idea, let us consider a second-order tensor, $\boldsymbol{x} \in \mathbb{R}^{n \times m}$, $n \leq m$. We wish to express tensor $\boldsymbol{x}$ as a sum of rank-one product terms plus an approximation error, i.e.,

$$\boldsymbol{x} = \sum_{r=1}^{R} c_r \boldsymbol{u}_r \otimes \boldsymbol{v}_r + \boldsymbol{e} = \sum_{r=1}^{R} c_r \boldsymbol{u}_r \boldsymbol{v}_r^T + \boldsymbol{e}, \tag{3.33}$$

where c_r, $r = 1,\ldots,R$, are constant coefficients, $\boldsymbol{u}_r \in \mathbb{R}^n$ and $\boldsymbol{v}_r \in \mathbb{R}^m$, $r = 1,2,\ldots,N$, are factor vectors, and tensor $\boldsymbol{e} \in \mathbb{R}^{n \times m}$ represents the approximation residue. It is easy to see that the original full tensor $\boldsymbol{x}$ contains nm entries, while the storage cost of the factor tensors, $\boldsymbol{u}_i$ and $\boldsymbol{v}_i$, reduces to $(m+n)R$. We say that (3.33) is the *optimal decomposition* if it has the minimal ε-rank R that keeps the error $\|\boldsymbol{e}\|_2 \leq \varepsilon$, for some $\varepsilon > 0$.

Proposition 3.12. *Consider the SVD of matrix* $\boldsymbol{x} \in \mathbb{R}^{n\times m}$, $n \le m$ *as*

$$\boldsymbol{x} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T = \sum_{r=1}^{n} \sigma_r \boldsymbol{u}_r \otimes \boldsymbol{v}_r = \sum_{r=1}^{n} \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^T, \tag{3.34}$$

for left singular matrix $\boldsymbol{U} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \in \mathbb{R}^{n\times n}$, *right singular matrix* $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n) \in \mathbb{R}^{n\times m}$, *and diagonal matrix* $\boldsymbol{\Lambda} = \mathrm{diag}(\sigma_i) \in \mathbb{R}^{n\times n}$. *The order of the singular values follows* $\sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_n$. *The truncated SVD after* $R < n$ *terms,*

$$\boldsymbol{y} = \sum_{r=1}^{R} \sigma_r \boldsymbol{u}_r \otimes \boldsymbol{v}_r = \sum_{r=1}^{R} \sigma_r \boldsymbol{u}_r \boldsymbol{v}_r^T,$$

yields an optimal ε*-rank* R *approximation with error tolerance* $\varepsilon = \|\boldsymbol{x} - \boldsymbol{y}\|_2$.

The above result is usually referred to as the Eckart-Young-Mirsky theorem (Eckart & Young, 1936). One higher-dimensional generalisation of the theorem is the higher order singular value decomposition (HOSVD) (De Lathauwer et al., 2000a), which computes singular value decompositions independently for each coordinate. This gives reliability and a quasi-optimal accuracy/rank ratio. Khoromskij & Khoromskaia (2009) extended this technique to the so-called Reduced HOSVD (RHOSVD), which finds the best (non-linear) Tucker approximation from large-rank canonical inputs via solving a dual maximization problem. The RHOSVD was justified with error estimate, and yields linear complexity in the grid size and the input rank.

Another generalisation of Proposition 3.12 is the MPS or TT-style tensor format, brought by a sequence of the (truncated) SVD. It was first invented by White (1992, 1993), and has been further developed for more than 20 years in quantum chemistry and in quantum information theory. The alternating direction concept for the linear structure of MPS/TT is essentially more powerful than the alternating least square for the canonical and Tucker format. Out of the alternating direction concept, the DMRG method has been extensively used to simulate the statics and dynamics of one-dimensional strongly correlated quantum lattice systems (Schollwöck, 2011a,b),

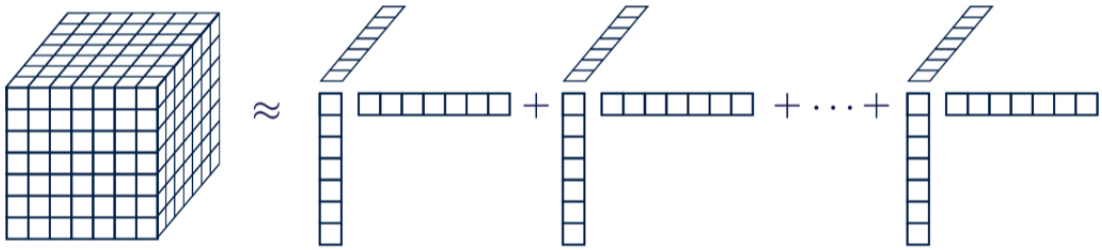

Figure 3.3: *A graphic illustration of the canonical tensor decomposition.*

which was later adopted in the numerical community as well (Holtz et al., 2012). It has many impressive modifications and improvements, such as the time evolution of MPS deveoped by Vidal (2003).

In what follows, we discuss the basic algebraic operations of the canonical, Tucker and TT formats.

3.2.2.2 The Canonical Polyadic and Tucker decomposition

An obvious generalisation of the matrix decomposition (3.33) is the so-called canonical Polyadic decomposition first introduced by Hitchcock (1927, 1928).

Definition 3.13. *An N-th order tensor $\boldsymbol{x} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ is said to be represented in the canonical format if it is expressed (or approximated) as a sum of rank-one tensors, i.e.,*

$$\boldsymbol{x} \approx \sum_{r=1}^{R} \boldsymbol{x}_1^{(r)} \otimes \boldsymbol{x}_2^{(r)} \otimes \cdots \otimes \boldsymbol{x}_N^{(r)}, \quad \boldsymbol{x}_i^r \in \mathbb{R}^{n_i}, \quad i = 1,2,\ldots,N, \quad r = 1,2,\ldots,R. \tag{3.35}$$

The number of summands, R, is called the canonical rank, and $\boldsymbol{x}_i^{(r)}$ are canonical factors. Similarly, an N-th order tensor matrix, $\boldsymbol{A} \in \mathbb{R}^{(n_1 \times \cdots \times n_N) \times (m_1 \times \cdots \times m_N)}$, is represented in the canonical format if it holds

$$\boldsymbol{A} = \sum_{r=1}^{R} \boldsymbol{A}_1^{(r)} \otimes \boldsymbol{A}_2^{(r)} \otimes \cdots \otimes \boldsymbol{A}_N^{(r)}, \quad \boldsymbol{A}_i^{(r)} \in \mathbb{R}^{n_i \times m_i}, \quad i = 1,2,\ldots,N, \quad r = 1,2,\ldots,R. \tag{3.36}$$

An illustration of the canonical representation of a three-dimensional tensor is

given in Figure 3.3. The above representation admits linear complexities in high-dimensional algebraic operations. It is possible to prove the existence of low-rank approximations for certain classes of tensors, especially if the tensors arise from the grid evaluation of some special types of functions, such as the product of Gaussians. And for tensor matrices, low-rank representations may be analytical derived for the discretisations of some differential operators, as we will derive for the Fokker-Planck operator in Section § 3.3. Given the canonical decomposition, we show next how basic arithmetic operations can be performed in such a representation, and for further details about the tensor arithmetic, we refer the readers to the discussion in Khoromskaia (2010).

Proposition 3.14. *Let tensors* $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ *admit canonical representations* (3.35) *of ranks* $R_{\boldsymbol{x}}$ *and* $R_{\boldsymbol{y}}$*, respectively. Let tensor matrix* $\boldsymbol{A} \in \mathbb{R}^{(m_1 \times \cdots \times m_N) \times (n_1 \times \cdots \times n_N)}$ *admit canonical representation* (3.36) *of rank* $R_{\boldsymbol{A}}$*. Define* $n \geq n_i, m_i$*,* $i = 1, 2, \ldots, N$*, the following arguments hold:*

(*i*) *Storage Cost: the number of entries of full tensor* $\boldsymbol{x}$ *for storage is* $\mathcal{O}(n^N)$*, and is reduced to* $\mathcal{O}(nNR_{\boldsymbol{x}})$ *in the canonical format; the number of entries of full tensor matrix* $\boldsymbol{A}$ *for storage is* $\mathcal{O}(n^{2N})$*, and is reduced to* $\mathcal{O}(n^2 N R_{\boldsymbol{A}})$ *in the canonical format.*

(*ii*) *Vector addition: the sum of tensors* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *in the canonical form is performed through merging and reordering the rank-one terms; there is no appreciable computational cost, but the canonical rank of the sum equals the sum of the ranks of the summands, i.e.,* $R_{\boldsymbol{x}+\boldsymbol{y}} = R_{\boldsymbol{x}} + R_{\boldsymbol{y}}$*.*

(*iii*) *Multiplied by a scalar: multiplying a scalar* $s \in \mathbb{R}$ *to tensor* $\boldsymbol{x}$ *in the canonical format is performed by multiplying* s *to the factor vectors,* $\boldsymbol{x}_i^{(r)}$*,* $r = 1, 2, \ldots, R$*,* $\forall i \in \{1, 2, \ldots, N\}$*; the computational cost is* $\mathcal{O}(nR)$*.*

(*iv*) *Vector inner product: the inner product of tensors* $\boldsymbol{x}$ *and* $\boldsymbol{y}$ *in the canonical*

format is performed by computing the inner products of the factor tensors as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{r_1=1}^{R_{\boldsymbol{x}}} \sum_{r_2=1}^{R_{\boldsymbol{y}}} \prod_{i=1}^{N} \left\langle \boldsymbol{x}_i^{(r_1)}, \boldsymbol{y}_i^{(r_2)} \right\rangle,$$

which requires $\mathcal{O}(nNR_{\boldsymbol{x}}R_{\boldsymbol{y}})$ operations.

(*v*) *Matrix-vector product: multiplying $\boldsymbol{A}$ by $\boldsymbol{x}$ from the right is performed via*

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} = \sum_{r_1=1}^{R_{\boldsymbol{x}}} \sum_{r_1=1}^{R_{\boldsymbol{y}}} \bigotimes_{i=1}^{N} \boldsymbol{A}_i^{(r_1)} \boldsymbol{x}_i^{(r_2)}.$$

The computational cost is $\mathcal{O}(n^2NR_{\boldsymbol{A}}R_{\boldsymbol{x}})$, and the canonical rank of the tensor $\boldsymbol{y}$ is $R_{\boldsymbol{y}} = R_{\boldsymbol{A}}R_{\boldsymbol{x}}$.

(*v*) *Mode-k product: the mode-k product of $\boldsymbol{x}$ by a matrix $\boldsymbol{B} \in \mathbb{R}^{n_k \times m_k}$ is performed via multiplying $\boldsymbol{B}$ to the k-th canonical factor, i.e.,*

$$\boldsymbol{y} = \boldsymbol{x} \times_k \boldsymbol{B} = \sum_{r=1}^{R_{\boldsymbol{x}}} \boldsymbol{x}_1^r \otimes \cdots \otimes \boldsymbol{x}_{k-1}^r \otimes \left(\boldsymbol{B}^T \boldsymbol{x}_k^r \right) \otimes \boldsymbol{x}_{k+1}^r \otimes \cdots \otimes x_N^r,$$

which requires $\mathcal{O}(n^2R_{\boldsymbol{x}})$ operations.

In Proposition 3.14, we see a common pattern that the tensor decomposition allows us to combine one-dimensional operations to achieve higher-dimensional operations. The result is another canonical tensor with a substantial larger separation rank. Thus, at each operation, it is essential to have some data compression procedure to reduce rank to avoid its untenable growth.

However, there is lack of stable algorithms to reduce the separation rank of the canonical tensors. Although the canonical decomposition of a tensor itself is usually unique (Sidiropoulos & Bro, 2000), the problem of determining the optimal rank has been proven to be NP-hard (Håstad, 1990). The existing techniques for truncating the canonical rank are mainly based on the Greedy-type optimisation algorithms (Espig & Hackbusch, 2012; Beylkin & Mohlenkamp, 2005), but they may converge

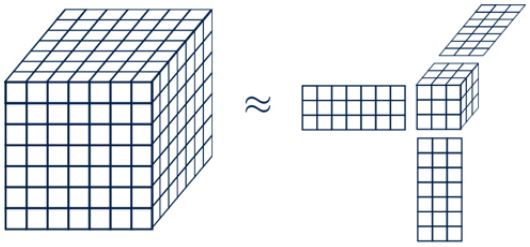

Figure 3.4: *A graphic illustration of the Tucker decomposition.*

slowly or get stuck into the local minimums.

Therefore, a desirable generalisation of the two-dimensional decomposition in Proposition 3.12 is to use SVD-based algorithms to reduce the separation rank. The first tensor formats of this type is the so-called Tucker tensor decomposition (Tucker, 1966).

Definition 3.15. *An N-th order tensor $\boldsymbol{x} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ is said to be represented in the Tucker format if it is expressed as*

$$\boldsymbol{x} = \boldsymbol{c} \times_1 \boldsymbol{B}_1 \times_2 \boldsymbol{B}_2 \times_3 \cdots \times_N \boldsymbol{B}_N, \tag{3.37}$$

where $\boldsymbol{c} \in \mathbb{R}^{R_1 \times \cdots \times R_N}$ is the core tensor, and $\boldsymbol{B}_i = \left(\boldsymbol{b}_1^{(i)}, \boldsymbol{b}_2^{(i)}, \ldots, \boldsymbol{b}_N^{(i)}\right)^T \in \mathbb{R}^{n_i \times R_i}$ are mode-i factor matrices, $i = 1, 2, \ldots, N$. The N-tuple $(R_1, \ldots, R_N)$ is called the Tucker rank of $\boldsymbol{x}$.

A graphic illustration of the Tucker format is given in Figure 3.4. If the core tensor $\boldsymbol{c}$ is a diagonal tensor, where $c_{i_1, i_2, \ldots, i_N} \neq 0$ only along the main diagonal, the Tucker decomposition (3.37) is reduced to the canonical tensor (3.35). To see how the matrix SVD can play a role in the Tucker format, we write (3.37) in the matricised form as

$$\boldsymbol{x}_{(i)} = \boldsymbol{B}_i \boldsymbol{c}_{(i)} (\boldsymbol{B}_1 \otimes \cdots \otimes \boldsymbol{B}_{i-1} \otimes \boldsymbol{B}_{i+1} \otimes \cdots \otimes \boldsymbol{B}_N)^T,$$

where the subscript, $\cdot_{(i)}$, denotes the mode$-i$ tensor unfolding introduced in Section § 3.2.1. The above formula is of similar form as the matrix SVD in (3.34), thus we may

use this techniques to compute and compress the Tucker tensors. The procedure is formally called the higher-order singular value decomposition (HOSVD), whose main building blocks are (De Lathauwer et al., 2000a):

1. Input a full N-th order tensor $\boldsymbol{x} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$, and its mode-$k$ unfoldings, $\boldsymbol{x}_{(k)}$, $k = 1, 2, \ldots, N$.
2. For each $k = 1, 2, \ldots, N$, compute the matrix SVD of the mode-k unfolding, $\boldsymbol{x}_{(k)} = \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^T$, as in (3.34). A truncation after $R_k < n_k$ terms may be performed, and then store the truncated left singular matrix as $\boldsymbol{B}_k = \left(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{R_k}\right)$.
3. Compute the core tensor as $\boldsymbol{c} = \boldsymbol{x} \times_1 \boldsymbol{B}_1^T \times_2 \boldsymbol{B}_2^T \times_3 \cdots \times_N \boldsymbol{B}_N^T$.

The resulting factor matrices as well as the core tensor are all orthogonal. If the Tucker decomposition is exact, the procedure eliminates the redundancies in the principle components in the factor matrices, according to the corresponding singular value. If further approximation is made by truncating less significant components with small singular values, the above HOSVD gives a quasi-optimal approximation as given in the following proposition.

Proposition 3.16. *Consider tensor $x \in \mathbb{R}^{n_1 \times \cdots \times n_N}$, and its approximation $\boldsymbol{y} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ in the Tucker format* (3.37) *as a result of the truncated HOSVD, with the Tucker rank-$(R_1, \ldots, R_N)$. Then $\boldsymbol{y}$ is a quasi-optimal Tucker approximation of $\boldsymbol{x}$, with*

$$\|\boldsymbol{x} - \boldsymbol{y}\|_F \leq \sqrt{N}\varepsilon, \quad \text{for} \quad \varepsilon = \min_{\boldsymbol{z}} \|\boldsymbol{x} - \boldsymbol{z}\|_F,$$

where $\boldsymbol{z}$ is another Tucker tensor of the same rank, and $\|\cdot\|_F$ denote the Frobenius norm defined as $\|\boldsymbol{x}\|_F = \|\mathrm{vec}(\boldsymbol{x})\|_2$.

See De Lathauwer et al. (2000a,b) for the proof. Note Tucker decompositions are not unique, and the previously described HOSVD algorithm, although widely known, yet yields one of many structure types of the Tucker format, and the computational is still $\mathcal{O}(n^{N+1})$. Alternative algorithms have been proposed, and we refer the reader

to the review Kolda & Bader (2009) for details.

Because of the "connectivity" constraint, the exponential scaling in dimensions is not fully eliminated in the Tucker format. Here, the "connectivity" constraint refers to the fact that the Tucker format is build upon a core tensor which connects the factor matrices in each dimension, and thus there is no way to reduce the dimensionality of the core tensor. Despite the fact that the size of the core tensor may be significantly smaller compared to the original full tensor, such separated representation can still become strained as the number of dimensions becomes larger.

3.2.2.3 Tensor train decomposition

A sequential recurrent application of the Tucker decomposition would result in a hierarchical tensor representation in a tree structure (Hackbusch & Kühn, 2009). The resulting representation still achieves linear scaling in dimensions as the canonical format, while admits SVD-based rank truncation procedure like the Tucker format. The tensor train (TT) format, first proposed by Oseledets & Tyrtyshnikov (2009); Oseledets (2011), is one of the simplest form of the this type.

Definition 3.17. *A tensor* $\boldsymbol{x} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ *is said to be represented in the tensor train (TT) format, or matrix product states (MPS), if it is expressed as*

$$x(i_1, \ldots, i_N) = \sum_{r_1=1}^{R_1} \cdots \sum_{r_{N-1}}^{R_{N-1}} x_1(i_1, r_1) x_2(r_1, i_2, r_2) \cdots x_N(r_{N-1}, i_N), \tag{3.38}$$

where $\boldsymbol{x}_i \in \mathbb{R}^{R_{i-1} \times n_i \times R_i}$ *are called TT cores (or blocks), the* $(N-1)$*-tuple indices* $(R_1, \ldots, R_{N-1})$ *are called TT rank, and* $R_0 = R_N = 1$.

The name as matrix product states has been widely known in quantum physics 20 years before it was discovered in the numerical linear algebra (Affleck et al., 1987; Fannes et al., 1992a). If $R_1 = \cdots = R_{N-1} = 1$, the TT representation (3.38) coincides

with the canonical representation (3.35) as a rank-one tensor (3.31). In fact, the rank-one tensor is invariant under all tensor representations.

Tensor matrices acting on tensors can also be represented as a tensor train, which is usually referred as TT matrices defined as below.

Definition 3.18. *A tensor matrix* $\boldsymbol{A} \in \mathbb{R}^{(n_1 \times \cdots \times n_N) \times (m_1 \times \cdots \times m_N)}$ *is said to be represented in the TT matrix format, or matrix product operator (MPO), if it is expressed as*

$$A(i_1,\ldots,i_N;j_1,\ldots,j_N) = \sum_{r_1=1}^{R_1} \cdots \sum_{r_{N-1}=1}^{R_{N-1}} A_1(i_1,j_1,r_1)A_2(r_1,i_2,j_2,r_2)\cdots \tag{3.39}$$
$$\cdots A_{N-1}(r_{N-2},i_{N-1},j_{N-1},r_{N-1})A_N(r_{N-1},i_N,j_N),$$

where $\boldsymbol{A}_i \in \mathbb{R}^{R_{i-1} \times n_i \times m_i \times R_i}$ *are called TT matrix cores (or blocks), and the* $(N-1)$*-tuple indices* $(R_1,\ldots,R_{N-1})$ *are called TT rank, and* $R_0 = R_N = 1$.

In (3.39), each core matrix $\boldsymbol{A}_d$ is in fact a four-dimensional array, and its entry, written as $A_d(r_{d-1},i_d,j_k,r_d)$, is specified by two *rank indices* r_{d-1} and r_d ranging from 1 to R_{d-1} and from 1 to R_d and two *mode indices* ranging from 1 to m_d and from 1 to n_d. One can reshape $\boldsymbol{A}_d$ into a two-level matrix form, and represent $\boldsymbol{A}_d$ by the block matrices $\boldsymbol{A}_d^{(\alpha,\beta)} \in \mathbb{R}^{m_d \times n_d}$, $1 \le \alpha \le R_{d-1}$ and $1 \le \beta \le R_d$, through the form

$$\boldsymbol{A}_d = \begin{bmatrix} \boldsymbol{A}_d^{(1,1)} & \cdots & \boldsymbol{A}_d^{(1,R_d)} \\ \vdots & \ddots & \vdots \\ \boldsymbol{A}_d^{(R_{d-1},1)} & \cdots & \boldsymbol{A}_d^{(R_{d-1},R_d)} \end{bmatrix},$$

and element-wise we have $A_d(r_{d-1},i_d,j_d,r_d) = A^{(r_{d-1},r_d)}(i_d,j_d)$. Then, in the TT format, the adjacent TT cores can be regarded as being connected by the strong Kronecker product (or the rank core product), $\rtimes\!\ltimes$, defined as below (Kazeev & Khoromskij, 2012).

Definition 3.19 (rank product of cores (Kazeev et al., 2013)). *Consider the TT matrix cores* $\boldsymbol{A}_d$ *and* $\boldsymbol{A}_{d+1}$ *composed of blocks* $\boldsymbol{A}_d^{(r_{d-1},r_d)}$ *and* $\boldsymbol{A}_{d+1}^{(r_d,r_{d+1})}$, $1 \le r_\ell \le R_\ell$, $d-1 \le \ell \le d+1$, $1 \le d \le N-1$. *The rank product* $\boldsymbol{A}_d \rtimes\!\ltimes \boldsymbol{A}_{d+1}$ *yields another TT core* $\boldsymbol{A}_{d,d+1}$ *composed of block matrices*

$$\boldsymbol{A}_{d,d+1}^{(r_{d-1},r_{d+1})} = \sum_{r_d=1}^{R_d} \boldsymbol{A}_d^{(r_{d-1},r_d)} \otimes \boldsymbol{A}_{d+1}^{(r_d,r_{d+1})},$$

with mode size $m_d m_{d+1} \times n_d n_{d+1}$.

The rank product $\rtimes\!\ltimes$ works in a similar way as the standard matrix product, and the multiplications between the entries (block matrices) are by means of the Kronecker products. From Definition 3.19, we can recast the TT matrix representation in (3.39) as

$$\boldsymbol{A} = \boldsymbol{A}_1 \rtimes\!\ltimes \boldsymbol{A}_2 \rtimes\!\ltimes \cdots \rtimes\!\ltimes \boldsymbol{A}_N.$$

Also, it is easy to see that the rank product $\rtimes\!\ltimes$ applies to link the core tensor in the TT tensor (vectors) by letting $m_d = 1$, $1 \le d \le N$ in Definition 3.19, and the TT vector representation in (3.38) can be written in the similar form as

$$\boldsymbol{x} = \boldsymbol{x}_1 \rtimes\!\ltimes \boldsymbol{x}_2 \rtimes\!\ltimes \cdots \rtimes\!\ltimes \boldsymbol{x}_N.$$

Next we introduce the mode product of cores (Kazeev et al., 2013), which gives convenience in defining the algebraic operations in TT format as we shall see later.

Definition 3.20 (mode product of cores (Kazeev et al., 2013)). *Consider the TT matrix cores* $\boldsymbol{A}_d \in \mathbb{R}^{R_{d-1}^{\boldsymbol{A}} \times m_d \times s_d \times R_d^{\boldsymbol{A}}}$ *and* $\boldsymbol{B}_d \in \mathbb{R}^{R_{d-1}^{\boldsymbol{B}} \times s_d \times n_d \times R_d^{\boldsymbol{B}}}$ *composed of block matrices* $\boldsymbol{A}_d^{\left(i_{d-1}^{\boldsymbol{A}}, i_d^{\boldsymbol{A}}\right)}$ *and* $\boldsymbol{B}_d^{\left(j_{d-1}^{\boldsymbol{B}}, j_d^{\boldsymbol{B}}\right)}$, $1 \le i_d^{\boldsymbol{A}} \le R_d^{\boldsymbol{A}}$, $1 \le j_d^{\boldsymbol{A}} \le R_d^{\boldsymbol{B}}$, $1 \le d \le N$, *respectively. Then the mode product, denoted by* $\bullet$, *of* $\boldsymbol{A}_d$ *and* $\boldsymbol{B}_d$ *yields* $\boldsymbol{C}_d = \boldsymbol{A}_d \bullet \boldsymbol{B}_d \in \mathbb{R}^{\left(R_{d-1}^{\boldsymbol{A}} R_{d-1}^{\boldsymbol{B}}\right) \times m_d \times n_d \times \left(R_d^{\boldsymbol{A}} R_d^{\boldsymbol{B}}\right)}$,

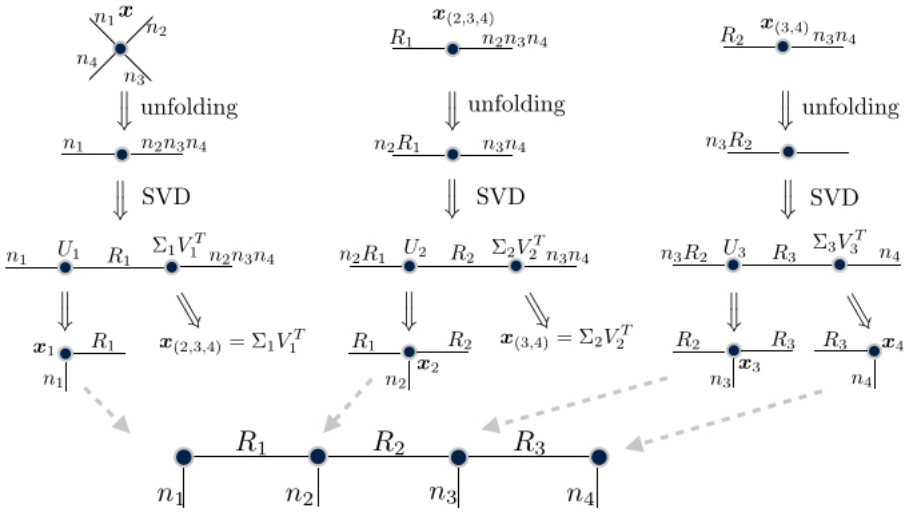

Figure 3.5: *A graphic illustration of the tensor train decomposition. The network representation of muti-dimensional data is used, where the node represents the tensor data and the number of edges linked to it refers to the number of dimensions. The size of each dimension is marked by the variable next to it.*

composed of block matrices

$$\boldsymbol{C}_d^{\left(i_{d-1}^{\boldsymbol{A}} i_{d-1}^{\boldsymbol{B}}, i_d^{\boldsymbol{A}} i_d^{\boldsymbol{B}}\right)} = \boldsymbol{A}_d^{\left(i_{d-1}^{\boldsymbol{A}}, i_d^{\boldsymbol{A}}\right)} \boldsymbol{B}_d^{\left(i_{d-1}^{\boldsymbol{B}}, i_d^{\boldsymbol{B}}\right)},$$

with $1 \le i_d^{\boldsymbol{A}} \le R_d^{\boldsymbol{A}}$, $1 \le j_d^{\boldsymbol{A}} \le R_d^{\boldsymbol{B}}$, $1 \le d \le N$.

The TT representation may be computed via sequential applications of matrix SVD/QR decomposition to unfolded matrices. A typical procedure of this type is called *TT-SVD* (Oseledets, 2011), and can be described as below:

1. Input a full N-th order tensor $\boldsymbol{x} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$, and define a temporary tensor $\boldsymbol{y} = \boldsymbol{x}$.
2. For $k = 1, \ldots, N-1$ do
   2a. Compute the SVD of the mode-1 unfolding matrix $\boldsymbol{y}_{(k)} = \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^T$ and truncate it after R_k terms;
   2b. Store the k-th TT core as $\boldsymbol{x}_k = \boldsymbol{Q}_k$;
   2c. Let $\boldsymbol{y} = \boldsymbol{\Sigma}_k \boldsymbol{V}_k^T$ and reshape $\boldsymbol{y}$ to size $(R_k n_{k+1}) \times n_{k+2} \times \cdots \times n_N$.

3. Store the N-th TT core as $\boldsymbol{x}_N = \boldsymbol{\Sigma}_{N-1}\boldsymbol{R}_{N-1}$.

Each truncated SVD splits one core tensors, and can be either accurate or approximate. A graphic representation of the TT decomposition is given in Figure 3.5. Similar to the Tucker decomposition, in the approximate case, the error generated by the TT-SVD algorithm is given in the following proposition.

Proposition 3.21. *Consider tensor* $\boldsymbol{x} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$*, and its approximation* $\boldsymbol{y} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ *in the TT format* (3.38) *as a result of the TT-SVD algorithm, with the TT-rank-*$(R_1, \ldots, R_{N-1})$*. Let* ε_k*,* $k = 1, 2, \ldots, N-1$*, be the Frobenius norm of the residue in the* k*-th truncated SVD. Then we have*

$$\|\boldsymbol{x} - \boldsymbol{y}\|_F \leq \sqrt{\sum_{k=1}^{N-1} \varepsilon_k^2}.$$

See Oseledets & Tyrtyshnikov (2010) for a proof. The practical implication of Proposition 3.21 is that, if the unfolded matrices are truncated with absolute error ε, then the error in the TT approximation will be $\sqrt{N-1}\varepsilon$. Reversely, in order to achieve an prescribed overall accuracy ε, the threshold for the truncated SVD in the TT-SVD algorithm can be set to $\varepsilon/\sqrt{N-1}$. There also exist other approaches to decompose a tensor into TT format that based on optimisation of some suitable cost function, including the alternating least squares (ALS) type of algorithms (Holtz et al., 2012; Rohwedder & Uschmajew, 2013) and the DMRG type of algorithms (Schollwöck, 2011b; Huckle et al., 2013).

3.2.2.4 Principle operations in the TT format

In this section, we collects the essential operations in the TT representation which we shall use in the later sections.

Definition 3.22. *Consider tensor matrices* $\boldsymbol{A} = \boldsymbol{A}_1 \bowtie \cdots \bowtie \boldsymbol{A}_N$ *and* $\boldsymbol{B} = \boldsymbol{B}_1 \bowtie \cdots \bowtie \boldsymbol{B}_N$*, where* $\boldsymbol{A}_d \in \mathbb{R}^{R^{\boldsymbol{x}}_{d-1} \times n_d \times R^{\boldsymbol{x}}_d}$ *and* $\boldsymbol{B}_d \in \mathbb{R}^{R^{\boldsymbol{y}}_{d-1} \times n_d \times R^{\boldsymbol{y}}_d}$*,* $1 \leq d \leq N$*. The direct sum,*

$\oplus$, *of the TT cores are defined as*

$$\boldsymbol{A}_d \oplus \boldsymbol{B}_d = \begin{bmatrix} \boldsymbol{A}_d & \\ & \boldsymbol{B}_d \end{bmatrix} \quad \text{for} \quad d = 2, \ldots, N-1,$$

and

$$\boldsymbol{A}_1 \oplus \boldsymbol{B}_1 = \begin{bmatrix} \boldsymbol{A}_1 & \boldsymbol{B}_1 \end{bmatrix} \quad \text{and} \quad \boldsymbol{A}_N \oplus \boldsymbol{B}_N = \begin{bmatrix} \boldsymbol{A}_N \\ \boldsymbol{B}_N \end{bmatrix}.$$

For a series of tensor matrices, $\boldsymbol{A}_\ell = \boldsymbol{A}_{\ell,1} \rtimes \cdots \rtimes \boldsymbol{A}_{\ell,N}$, $1 \le \ell \le L$, *the direct sum,* $\bigoplus$, *is defined as*

$$\bigoplus_{\ell=1}^{L} \boldsymbol{A}_{\ell,d} = \boldsymbol{A}_{1,d} \oplus \cdots \oplus \boldsymbol{A}_{L,d}.$$

Hence, the sum of the TT cores works in a similar way as the usual direct sum of matrices, except that the boundary TT cores (the first or the last TT core) are treated in a special way.

Proposition 3.23. *Consider tensors* $\boldsymbol{x} = \boldsymbol{x}_1 \rtimes \cdots \rtimes \boldsymbol{x}_N$ *and* $\boldsymbol{y} = \boldsymbol{y}_1 \rtimes \cdots \rtimes \boldsymbol{y}_N$, *where* $\boldsymbol{x}_d \in \mathbb{R}^{R^{\boldsymbol{x}}_{d-1} \times n_d \times R^{\boldsymbol{x}}_d}$ *and* $\boldsymbol{y}_d \in \mathbb{R}^{R^{\boldsymbol{y}}_{d-1} \times n_d \times R^{\boldsymbol{y}}_d}$, $1 \le d \le N$. *Consider tensor matrices* $\boldsymbol{A} = \boldsymbol{A}_1 \rtimes \cdots \rtimes \boldsymbol{A}_N$ *and* $\boldsymbol{B} = \boldsymbol{B}_1 \rtimes \cdots \rtimes \boldsymbol{B}_N$, *where* $\boldsymbol{A}_d \in \mathbb{R}^{R^{\boldsymbol{A}}_{d-1} \times n_d \times n_d \times R^{\boldsymbol{A}}_d}$ *and* $\boldsymbol{B}_d \in \mathbb{R}^{R^{\boldsymbol{B}}_{d-1} \times n_d \times n_d \times R^{\boldsymbol{B}}_d}$, $1 \le d \le N$. *Then the following arguments hold:*

(*i*) *Storage cost: the number of entries of* $\boldsymbol{x}$ *for storage in TT format is* $\mathcal{O}(nNR^2)$; *the storage cost of* $\boldsymbol{A}$ *in TT matrix format is* $\mathcal{O}(n^2NR^2)$.

(*ii*) *Matrix addition: the sum of tensor matrices* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *yields another TT tensor* $\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B} = \boldsymbol{C}_1 \rtimes \cdots \rtimes \boldsymbol{C}_N$ *of rank-*$\left(R^{\boldsymbol{A}}_1 + R^{\boldsymbol{B}}_1, \ldots, R^{\boldsymbol{A}}_N + R^{\boldsymbol{B}}_N\right)$. *The TT core*

matrices are computed as

$$\boldsymbol{C}_d = \boldsymbol{A}_d \oplus \boldsymbol{B}_d, \quad 1 \le d \le N,$$

where $\oplus$ is defined in Definition 3.22. There is no appreciable computational cost.

(*iii*) *Multiplied by a scalar: multiplying a scalar $s \in \mathbb{R}$ to the TT matrix $\boldsymbol{A}$ is performed by multiplying s to the entries of one of the TT cores, i.e.,*

$$s\boldsymbol{x} = \boldsymbol{x}_1 \bowtie \cdots \bowtie (s\boldsymbol{x}_i) \bowtie \cdots \bowtie \boldsymbol{x}_N,$$

$\forall i \in \{1, 2, \ldots, N\}$; the computational cost is $\mathcal{O}(nR^2)$.

(*iv*) *Vector inner product: the inner product of TT tensors $\boldsymbol{x}$ and $\boldsymbol{y}$ is computed by*

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \boldsymbol{c}_1 \bowtie \cdots \bowtie \boldsymbol{c}_N,$$

where $\boldsymbol{c}_d \in \mathbb{R}^{(R^{\boldsymbol{x}}_{d-1} R^{\boldsymbol{y}}_{d-1}) \times (R^{\boldsymbol{x}}_d R^{\boldsymbol{y}}_d)}$ has entries $c_d(i,j) = \left\langle \boldsymbol{x}_d^{(i,j)}, \boldsymbol{y}_d^{(i,j)} \right\rangle$. This requires $\mathcal{O}(nNR^4)$ operations.

(*iv*) *Matrix-by-vector product: multiplying $\boldsymbol{A}$ by $\boldsymbol{x}$ from the right yields another TT tensor $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}$, with TT cores*

$$\boldsymbol{y}_d = \boldsymbol{A}_d \bullet \boldsymbol{x}_d \in \mathbb{R}^{\left(R^{\boldsymbol{A}}_{d-1} R^{\boldsymbol{x}}_{d-1}\right) \times n_d \times \left(R^{\boldsymbol{A}}_d R^{\boldsymbol{x}}_d\right)}, \quad 1 \le d \le N.$$

The complexity of the product is $\mathcal{O}(n^2 NR^4)$.

(*v*) *Matrix-by-matrix product: the product of TT matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ yields another TT matrix $\boldsymbol{C} = \boldsymbol{A}\boldsymbol{B} = \boldsymbol{C}_1 \bowtie \cdots \bowtie \boldsymbol{C}_N$, with TT matrix cores given by*

$$\boldsymbol{C}_d = \boldsymbol{A}_d \bullet \boldsymbol{B}_d \in \mathbb{R}^{\left(R^{\boldsymbol{A}}_{d-1} R^{\boldsymbol{x}}_{d-1}\right) \times n_d \times n_d \times \left(R^{\boldsymbol{A}}_d R^{\boldsymbol{x}}_d\right)}, \quad 1 \le d \le N.$$

The complexity of the product is $\mathcal{O}(n^3NR^4)$.

(*vi*) *Mode-d product: the mode-d product of* $\boldsymbol{x}$ *by a matrix* $\boldsymbol{C} \in \mathbb{R}^{n_d \times m_d}$ *is yields* $\boldsymbol{y} = \boldsymbol{x} \times_d \boldsymbol{C} = \boldsymbol{x}_1 \bowtie \cdots \bowtie \boldsymbol{y}_d \bowtie \cdots \bowtie \boldsymbol{x}_N$, *where the TT core* $\boldsymbol{y}_d$ *is composed of the block matrices*

$$\boldsymbol{y}_d^{(i,j)} = \left(\boldsymbol{x}_d^{(i,j)}\right)^T \boldsymbol{C},$$

for $1 \le i \le R_{d-1}^{\boldsymbol{x}}$ *and* $1 \le j \le R_d^{\boldsymbol{x}}$. *Its computational cost is* $\mathcal{O}(n^2R)$.

(*vii*) *Tensor product: the tensor product of* $\boldsymbol{A}$ *and* $\boldsymbol{B}$ *yields another TT matrix* $\boldsymbol{C}$ *of the form*

$$\boldsymbol{C} = \boldsymbol{A} \otimes \boldsymbol{B} = \boldsymbol{A}_1 \bowtie \cdots \bowtie \boldsymbol{A}_N \bowtie \boldsymbol{B}_1 \bowtie \cdots \bowtie \boldsymbol{B}_N,$$

and the TT ranks are $\left(R_1^{\boldsymbol{A}}, \ldots, R_{N-1}^{\boldsymbol{A}}, 1, R_1^{\boldsymbol{B}}, \ldots, R_N^{\boldsymbol{B}}\right)$.

(*viii*) *Diagonalisation: let* $\boldsymbol{A}$ *be a square diagonal matrix whose diagonal entries are* $\boldsymbol{x}$, *i.e.,* $\boldsymbol{A} = \mathsf{diag}(\boldsymbol{x})$. *If* $\boldsymbol{x} = \boldsymbol{x}_1 \bowtie \cdots \bowtie \boldsymbol{x}_N$, *then the same TT ranks applies to* $\boldsymbol{A} = \boldsymbol{A}_1 \bowtie \cdots \boldsymbol{A}_N$, *and the TT matrix cores* $\boldsymbol{A}_d$ *are composed of the diagonal block matrices*

$$\boldsymbol{A}_d^{(i,j)} = \mathsf{diag}\left(\boldsymbol{x}_d^{(i,j)}\right),$$

for $1 \le i \le R_{d-1}^{\boldsymbol{x}}$ *and* $1 \le j \le R_d^{\boldsymbol{x}}$.

We refer the readers to Oseledets (2011) for detailed considerations of the TT operations.

Tensor rank truncation. Algebraic operations (such as additions and matrix multiplications) may significantly increase the TT ranks, and subsequently increase the complexity as in Proposition 3.23. Hence re-compressing, or rank-truncating, the TT tensors is necessary after each algebraic operation. The main building block is as below (Oseledets, 2011):

1. Input a N-th order TT tensor $\boldsymbol{x} = \boldsymbol{x}_1 \bowtie \cdots \bowtie \boldsymbol{x}_N \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ with the TT format with rank-$(R_1, \ldots, R_{N-1})$.

2. For $k = 1, 2, \ldots, N-1$ do

   2a. Unfold the k-th TT core $\boldsymbol{\mathcal{X}}_k \in \mathbb{R}^{R'_{k-1} \times n_k \times R_k}$ into matrix $(\boldsymbol{\mathcal{X}}_k)_{(1,2)} \in \mathbb{R}^{(R'_{k-1} n_k) \times R_k}$;

   2b. Compute the truncated SVD $(\boldsymbol{\mathcal{X}}_k)_{(1,2)} \approx \boldsymbol{U}_k \boldsymbol{\Sigma}_k \boldsymbol{V}_k^T$ by ε after $R'_k < R_k$ terms;

   2c. Store the k-th TT core as $\boldsymbol{\mathcal{X}}_k := \boldsymbol{U}_k$, and reshape $\boldsymbol{\mathcal{X}}_k$ to size $R'_{k-1} \times n_k R'_k$;

   2d. Recompute the $(k+1)$-th TT core as $\boldsymbol{\mathcal{X}}_{k+1} := \boldsymbol{\mathcal{X}}_{k+1} \times_1 \left(\boldsymbol{\Sigma}_k \boldsymbol{V}_k^T\right) \in \mathbb{R}^{R'_k \times n_{k+1} \times R_{k+1}}$.

One left-to-right sweep would yield a tensor of smaller TT ranks, and the TT cores are left-orthogonal, i.e., the mode-$(1,2)$ unfolding of $\boldsymbol{\mathcal{X}}_i$, $i = 1, 2, \ldots, N-1$, are orthogonal. A full rank truncation procedure contains also a right-to-left sweep which is similar to the steps described above but the loop starts from the last core back to the first. The procedure requires computing $2(N-1)$ matrix SVDs of size bounded by $(nR) \times R$, thus the complexity of the TT truncation is $\mathcal{O}(nNR^3)$. Note that the TT matrices (3.39) can be truncated using the same procedure by grouping the row and column indices.

Canonical-to-TT transformation. Usually, the function-related tensors may allow explicit canonical decompositions. One may want to convert the canonical tensor into a TT tensor. This conversion is based on the discovery of the canonical-Tucker conversion and the canonical-Tucker-canonical conversion by Khoromskij & Khoromskaia (2009), which allows the resulting TT ranks from the canonical-TT conversion to be estimated. Since a rank-one tensor is a rank-$(1, \ldots, 1)$ TT tensor, a canonical tensor can be written as a sum of rank-one tensors. Then from Proposition 3.23, the conversion has no appreciable computational cost, and the following result can be derived.

Proposition 3.24 (Oseledets (2011))**.** *If a tensor* $\boldsymbol{\mathcal{A}}$ *admits a canonical decomposition* (3.35) *of rank* R *and accuracy* ε*, then there exists a TT decomposition* (3.38) *of rank-*$(R_1, \ldots, R_{N-1})$ *with accuracy* $\sqrt{N-1}\varepsilon$*, where* $R_i \leq R$*,* $i = 1, 2, \ldots, N-1$.

3.2.2.5 Quantised tensor train decomposition

As opposed to tensorisation which introduce physical dimensions with indices from lower dimensional data, quantisation further splits the indices corresponding to the physical dimensions into many sub-indices referring to the "virtual" dimensions. Then, the virtual dimensions may be decomposed to achieve further storage compression.

If the size n_d of d-th mode of tensor $\boldsymbol{x}$ can be expressed as $n_d = q^{l_d}$ for some $q, l_d \in \mathbb{N}$, then the index $i_d \in \{1,2,\dots,n_d\}$ can be quantised into an l_d-tuple indices $(i_{d,1}, i_{d,2}, \dots, i_{d,l_d})$. The mapping may be defined as $i_d = \sum_{\ell=1}^{l_d} (i_{d,\ell} - 1) q^{\ell}$. Usually, we wish to create as many virtual dimensions as possible, and a fine quantisation with small value of q is desirable.

The concept of quantisation was first introduced into tensor computations by Khoromskij (2009a, 2011); Oseledets (2010). When the TT decomposition is applied to a quantised tensor, the resulting separated representation is referred as the quantised tensor train (QTT).

Definition 3.25. *A tensor $\boldsymbol{x} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$, $n_d = q^{l_d}$, $q \in \mathbb{N}$, $d = 1,2,\dots,N$, is said to be represented in the quantised tensor train (QTT) format, if it is represented in the TT format* (3.38) *of rank-$(R_1, R_2, \dots, R_{N-1})$, and each core of the TT tensor is further decomposed into*

$$\boldsymbol{x}_d = \boldsymbol{x}_{d,1} \bowtie \boldsymbol{x}_{d,2} \bowtie \cdots \bowtie \boldsymbol{x}_{d,l_d}, \quad d = 1,2,\dots,N, \tag{3.40}$$

where the quantised core tensors $\boldsymbol{x}_{d,\ell}$ are of size $R_{d,\ell-1} \times q \times R_{d,\ell}$ with $R_{d,0} = R_{d-1}$ and $R_{d,l_d} = R_d$, $d = 1,2,\dots,N$. The QTT ranks of $\boldsymbol{x}$ is

$$\big(R_{1,1},\dots,R_{1,l_1-1}\big), R_1, \big(R_{2,1},\dots,R_{2,l_2-1}\big), R_2, \dots, R_{N-1}, \big(R_{N,1},\dots,R_{N,l_N}\big). \tag{3.41}$$

Similarly, a tensor matrix $\boldsymbol{A} \in \mathbb{R}^{(n_1 \times \cdots \times n_N) \times (n_1 \times \cdots \times n_N)}$ *is said to be represented in the QTT matrix format, if the core tensors in its TT matrix decomposition* (3.39) *is further decomposed via*

$$\boldsymbol{A}_d = \boldsymbol{A}_{d,1} \bowtie \boldsymbol{A}_{d,2} \bowtie \cdots \bowtie \boldsymbol{A}_{d,l_d}, \quad d = 1,2,\ldots,N, \tag{3.42}$$

where the quantised core tensors $\boldsymbol{A}_{d,\ell}$ *are of size* $R_{d,\ell-1} \times q \times q \times R_{d,\ell}$, $\ell = 1,2,\ldots,l_d$, $d = 1,2,\ldots,N$.

It can be easily verified that the storage requirement of a QTT tensor vector, in addition to the linear scaling in dimensions, scales logarithmically with the volume size, i.e., $\mathcal{O}\left(NR^2 \log_q n\right)$. The complexity in the algebraic operations (such as vector inner product and matrix-vector multiplication) is reduced accordingly.

Since the QTT tensors refer to a special form of TT tensors, thus the QTT decomposition, as well as the rank truncation, are performed in the same way as the TT decomposition, see Oseledets (2010). More often, one tends to introduce as many virtual dimensions as possible, i.e., set $q = 2$ in Definition 3.25, and the sub-indices in the QTT representation is binary.

It is not always beneficial to use the quantised representation over the original TT format, because the QTT ranks can be higher. Fortunately, many function-related data arrays and operator-related matrices have low-rank QTT structure, see for example Kazeev & Khoromskij (2012); Khoromskaia et al. (2011); Khoromskij & Oseledets (2010a); Kazeev et al. (2013); Oseledets (2013); Khoromskaia & Khoromskij (2014a). These techniques have been applied to solve the parametric chemical master equation, see Dolgov & Khoromskij (2013a, 2015). Below we introduce the low-rank QTT structures of certain elementary matrices (vectors) that will be used later in construct the QTT representation of the CFPE operator. As the simplest case, we start with the identity matrix and some elementary functions.

Proposition 3.26 ((Khoromskij, 2011)). *The identity matrix $\boldsymbol{I} \in \mathbb{R}^{n\times n}$, $n = 2^l$, has the rank-1 QTT structure, $\boldsymbol{I} = \Psi_I \bowtie \Psi_I \bowtie \cdots \bowtie \Psi_I$, where $\Psi_I = \left[\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right]$, $1 \le d \le l$. The equidistant sampling of the exponential function and trigonometric function allows the QTT representation of rank 1 and 2, respectively. Further, the QTT rank of the vector from the equidistant sampling of a polynomial function does not exceed $m+1$, where m is the polynomial degree.*

Next, we introduce the low-rank QTT representation of the tridiagonal Toeplitz matrix.

Proposition 3.27 (Lemma 3.1 in Kazeev & Khoromskij (2012)). *Assume that $l \ge 2$ and a, b, c are numbers. Then the matrix*

$$\boldsymbol{A} = \begin{bmatrix} a & b & & & \\ c & a & \ddots & & \\ & \ddots & \ddots & \ddots & \\ & & \ddots & a & b \\ & & & c & a \end{bmatrix}$$

of size $2^l \times 2^l$ has the following QTT representation of ranks-$(3,\dots,3)$:

$$\boldsymbol{A} = \begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \end{bmatrix} \bowtie \begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \\ & \Psi_J & \\ & & \Psi'_J \end{bmatrix}^{\bowtie(l-2)} \bowtie \begin{bmatrix} a\Psi_I + b\Psi_J + c\Psi'_J \\ c\Psi_J \\ b\Psi'_J \end{bmatrix},$$

where

$$\Psi_I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \Psi_I = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}. \tag{3.43}$$

Finally, we give the QTT structure of the data array of polynomial functions.

Proposition 3.28 (Proposition 2.8 in Khoromskij (2011) and Theorem 6 in Oseledets (2013))**.** *Consider vector* $\boldsymbol{f} \in \mathbb{R}^n$*,* $n = 2^\ell$*, obtained by the equidistant sampling of a* p*-th order polynomial* $f(x) = \sum_{\ell=0}^{p} c_\ell x^\ell$*, and* $f_i = f(a + ih)$*,* $1 \le i \le n$*. The QTT decomposition of* $\boldsymbol{f}$ *is of rank-*$(p+1, \dots, p+1)$*, given by*

$$\boldsymbol{f} = \begin{bmatrix} \boldsymbol{f}_1^{(1,1)} & \cdots & \boldsymbol{f}_1^{(1,p+1)} \end{bmatrix} \bowtie \begin{bmatrix} \boldsymbol{f}_2^{(1,1)} & & \\ \vdots & \ddots & \\ \boldsymbol{f}_2^{(p+1,1)} & \cdots & \boldsymbol{f}_2^{(p+1,p+1)} \end{bmatrix} \bowtie \cdots$$
$$\bowtie \begin{bmatrix} \boldsymbol{f}_{l-1}^{(1,1)} & & \\ \vdots & \ddots & \\ \boldsymbol{f}_{l-1}^{(p+1,1)} & \cdots & \boldsymbol{f}_{l-1}^{(p+1,p+1)} \end{bmatrix} \bowtie \begin{bmatrix} \boldsymbol{f}_l^{(1,1)} \\ \vdots \\ \boldsymbol{f}_l^{(p+1,1)} \end{bmatrix}$$

where

$$\boldsymbol{f}_1^{(1,j)} = \sum_{\ell=j-1}^{p} c_\ell C_\ell^{j-1} \left(a \begin{pmatrix} 1 \\ 1 \end{pmatrix} + h \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right)^{\ell-j+1}, \quad \boldsymbol{f}_l^{(i,1)} = 2^{(d-1)(i-1)} h^{i-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad i, j = 1, \dots, p+1,$$
$$\boldsymbol{f}_d^{(i,j)} = C_{i-1}^{i-j} (2^{d-1} h)^{i-j} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \quad j \le i \le p+1, \quad 1 \le j \le p+1,$$

with binomial coefficient $C_\ell^j = \ell!/(j!(\ell - j)!)$*.*

3.3 Low-rank QTT structure of the CFPE operator

3.3.1 Explicit QTT representation of the CFPE operator

In the current section, we construct the explicit QTT representation of the matrix from the difference approximation of the CFPE operator described in Section § 3.1.2.1.

3.3.1.1 Difference operators

Let $\boldsymbol{Q}_d, \boldsymbol{Q}_{\hat{d}} \in \mathbb{R}^{(n_1 \times \cdots \times n_N) \times (n_1 \times \cdots \times n_N)}$ be the tensor matrices corresponding to the upwind difference and the downwind difference operators defined in (3.7) for approximating first-order derivatives, $\partial/\partial x_d$, in the d-th coordinate. It is readily verified that they are rank-one separable of the form,

$$\begin{aligned} \boldsymbol{Q}_d &= \boldsymbol{I}_1 \otimes \cdots \otimes \boldsymbol{I}_{d-1} \otimes \frac{1}{h_d} \boldsymbol{E}_d \otimes \boldsymbol{I}_{d+1} \otimes \cdots \otimes \boldsymbol{I}_N, \\ \boldsymbol{Q}_{\hat{d}} &= \boldsymbol{I}_1 \otimes \cdots \otimes \boldsymbol{I}_{d-1} \otimes \frac{1}{h_d} \boldsymbol{E}_{\hat{d}} \otimes \boldsymbol{I}_{d+1} \otimes \cdots \otimes \boldsymbol{I}_N, \end{aligned} \tag{3.44}$$

where $I_i \in \mathbb{R}^{n_i \times n_i}$, $i = 1, 2, \ldots, N$, denote identity matrices and

$$\boldsymbol{E}_d = \begin{pmatrix} -1 & 1 & & & \\ & -1 & \ddots & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \\ & & & & -1 \end{pmatrix} \quad \text{and} \quad \boldsymbol{E}_{\hat{d}} = \begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ & \ddots & \ddots & & \\ & & \ddots & 1 & \\ & & & -1 & 1 \end{pmatrix}. \tag{3.45}$$

Applying Proposition 3.21, matrices $\boldsymbol{Q}_d$ and $\boldsymbol{Q}_{\hat{d}}$ can be converted to the TT format with rank-$(1, \ldots, 1)$, and the QTT structure can be directly inferred from Propositions 3.26 and 3.27.

Lemma 3.29. *The upwind and downwind difference operators, $\boldsymbol{Q}_d$ and $\boldsymbol{Q}_{\hat{d}}$ in* (3.44)*, of size $(n_1 \cdots n_N) \times (n_1 \cdots n_N)$, $n_d = 2^{l_d}$, $1 \le d \le N$, have the following QTT representa-*

tion

$$\boldsymbol{Q}_d = (\Psi_I \bowtie \cdots \bowtie \Psi_I) \bowtie \cdots$$
$$\bowtie \left(\begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \end{bmatrix} \bowtie \begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \\ & \Psi_J & \\ & & \Psi'_J \end{bmatrix}^{\bowtie(l_d-2)} \bowtie \begin{bmatrix} -\frac{1}{h_d}\Psi_I + \frac{1}{h_d}\Psi_J \\ \\ \frac{1}{h_d}\Psi'_J \end{bmatrix} \right) \bowtie \cdots$$
$$\bowtie (\Psi_I \bowtie \cdots \bowtie \Psi_I),$$

and

$$\boldsymbol{Q}_{\hat{d}} = (\Psi_I \bowtie \cdots \bowtie \Psi_I) \bowtie \cdots$$
$$\bowtie \left(\begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \end{bmatrix} \bowtie \begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \\ & \Psi_J & \\ & & \Psi'_J \end{bmatrix}^{\bowtie(l_d-2)} \bowtie \begin{bmatrix} \frac{1}{h_d}\Psi_I - \frac{1}{h_d}\Psi'_J \\ -\frac{1}{h_d}\Psi_J \\ \\ \end{bmatrix} \right) \bowtie \cdots$$
$$\bowtie (\Psi_I \bowtie \cdots \bowtie \Psi_I),$$

and the QTT ranks are given by

$$(1,\ldots,1),1,\ldots,1,(3,\ldots,3),1,\ldots,1,(1,\ldots,1). \tag{3.46}$$

Similarly, the central difference operator $\boldsymbol{Q}_{d+\hat{d}} = \frac{1}{2}\left(\boldsymbol{Q}_d + \boldsymbol{Q}_{\hat{d}}\right) \in \mathbb{R}^{(n_1\cdots n_N)\times(n_1\cdots n_N)}$ for approximating the first-order derivative, $\partial/\partial x_d$, of the drift part in (3.12) has the rank-one structure

$$\boldsymbol{Q}_{d+\hat{d}} = \boldsymbol{I}_1 \otimes \cdots \otimes \boldsymbol{I}_{d-1} \otimes \frac{1}{2h_d}\boldsymbol{E}_{d+\hat{d}} \otimes \boldsymbol{I}_{d+1} \otimes \cdots \otimes \boldsymbol{I}_N, \tag{3.47}$$

where $\boldsymbol{E}_{d+\hat{d}} = \boldsymbol{E}_d + \boldsymbol{E}_{\hat{d}}$. The QTT structure of $\boldsymbol{Q}_{d+\hat{d}}$ can be derived in a similar man-

ner as in Lemma 3.29.

Lemma 3.30. *The central difference operator* $\boldsymbol{Q}_{d+\hat{d}}$ *in* (3.47) *of size* $(n_1\cdots n_N)\times(n_1\cdots n_N)$, $n_d = 2^{l_d}$, $1 \le d \le N$, *have the following QTT representation with the same QTT ranks as in* (3.46)*:*

$$\begin{aligned}\boldsymbol{Q}_{d+\hat{d}} = {} & (\Psi_I \bowtie \cdots \bowtie \Psi_I) \bowtie \cdots \\ & \bowtie \left(\begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \end{bmatrix} \bowtie \begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \\ & \Psi_J & \\ & & \Psi'_J \end{bmatrix}^{\bowtie(l_d-2)} \bowtie \begin{bmatrix} \frac{1}{2h_d}\Psi_J - \frac{1}{2h_d}\Psi'_J \\ -\frac{1}{2h_d}\Psi_J \\ \frac{1}{2h_d}\Psi_J \end{bmatrix} \right) \bowtie \cdots \\ & \bowtie (\Psi_I \bowtie \cdots \bowtie \Psi_I).\end{aligned}$$

Also, we define tensor matrix $\boldsymbol{Q}_{d_1,\hat{d}_2}$ for approximating the second-order mixed derivatives, $\partial^2/\partial x_{d_1}\partial x_{d_2}$, where the d_1-th direction applies the upwind difference scheme and the d_2-th direction applies the downwind scheme, $d_1 \neq d_2$. It can be shown that it has the following rank-one form

$$\boldsymbol{Q}_{d_1,\hat{d}_2} = \boldsymbol{Q}_{d_1}\boldsymbol{Q}_{\hat{d}_2} = \boldsymbol{I}_1 \otimes \cdots \otimes \frac{1}{h_{d_1}}\boldsymbol{E}_{d_1} \otimes \cdots \otimes \frac{1}{h_{d_2}}\boldsymbol{E}_{\hat{d}_2} \otimes \cdots \otimes \boldsymbol{I}_N, \tag{3.48}$$

where matrices $\boldsymbol{E}_d$ and $\boldsymbol{E}_{\hat{d}}$ are defined in (3.45).

Lemma 3.31. *The QTT-structured representation of tensor matrix* $\boldsymbol{Q}_{d_1,\hat{d}_2}$ *in* (3.48) *is*

given by

$$
\begin{aligned}
\boldsymbol{Q}_{d_1,\hat{d}_2} = (\Psi_I \bowtie \cdots \bowtie \Psi_I) \bowtie \cdots \\
\bowtie \left(\begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \end{bmatrix} \bowtie \begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \\ & \Psi_J & \\ & & \Psi'_J \end{bmatrix}^{\bowtie (l_{d_1}-2)} \bowtie \begin{bmatrix} -\frac{1}{h_{d_1}}\Psi_I + \frac{1}{h_{d_1}}\Psi_J \\ \\ \frac{1}{h_{d_1}}\Psi'_J \end{bmatrix} \right) \bowtie \cdots \\
\bowtie \left(\begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \end{bmatrix} \bowtie \begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \\ & \Psi_J & \\ & & \Psi'_J \end{bmatrix}^{\bowtie (l_{d_2}-2)} \bowtie \begin{bmatrix} \frac{1}{h_{d_2}}\Psi_I - \frac{1}{h_{d_2}}\Psi'_J \\ -\frac{1}{h_{d_2}}\Psi_J \\ \\ \end{bmatrix} \right) \bowtie \cdots \\
\bowtie (\Psi_I \bowtie \cdots \bowtie \Psi_I),
\end{aligned}
$$

and the QTT ranks are given by

$$(1,\ldots,1),1,\ldots,1,(3,\ldots,3),1,\ldots,1,(3,\ldots,3),1,\ldots,1,(1,\ldots,1).$$

Note that the results in Lemma 3.31 can be analogously to tensor matrices $\boldsymbol{Q}_{\hat{d}_1,d_2}$, $\boldsymbol{Q}_{d_1,d_2}$ and $\boldsymbol{Q}_{\hat{d}_1,\hat{d}_2}$, $d_1 \neq d_2$. In case $d_1 = d_2$, we consider the N-dimensional Laplacian operator of the form (Kazeev & Khoromskij, 2012)

$$\boldsymbol{Q}_{d,d} = \boldsymbol{I}_1 \otimes \cdots \otimes \boldsymbol{I}_{d-1} \otimes \boldsymbol{\Delta}_d \otimes \boldsymbol{I}_{d+1} \otimes \cdots \otimes \boldsymbol{I}_N, \tag{3.49}$$

where $\boldsymbol{\Delta}_d$ denotes the Laplacian matrix of size $n_d \times n_d$, and its QTT representation is giving in the following lemma.

Lemma 3.32. *The tensor matrix* $\boldsymbol{Q}_{dd}$ *in* (3.49) *has the following QTT representation:*

$$\begin{aligned}\boldsymbol{Q}_{d,d} = &(\Psi_I \bowtie \cdots \bowtie \Psi_I) \bowtie \cdots \\ &\bowtie \left(\begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \end{bmatrix} \bowtie \begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \\ & \Psi_J & \\ & & \Psi'_J \end{bmatrix}^{\bowtie(l_d-2)} \bowtie \begin{bmatrix} -\frac{2}{h_d^2}\Psi_I + \frac{1}{h_d^2}\Psi_J + \frac{1}{h_d^2}\Psi'_J \\ \frac{1}{h_d^2}\Psi_J \\ \frac{1}{h_d^2}\Psi'_J \end{bmatrix} \right) \bowtie \cdots \\ &\bowtie (\Psi_I \bowtie \cdots \bowtie \Psi_I),\end{aligned}$$

and the QTT ranks are the same as in (3.46)*.*

3.3.1.2 Propensity functions

The evaluation of the j-th propensity functions, α_j defined in (2.2) over the tensor grid $\omega_{\boldsymbol{h}}$ in (3.6) naturally yields an N-th order tensor vector, denoted by $\boldsymbol{\alpha}_j$, with entries

$$\boldsymbol{\alpha}_j = \left(\alpha_j(x_{1,i_1}, x_{2,i_2}, \ldots, x_{N,i_N})\right)_{1\le i_d \le n_d, 1\le d\le N},$$

where x_{d,i_d} refers to the i_d-th node value in the d-th direction. The canonical representations for various classes of function related tensors, with rank estimates, have been discovered by Khoromskij (2006). Based on the author's results, one can show that, due to the product structure of α_j's the propensity function α_j, tensor $\boldsymbol{\alpha}_j$ has the rank-one separable form, and we write it as

$$\boldsymbol{\alpha}_j = k_j \exp\left[\left(1 - \sum_{i=1}^{N} \nu_{j,i}^-\right)\log V\right] \boldsymbol{\alpha}_{j,1} \otimes \boldsymbol{\alpha}_{j,2} \otimes \cdots \otimes \boldsymbol{\alpha}_{j,N},$$

where the single-dimensional factor vectors $\boldsymbol{\alpha}_{j,d}$ have entries

$$\alpha_{j,d}(\ell,\ell) = \left(\nu_{j,i}^-\right)! \binom{x_{i,\ell}}{\nu_{j,i}^-}, \quad 1 \le \ell \le n_d.$$

That is each factor vector $\boldsymbol{\alpha}_{j,d}$ is a equidistant sampling of a $\left(\nu_{j,i}^{-}\right)$-th order polynomial and the QTT structure can be inferred from Proposition 3.28. We summarise the result in the following lemma.

Lemma 3.33. *Consider $\boldsymbol{\alpha}_j \in \mathbb{R}^{(n_1\cdots n_N)\times(n_1\cdots n_N)}$, $n_d = 2^{l_d}$, $1 \le d \le N$, as the grid evaluation of α_j over grid $\omega_{\boldsymbol{h}}$. Its QTT-structured representation is of the form*

$$\boldsymbol{\alpha}_j = \left(\boldsymbol{\alpha}_{j,1,1} \bowtie \cdots \bowtie \boldsymbol{\alpha}_{j,1,l_1}\right) \bowtie \left(\boldsymbol{\alpha}_{j,2,1} \bowtie \cdots \bowtie \boldsymbol{\alpha}_{j,2,l_2}\right) \bowtie \cdots \bowtie \left(\boldsymbol{\alpha}_{j,N,1} \bowtie \cdots \bowtie \boldsymbol{\alpha}_{j,N,l_N}\right), \tag{3.50}$$

where

$$\boldsymbol{\alpha}_{j,d,1}^{(1,j_d)} = \sum_{\ell=j_d-1}^{p} \begin{bmatrix} \nu_{j,i}^{-} \\ \ell \end{bmatrix} C_{\ell}^{j_d-1} \left(x_{d,1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + h_d \begin{pmatrix} 0 \\ 1 \end{pmatrix} \right)^{\ell - j_d + 1}, \quad \boldsymbol{\alpha}_{j,d,l_d}^{(i_d,1)} = 2^{(d-1)(i_d-1)} h_d^{i_d-1} \begin{pmatrix} 0 \\ 1 \end{pmatrix},$$

$$\boldsymbol{\alpha}_{j,d,v_d}^{(i_d,j_d)} = \begin{cases} C_{i_d-1}^{i_d-j_d} (2^{d-1} h_d)^{i_d-j_d} \begin{pmatrix} 0 \\ 1 \end{pmatrix}, & i_d \ge j_d, \\ \begin{pmatrix} 0 \\ 0 \end{pmatrix} & i_d < j_d, \end{cases}$$

for $i_d, j_d = 1, \ldots, \nu_{j,d}^{-} + 1$, $1 \le d \le N$, $1 \le j \le M$. The corresponding QTT ranks are

$$\left(\nu_{j,1}^{-}+1, \ldots, \nu_{j,1}^{-}+1\right), 1, \left(\nu_{j,2}^{-}+1, \ldots, \nu_{j,2}^{-}+1\right), 1, \ldots, 1, \left(\nu_{j,N}^{-}+1, \ldots, \nu_{j,N}^{-}+1\right).$$

Remark 3.34. From Lemma 3.33, we see that the QTT-structure of propensity tensor, $\boldsymbol{\alpha}_j$, has the maximum QTT rank bounded by $\max_{1\le i\le N} \nu_{j,i}^{-}$.

To make the propensity function as a multiplication operator, one usually converts

$\boldsymbol{\alpha}_j$ into a diagonal tensor matrix $\boldsymbol{H}_j \in \mathbb{R}^{(n_1 \cdots n_N)\times(n_1 \cdots n_N)}$ with diagonal entries

$$H_j(i_1,\ldots,i_N;i_1,\ldots,i_N) = \alpha_j(i_1,\ldots,i_N). \tag{3.51}$$

Then applying property (viii) of Propensity 3.23 to Lemma 3.33, we obtain the QTT representation of $\boldsymbol{H}_j$.

Lemma 3.35. *Consider tensor matrix $\boldsymbol{H}_j$, defined element-wise in* (3.51)*. The QTT representation is of the same form as* (3.50)*, i.e.,*

$$\boldsymbol{H}_j = \left(\boldsymbol{H}_{j,1,1} \bowtie \cdots \bowtie \boldsymbol{H}_{j,1,l_1}\right) \bowtie \left(\boldsymbol{H}_{j,2,1} \bowtie \cdots \bowtie \boldsymbol{H}_{j,2,l_2}\right) \bowtie \cdots \bowtie \left(\boldsymbol{H}_{j,N,1} \bowtie \cdots \bowtie \boldsymbol{H}_{j,N,l_N}\right),$$

where the QTT matrix cores $\boldsymbol{H}_{j,d,v_d}$ are composed of diagonal block matrices

$$\boldsymbol{H}_{j,d,v_d}^{(i_d,j_d)} = \mathsf{diag}\left(\boldsymbol{\alpha}_{j,d,v_d}^{(i_d,j_d)}\right) \in \mathbb{R}^{2\times 2}, i_d, j_d = 1,2,\ldots,\nu_{j,d}^- + 1.$$

3.3.1.3 The CFPE operator

Now let us apply the QTT representation of the elementary operators, as described in the previous section, to derive a low-rank QTT representation of the matrix from the finite difference approximation of the CFPE operator.

We start with rewriting the difference operators introduced in Section § 3.1.2.1 in the tensor matrix form, as

$$\begin{aligned} \boldsymbol{\Lambda}_d &= \boldsymbol{Q}_{d+\hat{d}} \boldsymbol{F}_d, \\ \boldsymbol{\Lambda}_{d,d} &= \boldsymbol{Q}_{d,d} \boldsymbol{G}_{d,d}, \end{aligned} \tag{3.52}$$

for $d = 1,2,\ldots,N$, and

$$\begin{aligned} \boldsymbol{\Lambda}_{d_1,d_2}^+ &= \frac{1}{2}\left(\boldsymbol{Q}_{d_1,d_2} + \boldsymbol{Q}_{\hat{d}_1,\hat{d}_2}\right) \boldsymbol{G}_{d_1,d_2}^+, \\ \boldsymbol{\Lambda}_{d_1,d_2}^- &= \frac{1}{2}\left(\boldsymbol{Q}_{d_1,\hat{d}_2} + \boldsymbol{Q}_{\hat{d}_1,d_2}\right) \boldsymbol{G}_{d_1,d_2}^-, \end{aligned} \tag{3.53}$$

for $d_1, d_2 = 1, 2, \ldots, N$. The tensor matrices, $\mathbf{G}^{\pm}_{d_1,d_2} \in \mathbb{R}^{(n_1 \times \cdots \times n_N) \times (n_1 \times \cdots \times n_N)}$, refer to the positive and negative summation of diffusion coefficients, $D^{+}_{d_1,d_2}$ and $D^{-}_{d_1,d_2}$, in (3.10) of the form

$$\begin{aligned}
\boldsymbol{G}^{+}_{d_1,d_2} &= \sum_{r=1}^{M} \nu_{r,d_1} \nu_{r,d_2} \boldsymbol{H}_r \mathbb{1}_{\nu_{r,d_1}\nu_{r,d_2}>0}, \\
\mathbf{G}^{-}_{d_1,d_2} &= \sum_{r=1}^{M} \nu_{r,d_1} \nu_{r,d_2} \mathbf{H}_r \mathbb{1}_{\nu_{r,d_1}\nu_{r,d_2}<0}.
\end{aligned}$$

The tensor matrix $\mathbf{F}_d \in \mathbb{R}^{(n_1 \times \cdots \times n_N) \times (n_1 \times \cdots \times n_N)}$ corresponds to the drift coefficients, μ_d, in (2.11) as

$$\boldsymbol{F}_d = \sum_{j=1}^{M} \nu_{j,d} \boldsymbol{H}_j,$$

where $\boldsymbol{H}_j$, $j = 1, 2, \ldots, M$ are given in (3.51).

Combining the above operators in the same form as in (3.13), we have the construct the QTT representation of the matrix from the finite difference approximation of the CFPE operator, written as

$$\boldsymbol{A}_{\boldsymbol{h},t} \equiv \sum_{d=1}^{N} (\boldsymbol{\Lambda}_d + \boldsymbol{\Lambda}_{d,d}) + \sum_{\substack{d_1,d_2=1 \\ d_1 \neq d_2}}^{N} (\boldsymbol{\Lambda}^{+}_{d_1,d_2} + \boldsymbol{\Lambda}^{-}_{d_1,d_2}). \tag{3.54}$$

Here $\boldsymbol{A}_{\boldsymbol{h},t}$ is an exact permuted reformulation of $\boldsymbol{A}_{\boldsymbol{h}}$ in (3.13) represented in the QTT format. And we arrive at the tensor-formatted version of the matrix principle eigenvalue problem (3.13) as

$$\boldsymbol{A}_{\boldsymbol{h},t} \boldsymbol{p}_{\boldsymbol{h},t} = \lambda_{\boldsymbol{h},t} \boldsymbol{p}_{\boldsymbol{h},t}, \tag{3.55}$$

where $\boldsymbol{p}_{\boldsymbol{h},t}$ is the QTT representation of the "long" vector $\boldsymbol{p}_{\boldsymbol{h}}$.

An immediate result of the formulation in (3.54) is the canonical rank of $\boldsymbol{A}_{\boldsymbol{h},t}$, given in the following proposition.

Proposition 3.36. *Consider tensor matrix* $\boldsymbol{A}_{\boldsymbol{h},t} \in \mathbb{R}^{(n_1 \cdots n_N)\times(n_1 \times n_N)}$, $n_d = 2^{l_d}$, $1 \le d \le N$, *defined in* (3.54). *It admits the canonical matrix representation with the canonical rank bound*

$$R^{\boldsymbol{A}_{\boldsymbol{h},t}} \le 2MN + 2N^2,$$

where M *and* N *are the numbers of reactions and species, respectively, as defined in* (2.1).

Proof. We have already shown in Sections § 3.3.1.1 and § 3.3.1.2 that the elementary operators involved in the CFPE are of rank-one form, thus the rank bound can be directly derived from the explicit structure of $\boldsymbol{A}_{\boldsymbol{h},t}$ in (3.54). □

Then, applying Proposition 3.24 to the canonical rank bound in Proposition 3.36, we obtain the TT rank bound of $\boldsymbol{A}_{\boldsymbol{h},t}$ given in the following proposition.

Proposition 3.37. *There exists a TT matrix representation* (3.39) *of* $\boldsymbol{A}_{\boldsymbol{h},t}$ *in* (3.54) *with TT ranks-*$(R_1^{\boldsymbol{A}_{\boldsymbol{h},t}}, R_2^{\boldsymbol{A}_{\boldsymbol{h},t}}, \ldots, R_{N-1}^{\boldsymbol{A}_{\boldsymbol{h},t}})$ *bounded by* $R_d^{\boldsymbol{A}_{\boldsymbol{h},t}} \le R^{\boldsymbol{A}_{\boldsymbol{h},t}}$, $d = 1, 2, \ldots, N-1$, *where* $R^{\boldsymbol{A}_{\boldsymbol{h},t}}$ *is the canonical rank in Proposition 3.36.*

Finally, we give the QTT structure of $\boldsymbol{A}_{\boldsymbol{h},t}$ in the following theorem.

Theorem 3.38. *Consider tensor matrix* $\boldsymbol{A}_{\boldsymbol{h},t} \in \mathbb{R}^{(n_1 \cdots n_N)\times(n_1 \times n_N)}$, $n_d = 2^{l_d}$, $1 \le d \le N$, *defined in* (3.54). *It admits the QTT-structured representation of the form*

$$\boldsymbol{A}_{\boldsymbol{h},t} = \left(\boldsymbol{A}_{\boldsymbol{h},t,1,1} \bowtie \boldsymbol{A}_{\boldsymbol{h},t,1,2} \bowtie \cdots \bowtie \boldsymbol{A}_{\boldsymbol{h},t,1,l_1}\right) \bowtie \cdots \bowtie \left(\boldsymbol{A}_{\boldsymbol{h},t,N,1} \bowtie \boldsymbol{A}_{\boldsymbol{h},t,N,2} \bowtie \cdots \bowtie \boldsymbol{A}_{\boldsymbol{h},t,N,l_N}\right), \tag{3.56}$$

where[†]

$$
\begin{aligned}
\boldsymbol{A}_{\boldsymbol{h},t,d,v_d} = \bigoplus_{j=1}^{M} \Bigg\{ & \left[\bigoplus_{i=1}^{N} \left(\nu_{j,i} \left(\Psi_I \mathbb{1}_{i\neq d} + \Phi_{d+\hat{d},v_d} \mathbb{1}_{i=d} \right) \right) \oplus \left(\nu_{j,i}^2 \left(\Psi_I \mathbb{1}_{i\neq d} + \Phi_{d,d,v_d} \mathbb{1}_{i=d} \right) \right) \right] \\
& \oplus \left[\bigoplus_{\substack{i_1,i_2=1 \\ i_1\neq i_2}}^{N} \frac{\nu_{j,i_1}\nu_{j,i_2}\mathbb{1}_{\nu_{j,i_1}\nu_{j,i_2}>0}}{2} \left(\left(\Phi_{d,v_d} \oplus \Phi_{\hat{d},v_d} \right) \mathbb{1}_{i_1=d \vee i_2=d} \right) \right] \\
& \oplus \left[\bigoplus_{\substack{i_1,i_2=1 \\ i_1\neq i_2}}^{N} \frac{\nu_{j,i_1}\nu_{j,i_2}\mathbb{1}_{\nu_{j,i_1}\nu_{j,i_2}<0}}{2} \left(\left(\Phi_{d,v_d} \oplus \Phi_{\hat{d},v_d} \right) \mathbb{1}_{i_1=d} + \left(\Phi_{\hat{d},v_d} \oplus \Phi_{d,v_d} \right) \mathbb{1}_{i_2=d} \right) \right] \\
& \oplus \left[\bigoplus_{\substack{i_1,i_2=1 \\ i_1\neq i_2}}^{N} \frac{\nu_{j,i_1}\nu_{j,i_2}\mathbb{1}_{i_1\neq d \wedge i_2\neq d}}{2} \left(\Psi_I \oplus \Psi_I \right) \right] \Bigg\} \\
& \bullet \left\{ \left[\left(k_j V^{1-\sum_{i=1}^{N} \nu_{j,i}^{-}} - 1 \right) \mathbb{1}_{d=N \wedge v_d=l_N} + 1 \right] \boldsymbol{H}_{j,d,v_d} \right\},
\end{aligned}
\tag{3.57}
$$

[†] Here, we have used the indicator function, $\mathbb{1}_{\mathscr{Y}} : \mathscr{X} \mapsto \{0,1\}$, of a subset $\mathscr{Y}$ of a set $\mathscr{X}$, formally defined as

$$
\mathbb{1}_{\mathscr{Y}}(x) := \begin{cases} 0, & x \notin \mathscr{Y} \\ 1, & x \in \mathscr{Y} \end{cases}.
$$

For notational simplicity, we drop the argument and the bracket, and write the simplified version, for example, write $\mathbb{1}_{i\neq d}$ instead of $\mathbb{1}_{\mathbb{N}\setminus\{d\}}(i)$. Also we use the shortened notation, $\mathbb{1}_{i_1\neq d \wedge i_2\neq d}$, to represent the product $\mathbb{1}_{i_1\neq d} \cdot \mathbb{1}_{i_2\neq d}$, and use $\mathbb{1}_{i_1\neq d \vee i_2\neq d}$ to represent $\max\left(\mathbb{1}_{i_1\neq d}, \mathbb{1}_{i_2\neq d}\right)$.

for $1 \le v_d \le l_d - 1$, $1 \le d \le N$, *and*

$$\Phi_{d+\hat{d},1} = \Phi_{d,1} = \Phi_{\hat{d},1} = \Phi_{d,d,1} = \begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \end{bmatrix},$$

$$\Phi_{d+\hat{d},v_d} = \Phi_{d,v_d} = \Phi_{\hat{d},v_d} = \Phi_{d,d,v_d} = \begin{bmatrix} \Psi_I & \Psi'_J & \Psi_J \\ & \Psi_J & \\ & & \Psi'_J \end{bmatrix}, \quad v_d = 2,3,\ldots,l_d-1,$$

$$\Phi_{d,l_d} = \begin{bmatrix} -\frac{1}{h_d}\Psi_I + \frac{1}{h_d}\Psi_J \\ \\ \frac{1}{h_d}\Psi'_J \end{bmatrix}, \quad \Phi_{\hat{d},l_d} = \begin{bmatrix} \frac{1}{h_d}\Psi_I - \frac{1}{h_d}\Psi'_J \\ -\frac{1}{h_d}\Psi_J \\ \\ \end{bmatrix} \quad \Phi_{d+\hat{d},l_d} = \begin{bmatrix} \frac{1}{2h_d}\Psi_J - \frac{1}{2h_d}\Psi'_J \\ -\frac{1}{2h_d}\Psi_J \\ \frac{1}{2h_d}\Psi_J \end{bmatrix},$$

$$\Phi_{d,d,l_d} = \begin{bmatrix} -\frac{2}{h_d^2}\Psi_I + \frac{1}{h_d^2}\Psi_J + \frac{1}{h_d^2}\Psi'_J \\ \frac{1}{h_d^2}\Psi_J \\ \frac{1}{h_d^2}\Psi'_J \end{bmatrix}.$$

The summation operations of TT cores, $\oplus$ *and* $\bigoplus$*, are defined in Definition 3.22. The corresponding QTT ranks, denoted by*

$$\left(R_{1,1}^{\boldsymbol{A}_{\boldsymbol{h},t}},\ldots,R_{1,l_1}^{\boldsymbol{A}_{\boldsymbol{h},t}}\right), R_1^{\boldsymbol{A}_{\boldsymbol{h},t}}, \left(R_{2,1}^{\boldsymbol{A}_{\boldsymbol{h},t}},\ldots,R_{2,l_2}^{\boldsymbol{A}_{\boldsymbol{h},t}}\right), R_2^{\boldsymbol{A}_{\boldsymbol{h},t}},\ldots,R_{N-1}^{\boldsymbol{A}_{\boldsymbol{h},t}}, \left(R_{N,1}^{\boldsymbol{A}_{\boldsymbol{h},t}},\ldots,R_{N,l_2}^{\boldsymbol{A}_{\boldsymbol{h},t}}\right),$$

are bounded by $R_d^{\boldsymbol{A}_{\boldsymbol{h},t}} \le 2MN(N+1)$, $1 \le d \le N-1$, *and*

$$R_{d,v_d}^{\boldsymbol{A}_{\boldsymbol{h},t}} \le \sum_{j=1}^{M}\left(\nu_{j,d}^{-}+1\right)\left[\sum_{i=1}^{N}\left(2\mathbb{1}_{i\neq d}+6\mathbb{1}_{i=d}\right)+\sum_{\substack{i_1,i_2=1\\ i_1\neq i_2}}^{N}\left(2\mathbb{1}_{i_1\neq d\wedge i_2\neq d}+6\mathbb{1}_{i_1=d\vee i_2=d}\right)\right]. \quad (3.58)$$

Proof. Applying the properties of TT operations in the Proposition 3.23, one can directly derive the QTT-structured representation of $\boldsymbol{A}_{\boldsymbol{h},t}$ as we shown in (3.57), and the corresponding QTT ranks in (3.58) can be derived in the similar manner. Note that the inequality in (3.58) is used because certain terms in (3.54) are zero depending on the stoichiometric coefficients. □

Remark 3.39. A slightly crude upper rank bound can be derived from Theorem 3.57. Let $\nu_{\max} = \max_{j,i} \nu_{j,i}^{-}$ and we have

$$R_{d,v_d}^{\boldsymbol{A}_{\boldsymbol{h},t}} \leq (\nu_{\max} + 1)M(2N^2 + 8N - 4). \tag{3.59}$$

Further, let $R_{d,v_d}^{\boldsymbol{A}_{\boldsymbol{h},t}} \leq R$ and assume $l_1 = \cdots = l_N = l$, the bound (3.59) ensure the leading order estimate of the storage requirements and complexity of $\boldsymbol{A}_{\boldsymbol{h},t}$ in QTT format to be:

$$\mathcal{O}(M^2 N^5 l), \tag{3.60}$$

instead of $\mathcal{O}(2^{Nl})$ for the full matrix representation.

3.3.2 Operator compression in the QTT format

Although the estimate (3.60) scales as the fifth power of the number of dimensions (the number of chemical species), we find in practice the "optimal" QTT ranks can be significant lower. To achieve a smaller rank, the TT rank procedure (see Section § 3.2.2.4) is applied to $\boldsymbol{A}_{\boldsymbol{h},t}$ to seek for an ε-approximation $\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$ with the "optimal" QTT ranks, satisfying

$$\|\boldsymbol{A}_{\boldsymbol{h},t} - \boldsymbol{A}_{\boldsymbol{h},\hat{t}}\| \leq \varepsilon \|\boldsymbol{A}_{\boldsymbol{h},t}\|,$$

where ε is the required accuracy level. Here, $\|\boldsymbol{p}\|$ denotes any vector norm for vector $\boldsymbol{p}$, and $\|\mathbf{A}\|$ denotes the corresponding matrix norm. Under certain assumptions (Hackbusch et al., 2005), a suboptimal rank bound $\hat{R}$ can scale with ε in the order of $\mathcal{O}(\log^2 \varepsilon^{-1})$.

The consequence of applying the TT rank truncation procedure is a modified prin-

ciple eigenvalue problem, written as

$$\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\boldsymbol{p}_{\boldsymbol{h},\hat{t}} \equiv (\boldsymbol{A}_{\boldsymbol{h},t} + \delta\boldsymbol{A}_{\boldsymbol{h},t})\boldsymbol{p}_{\boldsymbol{h},\hat{t}} = \lambda_{\boldsymbol{h},\hat{t}}\boldsymbol{p}_{\boldsymbol{h},\hat{t}}, \tag{3.61}$$

where $(\lambda_{\boldsymbol{h},\hat{t}}, \boldsymbol{p}_{\boldsymbol{h},\hat{t}})$ stands for the principal eigen-pair of $\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$, and $\delta\boldsymbol{A}_{\boldsymbol{h},t} = \boldsymbol{A}_{\boldsymbol{h},\hat{t}} - \boldsymbol{A}_{\boldsymbol{h},t}$ represents the perturbation caused by tensor truncation. It is of interest here to obtain bounds for the differences $|\lambda_{\boldsymbol{h},\hat{t}} - \lambda_{\boldsymbol{h},t}|$ and $\|\boldsymbol{p}_{\boldsymbol{h},\hat{t}} - \boldsymbol{p}_{\boldsymbol{h},t}\|$. Such error bounds can be analogically obtained as the perturbation bounds for the principal eigenvalues and eigenvectors of corresponding matrices (Deif, 1995). We state it as following theorem.

Theorem 3.40. *Let $(\lambda_{\boldsymbol{h},\hat{t}}, \boldsymbol{p}_{\boldsymbol{h},\hat{t}})$ and $(\lambda_{\boldsymbol{h},t}, \boldsymbol{p}_{\boldsymbol{h},t})$ be the principal eigenpairs for the QTT-formatted tensor matrices $\boldsymbol{A}_{\boldsymbol{h},t}$ in* (3.61) *and $\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$ in* (3.54)*, respectively. If $\|\delta\boldsymbol{A}_{\boldsymbol{h},t}\| \leq \varepsilon\|\boldsymbol{A}_{\boldsymbol{h},t}\|$, for sufficiently small ε, then we have*

$$|\lambda_{\boldsymbol{h},\hat{t}} - \lambda_{\boldsymbol{h},t}| \leq C_1\varepsilon, \quad \text{and} \quad \|\boldsymbol{p}_{\boldsymbol{h},\hat{t}} - \boldsymbol{p}_{\boldsymbol{h},t}\| \leq C_2\varepsilon, \tag{3.62}$$

where C_1, C_2 are positive constants.

3.3.2.1 Numerical study of tensor truncation error

Recall in Section § 3.1.3 that the CFPE (3.30) for the birth-death system (2.29) is discretised using the proposed difference scheme. Here, we follow the previous sections, and establish the QTT matrix representation (3.42) of the assembled tensor matrix $\boldsymbol{A}_{\boldsymbol{h},t}$ in (3.54) for the birth-death system, and test how the tensor truncation procedure of $\boldsymbol{A}_{\boldsymbol{h},t}$ would cause deflection in the principal eigenpair $(\lambda_{\boldsymbol{h},t}, \boldsymbol{p}_{\boldsymbol{h},t})$ of the eigenvalue problem (3.55) towards the principal eigenpair $(\lambda_{\boldsymbol{h},\hat{t}}, \boldsymbol{p}_{\boldsymbol{h},\hat{t}})$ of the approximated problem (3.61). For different tolerances in tensor rank truncation, we always first assemble the operator $\boldsymbol{A}_{\boldsymbol{h},t}$ in QTT format, apply TT-rounding algorithm (Oseledets, 2011) with the prescribed tolerance to truncate $\boldsymbol{A}_{\boldsymbol{h},t}$ to form $\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$, unfold tensor matrix representation of $\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$ into its standard matrix representation, and

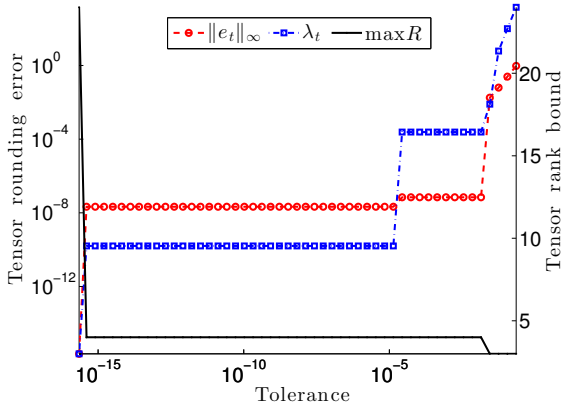

Figure 3.6: *Numerical illustration of the tensor truncation error in solving the CFPE of the birth-death process* (2.29)*. The model parameters are the same as in Figure 3.2. Left y-axis: the principal eigenvalue λ_t and the tensor rounding error, $\|\boldsymbol{p}_{\boldsymbol{h},t} - \boldsymbol{p}_{\boldsymbol{h},\hat{t}}\|$, where $\boldsymbol{p}_{\boldsymbol{h},t}$ is defined in* (3.55)*, and $\lambda_{\boldsymbol{h},\hat{t}}, \boldsymbol{p}_{\boldsymbol{h},\hat{t}}$ are defined in* (3.61)*. Right y-axis: maximum QTT ranks of the assembled tensor matrix $\boldsymbol{p}_{\boldsymbol{h},\hat{t}}$. The lower bound for x-axis is the machine epsilon 2.2204×10^{-16}, meaning that the tensor truncation is effectively exact at that point. The computational domain Ω is discretised by 2^{10} equidistant nodes.*

then use the Matlab `eigs` function to compute the principal eigenpair $(\lambda_{\boldsymbol{h},\hat{t}}, \boldsymbol{p}_{\boldsymbol{h},\hat{t}})$. In this way, we could exactly measure the error, $\boldsymbol{p}_{\boldsymbol{h},t} - \boldsymbol{p}_{\boldsymbol{h},\hat{t}}$, caused by tensor rounding of the QTT matrix $\boldsymbol{A}_{\boldsymbol{h},t}$.

In Figure 3.6, we could see the trend that the error in ℓ^∞-norm increases for larger tolerance values. We could also observe that, for ranges of tolerance values, the tensor rounding error $\|\boldsymbol{p}_{\boldsymbol{h},t} - \boldsymbol{p}_{\boldsymbol{h},\hat{t}}\|_\infty$, together with $\lambda_{\boldsymbol{h},\hat{t}}$, remains at a constant level. This is, however, not predicted by our Theorem 3.40. The reason is that the tensor rounding algorithm is SVD-based (see Section § 3.2.2.3 or Oseledets (2011)), meaning that the rank truncation is based on the magnitude of the singular values rather than certain matrix norm. Thus different tolerance values in tensor truncation procedure may generate the same result, especially when there exist large spectrum gaps. Also, Figure 3.6 shows the maximum QTT ranks are bounded by 24, which is the same as our prediction in Theorem 3.38. But even for very small tolerance values, the maximum rank of $\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$ is reduced to 4, by only causing an error of order 10^{-8} in the approximated principal eigenvector.

Note that the analytical solution for simple birth-death process has been derived in (2.30). The formula is a product of a exponential function, $\exp(-2x)$, and a polynomial function, $(k_1V+k_2x)^{4k_1V/k_2-1}$. From Proposition 3.26 and the product rule of the QTT ranks, the ranks of the exact QTT representation of the stationary distribution do not exceed $(k_1V+k_2x)^{4k_1V/k_2-1}$.

3.4 Steady-state computation

In Section § 3.3, we have shown the stiffness matrix $\boldsymbol{A}_{\boldsymbol{h}}$ in (3.13) can be formulated with tensor matrix $\boldsymbol{A}_{\boldsymbol{h},t}$ in (3.54), which admits explicit QTT decomposition. In the current section, we discussion how to solve the QTT-formatted principal eigenvalue problem. First of all, in Section § 3.4.1, we show there exist low-rank tensor approximations to the principal eigenfunctions under feasible conditions. Then, in Section § 3.4.2, we present an inverse scheme in tensor formats that aims to search for such low-rank approximations, and study the algebraic error of the proposed algorithm.

3.4.1 Existence of low-rank approximation

Before presenting any algorithm to solve the principle eigenvalue problem (3.61), we are interested to understand whether there exists a low-rank QTT approximation of the solution. We want to show there exists a QTT tensor $\boldsymbol{p}_{\boldsymbol{h},\hat{t}}$ in the form of (3.40) that gives an ε-approximation of the solution $\boldsymbol{p}_{\boldsymbol{h},\hat{t}}$, i.e., $\left\|\boldsymbol{p}_{\boldsymbol{h},\hat{t}}-\hat{\boldsymbol{p}}_{\boldsymbol{h},\hat{t}}\right\|\leq\varepsilon$.

To achieve this, we first show the existence of ε-approximation in the case where the solution $\boldsymbol{p}_{\boldsymbol{h},\hat{t}}$ is (approximately) a grid evaluation of the Gaussian distribution. Then, the obtained rank bound is extended to more general distributions.

3.4.1.1 Gaussian distributions

We start with the cases where the solution is approximately a Gaussian distribution. In single-dimensional cases, the rank bounds are subject to the following lemma.

Lemma 3.41 (Lemma 2.4 in Dolgov et al. (2012))**.** *Suppose uniform grid points* $-a = x_0 < x_1 < \cdots < x_n = a$, $x_i = -a + hi$, $n = 2^l$, *are given on an interval* $[-a,a]$ *and the vector* $\boldsymbol{p}$ *is defined by its elements* $p_i = \exp(-x_i^2/2\sigma^2)$, $i = 0,\ldots,N$. *Suppose in addition that* $\exp(-a^2/2\sigma^2) \le \epsilon$. *Then for all sufficiently small* $\epsilon > 0$ *there exists the QTT approximation* $\hat{\boldsymbol{p}}$ *with ranks bounded as*

$$R \le C\frac{a}{\sigma}\sqrt{\log\left(\frac{1}{\epsilon}\frac{\sigma}{1+a}\right)},$$

and the accuracy

$$\|\boldsymbol{p} - \hat{\boldsymbol{p}}\| \le \left(\frac{C}{\sigma}\sqrt{\log\left(\frac{1}{\epsilon}\frac{\sigma}{1+a}\right)} + 1\right)\epsilon,$$

where C *is a constant that does not depend on* a, σ, ϵ, *or* n.

Since the multidimensional Gaussian function is a product of one-dimensional counterparts, its canonical separation ranks are equal to 1. By the triangular inequality, an error bound of the QTT approximation in N dimensions can be derived. Thus, we extend Lemma 3.41 to the multidimensional cases as stated in the following lemma.

Lemma 3.42. *Consider an* N*-dimensional hypercube* $\Omega = \mathcal{I}_1 \times \cdots \times \mathcal{I}_N$ *with* $\mathcal{I}_d = [a_d, b_d]$ *discretised by uniform grid nodes* $(x_{1,i_1},\ldots,x_{N,i_N})$, *where* $x_{d,i_d} = a_d + h_d i_d$, $i_d = 1,\ldots,n_d$, $n = 2^{l_d}$, $d = 1,\ldots,N$. *Let tensor* $\boldsymbol{p}$ *is defined by its elements*

$$p_{i_1,\ldots,i_N} = \prod_{d=1}^{N} \exp\left(-\frac{(x_{d,i_d} - \mu_d)^2}{2\sigma_d^2}\right). \tag{3.63}$$

Suppose in addition that

$$\max_{1\le d\le N}\left(\sqrt{2\pi}\sigma_d - \int_{a_d}^{b_d}\exp\left(-\frac{(x_d-\mu_d)^2}{2\sigma_d^2}\right)\mathrm{d}x_d\right)\le\epsilon<2.$$

Then for sufficiently small $\epsilon>0$, there exists the QTT approximation $\hat{\boldsymbol{p}}$ with ranks bounded as

$$R_d=1 \quad\text{and}\quad R_{d,\ell}\le C\frac{L_d}{2\sigma_d}\sqrt{\log\left(\frac{1}{\epsilon}\frac{\sigma_d}{1+L_d/2}\right)}, \tag{3.64}$$

for $\ell=1,2,\ldots,l_d$, $d=1,2,\ldots,N$, with the accuracy of the QTT approximation

$$\|\boldsymbol{p}-\hat{\boldsymbol{p}}\|\le\max_d\left(\frac{C}{\sigma_d}\sqrt{\log\left(\frac{1}{\epsilon}\frac{\sigma_d}{1+L_d/2}\right)}+1\right)N\epsilon, \tag{3.65}$$

where $L_d=(b_d-a_d)/2$ and C is a constant that does not depend on L_d, σ_d, ϵ, n_d, or N.

Proof. The result can be obtained directly from Definition 3.25 and Lemma 3.41. □

For fixed h_d's, a more concentrated Gaussian distribution, with a larger ratio of L_d over σ_d, would give rise to smaller separation ranks in (3.64) and smaller approximation error in (3.65). On the other hand, this could also be achieved by increasing h_d's while fixing L_d's and σ_d's.

Remark 3.43. Tolerance ϵ implicitly impose restrictions on the choice of L_d and σ_d. By requiring

$$\exp\left(-\frac{\left(b_d-\mu_d\right)^2}{2\sigma_d^2}\right)=\exp\left(-\frac{\left(a_d-\mu_d\right)^2}{2\sigma_d^2}\right)\le\epsilon,$$

we have

$$L_d\ge 2\sqrt{2}\sigma_d\sqrt{\log\epsilon^{-1}},$$

such that

$$R_{d,\ell}\sim\mathcal{O}(\log(1/\epsilon)) \quad\text{and}\quad \|\boldsymbol{p}-\hat{\boldsymbol{p}}\|\sim\mathcal{O}(N\epsilon).$$

Then, an estimate for the approximation error with respect to the tensor ranks could be

$$\|\boldsymbol{p} - \hat{\boldsymbol{p}}\| \sim \mathcal{O}(N \exp(-R)).$$

3.4.1.2 Non-Gaussian distributions

In the following, we generalize the QTT approximation of the Gaussian solutions to the cases of more general classes of N-dimensional distributions. Let us consider the class of $\boldsymbol{p}$ in (3.40) equivalent to certain analytical functions $P(\boldsymbol{x})$ by grid point evaluation. We assume that $P(\boldsymbol{x})$ allows the efficient approximation in the set of Gaussian distributions on Ω. Then we prove the following error bound for the QTT approximation.

Proposition 3.44. *Let $\Omega = \mathcal{I}_1 \times \cdots \times \mathcal{I}_n$ with $\mathcal{I}_d = [a_d, b_d]$, $d = 1, \ldots, N$. Suppose that for a given continuous function $P : \Omega \to \mathbb{R}$, and given $\epsilon > 0$, there is an approximation by Gaussian sums such that*

$$\max_{\boldsymbol{x} \in \Omega} \left| P(\boldsymbol{x}) - \sum_{\ell=1}^{Z} c_\ell G\left(\boldsymbol{\mu}^{(\ell)}, \boldsymbol{\sigma}^{(\ell)}\right) \right| \leq \epsilon, \tag{3.66}$$

where $G\left(\boldsymbol{\mu}^{(\ell)}, \boldsymbol{\sigma}^{(\ell)}\right)$ is the N-dimensional Gaussian function defined by

$$G\left(\boldsymbol{\mu}^{(\ell)}, \boldsymbol{\sigma}^{(\ell)}\right) = \prod_{d=1}^{N} \exp\left(-\left(x_d - \mu_d^{(\ell)}\right)^2 / 2\left(\sigma_d^{(\ell)}\right)^2\right).$$

In addition, we assume that

$$\max_{\substack{1 \leq d \leq N \\ 1 \leq \ell \leq Z}} \sqrt{2\pi}\sigma_d^{(\ell)} - \int_{a_d}^{b_d} \exp\left(-\left(x_d - \mu_d^{(\ell)}\right)^2 / 2\left(\sigma_d^{(\ell)}\right)^2\right) \mathrm{d}x_d \leq \epsilon < 2.$$

Then, consider an N-dimensional tensor $\boldsymbol{p}$ defined by its entries $p_{i_1,\ldots,i_N} = P(x_{1,i_1}, \ldots, x_{N,i_N})$, for $i_d = 1, \ldots, n_d$, $d = 1, \ldots, N$, where $x_{d,i_d} = a_d + h_d i_d$, $h_d = (b_d - a_d)/n_d$. It allows an

QTT tensor $\hat{\boldsymbol{p}}$ *in the form of* (3.40)*, with ranks bounded by*

$$R_d = Z \quad \text{and} \quad R_{d,\ell} \le C_1 \frac{Z L_d}{\min_\ell \sigma_d^{(\ell)}} \sqrt{\log\left(\frac{1}{\epsilon} \frac{\max_\ell \sigma_d^{(\ell)}}{1+L_d/2}\right)}, \tag{3.67}$$

for $\ell = 1,\dots,l_d$*,* $d = 1,\dots,N$*, and the accuracy*

$$\|\boldsymbol{p} - \hat{\boldsymbol{p}}\| \le C_2 Z N \epsilon,$$

where C_1 *is a positive constant that does not depend on* L_d*,* $\sigma_d^{(\ell)}$*,* ϵ*,* n_d*,* N*, or* Z*, and* C_2 *is a positive constant that does not depend on* N*,* Z*, or* ϵ*.*

Proof. We define tensors $\mathbf{g}_\ell$ by grid evaluation of the N-dimensional Gaussian $G(\boldsymbol{\mu}^{(\ell)}, \boldsymbol{\sigma}^{(\ell)})$, for $\ell = 1,\dots,Z$, as in (3.63). From Lemma 3.42, there exists an QTT approximation $\hat{\mathbf{g}}_\ell$, with ranks $R^{(\ell)}$ bounded by (3.64) and accuracy given by (3.65). Hence, we define $\hat{\boldsymbol{p}}$, as a QTT approximation of $\boldsymbol{p}$, by $\hat{\boldsymbol{p}} = \sum_{\ell=1}^{Z} \hat{\mathbf{g}}$. Using the addition rules of QTT ranks in Proposition 3.23, the bounds in (3.67) can be justified.

To prove the accuracy, we use the triangular inequality

$$\|\boldsymbol{p} - \hat{\boldsymbol{p}}\| \le \left\| \boldsymbol{p} - \sum_{\ell=1}^{Z} \mathbf{g}_\ell \right\| + \sum_{\ell=1}^{Z} \left\| \mathbf{g}_\ell - \hat{\mathbf{g}}_\ell \right\|,$$

where the first term is bounded by the assumption (3.66), and a bound from the second term can be inferred from Remark 3.43. □

Note that the rank estimate in Proposition 3.44 specifically applies to Gaussian sums, and the results for other classes of analytical functions are presented and discussed in details in Khoromskij (2006, 2009a); Khoromskij & Khoromskaia (2009); Khoromskij (2011).

Remark 3.45. Similarly as in Remark 3.43, we can derive, from Proposition 3.44, that

$$R_{d,\ell} \sim \mathcal{O}(Z \log(1/\epsilon)) \quad \text{and} \quad \|\boldsymbol{p} - \hat{\boldsymbol{p}}\| \sim \mathcal{O}(ZN\epsilon).$$

Now, if we define tolerance ε by requiring $\|\boldsymbol{p}_{\boldsymbol{h},\hat{t}} - \hat{\boldsymbol{p}}_t\| \sim \mathcal{O}(\varepsilon)$, where $\boldsymbol{p}_{\boldsymbol{h},\hat{t}}$ and $\hat{\boldsymbol{p}}_t$ are defined in (3.61) and (3.40), respectively, then there exists a QTT representation $\hat{\boldsymbol{p}}_t$ whose separation ranks scale as

$$R \sim \mathcal{O}(Z \log(ZN/\varepsilon)), \tag{3.68}$$

where $R \geq R_d, R_{d,\ell}$, $\ell = 1, \ldots, l_d$, $d = 1, \ldots, N$.

Back to our question at the beginning of this section about the existence of a low-rank ϵ-approximation to the solution of (3.61). Remark 3.45 imposes a key condition that the eigenfunctions need to be well approximated by the sum of a minimum number of Gaussian functions. And the peaks of these Gaussian functions have to be significantly away from the boundary $\partial\Omega$. Also, Khoromskaia & Khoromskij (2014a) have proved that the QTT ranks for the large lattice sum of Gaussian-type functions are bonded uniformly in the number of summands. Unfortunately, priori conditions on the operator $\boldsymbol{A}_{\boldsymbol{h},t}$ are still unclear.

3.4.2 The higher-order inverse power iteration

Given the existence of the low-rank QTT ε-approximation of $\boldsymbol{p}_{\boldsymbol{h},\hat{t}}$ as in Proposition 3.44, we are now in the position to develop the QTT-formatted algorithms to compute $\boldsymbol{p}_{\boldsymbol{h},\hat{t}}$. We first introduce a higher order analogue of the inverse iterations as presented in Algorithm 2.

If we assume the tensor linear systems in step 2 of Algorithm 2 are solved precisely, and the classic result on the inverse iterations applies (Demmel, 1997). That is,

Algorithm 2 Inverse power method in tensor format.

Require: Tensor matrix $\boldsymbol{A}_{\boldsymbol{h},\hat{t}} \in \mathbb{R}^{(n_1\times\cdots\times n_N)\times(n_1\times\cdots\times n_N)}$ in the QTT format, an positive shift $\sigma_{\text{shift}} > 0$, an accuracy level $\varepsilon > 0$, and an initial guess $\boldsymbol{p}_0 \in \mathbb{R}^{n_1\times\cdots\times n_N}$ in the QTT format.

1: **for** $k = 1,2,\ldots$ till convergence **do**
2: $\quad\left(\boldsymbol{A}_{\boldsymbol{h},\hat{t}} - \sigma_{\text{shift}}\boldsymbol{I}\right)\boldsymbol{p}_k = \boldsymbol{p}_{k-1} + \boldsymbol{r}_k$, with residue $\|\boldsymbol{r}_k\| \le \varepsilon \left\|\boldsymbol{p}_k\right\|$;
3: $\quad\boldsymbol{p}_k \leftarrow \boldsymbol{p}_k/\|\boldsymbol{p}_k\|$;
4: **end for**

beginning with an initial tensor $\boldsymbol{p}_0$, the series $\left\{\boldsymbol{p}_k\right\}_{k=1,2,\ldots}$ satisfying

$$\left(\boldsymbol{A}_{\boldsymbol{h},\hat{t}} - \sigma_{\text{shift}}\boldsymbol{I}\right)\boldsymbol{p}_k = \boldsymbol{p}_{k-1}, \tag{3.69}$$

would converge to the eigenvector corresponding to the eigenvalue closest to the chosen shift σ_{shift}. For sufficiently small perturbation $\delta\boldsymbol{A}_{\boldsymbol{h},t}$ in (3.61), $\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$ is an M-matrix (see Definition 3.3). Then from Lemma 3.7, any non-negative σ_{shift} would lead to the correct convergence direction towards $(\lambda_{\boldsymbol{h},\hat{t}}, \boldsymbol{p}_{\boldsymbol{h},\hat{t}})$.

However, the assumption of zero the residue tensor $\boldsymbol{r}_k$ is not generally realistic. First, the computational time would increase significantly for the QTT-formatted linear solvers to achieve very high accuracy, and it is usually not necessary (as we will prove later). Second, the standard routine is to apply the tensor truncation after each inverse iteration to avoid uncontrollable growth of the QTT ranks of the intermediate solutions (Oseledets, 2011; Holtz et al., 2012). The rank truncation also contributes to the residue. Therefore, it is of interest to analyse the effect of the residues $\boldsymbol{r}_k$ on the final convergence.

Note that, in the current section, our discussion on solving the tensor linear system (3.69) would always be based on the assumption that the chosen tensor linear solver converges for the prescribed σ_{shift} value. But we will show in Section § 3.4.3 this is generally not the case, and additional efforts are need to choose a suitable σ_{shift} value.

3.4.2.1 Algebraic error

The error analysis of Algorithm 2 is analogous to the analysis for the inexact inverse power method (Golub & Ye, 2000). Let $(\lambda^i_{\boldsymbol{h},\hat{t}}, \boldsymbol{u}_i, \boldsymbol{v}_i)$ for $i = 1,2,\ldots,n$ be a complete set of eigen-triples of $\boldsymbol{A}_{\boldsymbol{h},\hat{t}} \in \mathbb{R}^{(n_1\times\cdots\times n_N)\times(n_1\times\cdots\times n_N)}$ satisfying $0 > \lambda^1_{\boldsymbol{h},\hat{t}} > \lambda^2_{\boldsymbol{h},\hat{t}} \geq \cdots$. For simplicity, let us define $n = n_1 \times \cdots \times n_N$, it follows that

$$\boldsymbol{u}_i^T \boldsymbol{A}_{\boldsymbol{h},\hat{t}} = \lambda^i_{\boldsymbol{h},\hat{t}} \boldsymbol{u}_i^T, \;\; \boldsymbol{A}_{\boldsymbol{h},\hat{t}} \boldsymbol{v}_i = \lambda^i_{\boldsymbol{h},\hat{t}} \boldsymbol{v}_i \;\text{ and }\; \boldsymbol{u}_i^T \boldsymbol{A}_{\boldsymbol{h},\hat{t}} \boldsymbol{v}_j = \delta_{i,j}, \;\text{ for }\; i,j = 1,2,\ldots,n,$$

where $\delta_{i,j}$ is the Kronecker symbol. We assume the solution after k inverse iterations can be expanded as a linear combination of right eigenvectors:

$$\boldsymbol{p}_k = \sum_{i=1}^{n} c_i^{(k)} \boldsymbol{v}_i, \;\text{ where }\; c_i^{(k)} = \boldsymbol{u}_i^T \boldsymbol{A}_{\boldsymbol{h},\hat{t}} \boldsymbol{p}_k. \tag{3.70}$$

Then a measure of the approximation of $\boldsymbol{p}_k$ to $\boldsymbol{v}_1$ for $c_1^{(k)} \neq 0$ can be defined as

$$t_k = \|(c_2^{(k)}, c_3^{(k)}, \ldots, c_n^{(k)})\| / |c_1^{(k)}|.$$

The convergence of t_k can be deduced from the corresponding matrix analysis, see Lemma 2 and 3 in (Golub & Ye, 2000), and for Algorithm 2, we state the tensor version as below.

Lemma 3.46. *Let $\rho = |(\lambda_1 - \sigma_{\text{shift}})/(\lambda_2 - \sigma_{\text{shift}})| < 1$, and $c_1^{(k)} \neq 0$. Let columns of tensor matrices $\boldsymbol{U}$ and $\boldsymbol{V}$ contain all left and right eigenvectors of $\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$, respectively. Then,*

$$t_k \leq \rho t_{k-1} + \varepsilon C, \tag{3.71}$$

where constant $C \leq \|\mathbf{U}^T\| \|\mathbf{V}^T\| (1 + t_0)^2$.

Then we have the convergence of the algebraic error in the following theorem.

Theorem 3.47. *Consider $\sigma_{\text{shift}} > 0$ in Algorithm 2, and the spectrum of $\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$ is distributed in the negative half plane. Let $\langle \boldsymbol{p}_{\boldsymbol{h},\hat{t}}, \boldsymbol{p}_k \rangle \neq 0$ and let $\boldsymbol{p}_k$ is normalised such that $\langle \boldsymbol{p}_{\boldsymbol{h},\hat{t}}, \boldsymbol{p}_k \rangle = 1$. Then, the convergence of $\boldsymbol{p}_k$ towards the principal eigenvector $\boldsymbol{p}_{\boldsymbol{h},\hat{t}}$ of $\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$ is given by*

$$\left\| \boldsymbol{p}_{\boldsymbol{h},\hat{t}} - \boldsymbol{p}_k \right\| \leq C_1 \rho^k + \varepsilon C_2 \frac{1-\rho^k}{1-\rho}, \tag{3.72}$$

where ρ is defined in Lemma 3.46, and C_1 and C_2 are constants that do not depend on parameters ρ and k.

Proof. It follows directly from (3.71) that, for $\forall k = 1, 2, \ldots$, we have

$$t_k \leq \rho t_{k-1} + \varepsilon C \leq \cdots \leq \rho^k t_0 + \rho^{k-1} \varepsilon C + \cdots + \varepsilon C = \rho^k t_0 + \varepsilon C \frac{1-\rho^k}{1-\rho}.$$

By requesting $\langle \boldsymbol{p}_{\boldsymbol{h},\hat{t}}, \boldsymbol{p}_k \rangle = 1$, we have $c_1^{(k)} = 1$. Then,

$$\left\| \boldsymbol{p}_{\boldsymbol{h},\hat{t}} - \boldsymbol{p}_k \right\| = \left\| \sum_{i=2}^{n} c_i^{(k)} \boldsymbol{v}_i \right\| \leq \sum_{i=2}^{n} c_i^{(k)} \left\| \boldsymbol{v}_i \right\| \leq t_k,$$

where $c_i^{(k)}$ and $\boldsymbol{v}_i$ are defined in (3.70). □

Remark 3.48. Theorem 3.47 suggests that, for small ρ, Algorithm 2 converges with algebraic error of $\mathcal{O}(\varepsilon)$. The inexactness of the tensor linear solvers and the tensor truncation procedure will gradually dominate the algebraic error as approaching to the final stage of computation.

3.4.2.2 Numerical simulations

The birth-death process. In Section § 3.3.2.1, the principle eigenvalue problem associated to the CFPE of the birth-death process was solved using the Matlab `eigs` function, while here, we keep all objects in the QTT format and apply the higher

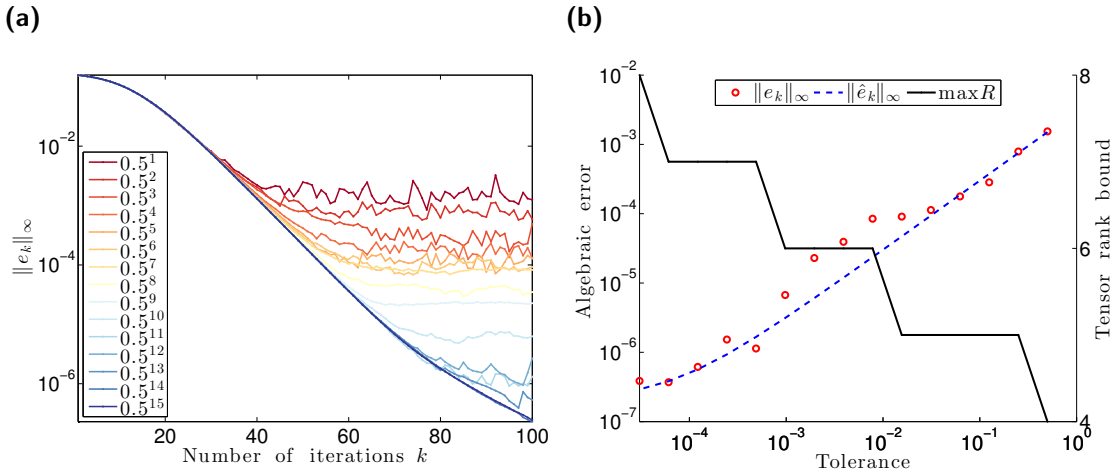

Figure 3.7: *Algebraic error analysis of the inverse power method in tensor format applied to the birth death process. Simulation parameters:* $\varepsilon = 10^{-10}$ *and* $\sigma_{\text{shift}} = 10$*. Other model parameters are the same as in Figure 3.6.* **(a)** *Algebraic error in* ℓ^∞*-norm for the first 100 inverse iterations in Algorithm 6.6. The tensor linear systems are solved using the AMEN under the tolerance values given in the legend.* **(b)** *Left y-axis: Algebraic error after 100 inverse iterations. The red circles mark the exact error, and the dashed curve refer to* $y = C_1 x + C_2$ *with* $C1 = 0.003$ *and* $C2 = 2 \times 10^{-7}$*. Right y-axis: The maximum QTT ranks in the final solution* $\boldsymbol{p}_k$*.*

order inverse iteration presented in Algorithm 2. Setting the shift value $\sigma_{\text{shift}} = 0.1$, we computed the algebraic error $\boldsymbol{p}_k - \boldsymbol{p}_{\boldsymbol{h},\hat{t}}$ for each of the 100 inverse iterations. The convergences of the algebraic error in ℓ^∞-norm under difference stopping tolerances for the tensor linear solver – AMEN (Dolgov & Savostyanov, 2013a,b) are shown in Figure 3.7a. We see that all simulations initially converge with similar rates, and the convergence comes to a halt after certain number of iterations. This matches our prediction in Theorem 3.47, that for small number of iterations k, the term $C_1\rho^k$ in (3.72) dominates the algebraic error $\|e_k\|$, while for large k, the dominance is taken over by the term $\varepsilon C_2 \frac{1-\rho^k}{1-\rho}$ in (3.72) caused by the inexactness of the tensor linear solver (Remark 3.48). In Fig. 3.7b, we only consider the algebraic error $\|e_k\|_\infty$ after $k = 100$ inverse iterations, and plot it against different tolerance in solving the tensor linear systems. The algebraic error agrees well with the reference line $y = C_1 x + C_2$ in the blue dashed curve, which is the direct result from (3.72) in Theorem 3.47 for fixed ρ and k. We could also observe the increase in the tensor rank as the tolerance

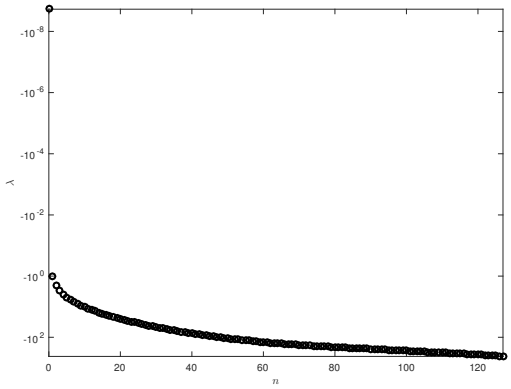

Figure 3.8: *The eigenvalues of the FDM operator associated to the CFPE of the birth-death process plotted in log scale.*

value decreases (solid curve in Fig. 3.7b), and the exponential scaling matches our prediction in Remark 3.45.

Application of the inverse power method (Algorithm 2) relies on the assumption that the principle eigenvalue λ_0 is well isolated from the rest of the spectrum, i.e., there exists only one stationary distribution to the CFPE. In Figure 3.8 we numerically validate the assumption via plotting the eigenvalues of the FDM matrix of the CFPE operator of the death-birth process, which shows a gap of approximately 1 between the principle and the second eigenvalues.

A 50-dimensional example. In this section, the simple birth-death process in (2.29) is extended to a 50-dimensional reaction chain, where molecules are allowed to transform themselves into different isometric forms, i.e.,

$$\emptyset \underset{k_2}{\overset{k_1}{\rightleftarrows}} S_1 \underset{k_4}{\overset{k_3}{\rightleftarrows}} S_2 \underset{k_6}{\overset{k_5}{\rightleftarrows}} \cdots \underset{k_{98}}{\overset{k_{97}}{\rightleftarrows}} S_{49} \underset{k_{100}}{\overset{k_{99}}{\rightleftarrows}} S_{50} \underset{k_{102}}{\overset{k_{101}}{\rightleftarrows}} \emptyset, \tag{3.73}$$

where S_i, $i = 1,2,\ldots,50$, can be interpreted as the i-th isometric form of species S, and k_j, $j = 1,2,\ldots,102$, are reaction constants. The corresponding stoichiometric

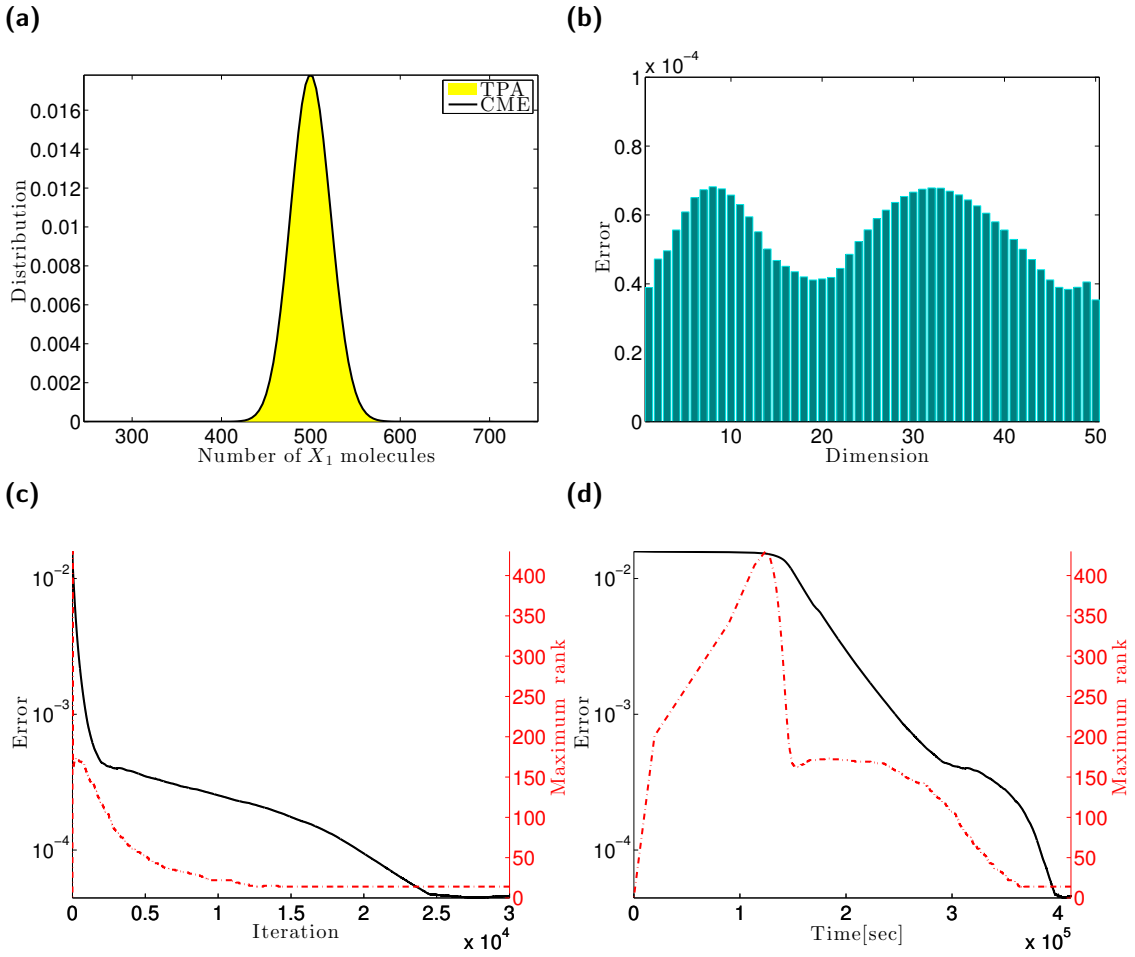

Figure 3.9: *Steady state simulation results of the* 50*-dimensional isomerization reaction chain* (3.99) *in QTT format. Parameters: reaction rates* $k_j = 1$ *for* $j \in [1,102]$*, volume* $V = 500$*, domain* $\Omega = [246,754]^{50}$*, grid size* $h = 4$*, tensor rounding tolerance* $\varepsilon_t = 10^{-10}$*, shift value* $\sigma = 40$*, and tolerance for tensor linear solver* $\varepsilon_k = 10^{-3}$*. Storage requirement of tensor matrix* $\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$ *is* 176144 *with maximum QTT rank* $R_{\max} = 13$*, and the final tensor solution* $\boldsymbol{p}_k$ *is* 61848 *with* $R_{\max} = 14$*.* **(a)** *The computed marginal distribution for* S_1 *species versus the Poisson distribution* π^1_{m} *in* (3.74) *as the exact solution of the CME.* **(b)** *Error in* ℓ^∞*-norm of the marginal distributions in all 50 dimensions.* **(c)** *Error convergence and maximum QTT rank against the number of inverse iterations. Error is measured in* ℓ^∞*-norm between the marginal distribution of* S_1*.* **(d)** *Error convergence and maximum QTT rank against the computational time.*

matrix $\boldsymbol{\nu} \in \mathbb{Z}^{102\times 50}$ is given by

$$\nu_{j,i} = \begin{cases} 1 & j = 2i-1 \ \text{ or } \ j = 2i+2, \\ -1 & j = 2i \ \text{ or } \ j = 2i+1, \\ 0 & \text{otherwise,} \end{cases} \qquad \text{for} \quad i = 1,\ldots,50 \text{ and } j = 1,\ldots,102.$$

Let $\boldsymbol{x} = [x_1, x_2, \ldots, x_{50}]^T$ be the state vector, then the propensity functions are of the form

$$\alpha_j(\boldsymbol{x}) = \begin{cases} k_j V & j = 1 \text{ or } 102, \\ k_j x_{j/2} & j \text{ is even} \\ k_j x_{(j-1)/2} & \text{otherwise,} \end{cases} \qquad \text{for} \quad j = 1, \ldots, 102.$$

Consider $\pi_{\mathrm{m}}(\boldsymbol{n})$, $\boldsymbol{n} \in \mathbb{N}^{50}$, be the solution of the stationary CME (2.5) for the isometrization reaction (3.99), it is a product Poisson distribution (Jahnke & Huisinga, 2007),

$$\pi_{\mathrm{m}}(\boldsymbol{n}) = \prod_{i=1}^{50} \pi_{\mathrm{m}}^i(n_i) = \prod_{i=1}^{50} \frac{\phi_i^{n_i}}{n_i!} e^{-\phi_i}, \tag{3.74}$$

where π_{m}^i, $i = 1, \ldots, 50$, stand for univariate Poisson functions whose mean ϕ_i equal to the stead state solution of a system of reaction rate equations.

In comparison, the analytical formula for the solution of the stationary CFPE (2.13) for the isomerization reactions (3.99) is not obvious. Furthermore, because of the very high dimensionality, we would not be able to identify individual sources of errors separately as we did in Sections § 2.2.3, § 3.1.3, § 3.3.2.1, or the first part of Section § 3.4.2.2. This is to say that we would only be able to compute the final tensor approximation $\boldsymbol{p}_k$, whose accuracy depends on the accuracies of all intermediate approximations. Whereas there will be no intermediate solutions available to guide us to choose appropriate parameters in each step. To address this issue, we notice that the isomerization reaction chain (3.99) could be viewed as an extension of the birth death process (2.29), and thus our previous error analysis for the death-birth process could potentially hint the choice of parameters for simulating the extended reaction chain (3.99).

Let us start with a target that we wish the final tensor solution $\boldsymbol{p}_k$ to be accurate to order 10^{-4} in approximating the exact Poisson distribution (3.74). This means that

we should pick simulation parameters such that all sources of errors are kept below 10^{-4}. We choose the system volume $V = 500$ to be consistent with the simulations of the 1D death-birth process. From Figure 2.1c, the modelling error at $V = 500$ would be far below 10^{-4}. The computational domain is chosen as $\Omega = \prod_{i=1}^{50} \Omega_i = [246, 754]^{50}$, and accordingly to Figure 3.2a, the domain size $|\Omega_i| = 508$ keeps the artificial boundary error below 10^{-4}. We choose the equidistant grid size $h = 4$ for all 50 dimensions, to keep the discretization error below 10^{-4} as suggested by Figure 3.2b. We choose the tolerance of tensor rounding procedure to be 10^{-10} such that the tensor rounding error in Figure 3.6 not significant. The tolerance for the tensor linear solver, AMEN, in the inverse iterations is chosen to be 10^{-3} to keep the algebraic error in Figure 3.7 below 10^{-4}.

The simulation results and performances are demonstrated in Figure 3.9. The computed tensor data agrees well with the exact Poisson distribution (see Figure 3.9a), and the accuracy meets the pre-set target 10^{-4} in all 50 dimensions (see Figure 3.9b). The error convergence in Figure 3.9c matches the prediction of Theorem 3.47 that the error first decreases monotonically and then the convergence come to a halt due to the inexactness in solving the tensor linear systems. But we also observe the maximum QTT rank increases dramatically in the initial inverse iterations. Larger tensor ranks give rise to quadratic complexity of basic arithmetic, and as a consequence, the first few iterations with large QTT ranks cost almost half of the computational time, while their effect in error convergence is not obvious at all (see Figure 3.9d). This refers to a major problem in the low-rank tensor computations that, even if the final solution admits low-rank approximations, the growth in the ranks of the intermediate solutions might still kills the simulation. This still remains as an open problem in this area.

Our 50-dimensional solution in QTT format has storage requirement to be 61848, whereas this number would be 2.3×10^{105} if stored in a standard vector. This demon-

strate the effectiveness of the tensor approach for analysing high-dimensional stochastic models of GRNs. It is also worth pointing out that the total computational time is as long as 4×10^5 seconds using a personal laptop. Such heavy simulation highlights the crucial role of the error analysis as we have presented in this thesis, because such understanding not only has informed us about the accuracy of the method, but has also guided us with feasible choices of simulations parameters such that unnecessary repetitions could be avoided.

3.4.3 The adaptive inverse power iterations

Previously, we have assumed σ_{shift} is chosen such that the tensor linear solver converges for system (3.69). In the current section we discuss the criteria for the convergence the AMEn method, and then propose an adaptive scheme that automatically choose suitable σ_{shift} value for convergence.

3.4.3.1 Criteria for the global convergence of the inverse iteration

The symmetric positive definite system. We start with reviewing the AMEn method, which was originally derived for solving the QTT-structured symmetric positive definite (SPD) system:

$$\boldsymbol{M}\boldsymbol{x} = \boldsymbol{b}, \tag{3.75}$$

where $\boldsymbol{M} \in \mathbb{R}^{n^N \times n^N}$ is symmetric positive definite and $\boldsymbol{b} \in \mathbb{R}^{n^N}$. Alternative numerical schemes for solving (3.75) tailored the (Q)TT format includes TT-GMRES (Dolgov, 2013), alternating least squares (ALS) (Holtz et al., 2012), modified alternating least squares (MALS) (Holtz et al., 2012), density matrix normalisation group (DMRG) (Oseledets & Dolgov, 2012), and alternating minimal energy (AMEn), and we refer the reader to a review given by Khoromskij (2012).

Algorithm 3 Alternating minimal energy (AMEn) method (Dolgov & Savostyanov, 2013b, Algorithm 1).

Require: Linear system $\boldsymbol{M}\boldsymbol{x} = \boldsymbol{b}$ in the QTT-format, and an initial guess $\boldsymbol{x}^{(0)} = \boldsymbol{x}_1 \bowtie \ldots \bowtie \boldsymbol{x}_N$.
1: **for** $k = 1, 2, \ldots, N$ **do**
2: Compute $\boldsymbol{v}_k = \arg\min_{\boldsymbol{x}_k} J\left(\boldsymbol{y}_1 \bowtie \ldots \bowtie \boldsymbol{y}_{k-1} \bowtie \boldsymbol{x}_k \bowtie \boldsymbol{x}_{k+1} \bowtie \ldots \bowtie \boldsymbol{x}_N\right)$.
3: Approximate the residue $\boldsymbol{z}^{(k)} = \boldsymbol{b}_{\geq k} - \boldsymbol{M}_{\geq k}\boldsymbol{u}^{(k)}$, where $\boldsymbol{u}^{(k)} = \boldsymbol{v}_k \bowtie x_{k+1} \bowtie \ldots \bowtie \boldsymbol{x}_N$, by $\tilde{\boldsymbol{z}}^{(k)} = \tilde{\boldsymbol{z}}_k \bowtie \ldots \bowtie \tilde{\boldsymbol{z}}_N$.
4: Expand the basis $\boldsymbol{y}_k = [\boldsymbol{v}_k \;\; \boldsymbol{z}_k]$, and $\boldsymbol{x}_{k+1} = \left[\begin{smallmatrix} \boldsymbol{x}_{k+1} \\ \boldsymbol{0} \end{smallmatrix}\right]$.
5: Recover the TT-orthogonality of $\boldsymbol{x}^{(k+1)} = \boldsymbol{y}_1 \bowtie \ldots \bowtie \boldsymbol{y}_k \bowtie \boldsymbol{x}_{k+1} \bowtie \ldots \bowtie \boldsymbol{x}_N$.
6: **end for**
7: **return** $\boldsymbol{x}^{(N)}$

The solution of (3.75) can be sought by solving the optimisation problem

$$\boldsymbol{x} = \arg \min_{\boldsymbol{y} \in \mathbb{R}^{n^N}} J_{\text{energy}}(\boldsymbol{y}), \tag{3.76}$$

with the objective (energy) function $J(\boldsymbol{y})$ defined by

$$J_{\text{energy}}(\boldsymbol{y}) = \|\hat{\boldsymbol{x}} - \boldsymbol{y}\|_{\boldsymbol{M}} = (\boldsymbol{y}, \boldsymbol{M}\boldsymbol{y}) - 2(\boldsymbol{y}, \boldsymbol{b}) + const, \tag{3.77}$$

where $\hat{\boldsymbol{x}} = \boldsymbol{M}^{-1}\boldsymbol{b}$, and $\|\boldsymbol{x}\|_{\boldsymbol{M}} = (\boldsymbol{x}, \boldsymbol{M}\boldsymbol{x})$ denotes the $\boldsymbol{M}$-norm of a vector $\boldsymbol{x}$, and $(\cdot, \cdot)$ refers to the inner product. The minimisation problem (3.76) is not tractable if the dimensionality is large. The multi-dimensional problem (3.76) is addressed in the AMEn by performing a greedy-type approach, where the optimisation is taken over one dimension at a time which the rest of the dimensions fixed.

The AMEn as described in Algorithm 3 proceeds as follows. Given an initial guess $\boldsymbol{x}^{(0)} = \tau(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ represented in the TT format, the algorithm recursively for $k = 1, 2, \ldots, N$ considers solving the following sub-linear system:

$$\boldsymbol{M}_{\geq k}\boldsymbol{x}_{\geq k} = \boldsymbol{b}_{\geq k}, \tag{3.78}$$

where

$$\boldsymbol{M}_{\geq k} = \left(\boldsymbol{X}_{<k}^{(k)}\right)^T \boldsymbol{M} \boldsymbol{X}_{<k}^{(k)}, \quad \boldsymbol{b}_{\geq k} = \left(\boldsymbol{X}_{<k}^{(k)}\right)^T \boldsymbol{b}, \quad \boldsymbol{X}_{<k}^{(k)} = \boldsymbol{x}_{<k}^{(k)} \otimes \boldsymbol{I}_{\geq k},$$
$$\boldsymbol{x}^{(k)} = \boldsymbol{y}_1 \bowtie \ldots \bowtie \boldsymbol{y}_{k-1} \bowtie \boldsymbol{x}_k \bowtie \ldots \bowtie \boldsymbol{x}_N, \quad \text{and} \quad \boldsymbol{x}_{<k}^{(k)} = \boldsymbol{y}_1 \bowtie \ldots \bowtie \boldsymbol{y}_{k-1}.$$

Note that $\boldsymbol{M}_{\geq k}$ has the same TT-ranks as $\boldsymbol{M}$, but the length is shorter by $(k-1)$ TT cores, i.e., (3.78) is a $(N-k+1)$-dimensional problem. The algorithm optimises J_{energy} in (3.77) over the k-th TT core $\boldsymbol{x}_k$ of $\boldsymbol{x}^{(k)}$ and obtain the minimiser $\boldsymbol{v}_k$. Defining $\boldsymbol{u}^{(k)} = \boldsymbol{v}_k \bowtie \boldsymbol{x}_{k+1} \bowtie \ldots \bowtie \boldsymbol{x}_N$, the residue $\boldsymbol{z}^{(k)} = \boldsymbol{b}_{\geq k} - \boldsymbol{M}_{\geq k} \boldsymbol{u}^{(k)}$ is approximated[‡] by $\tilde{\boldsymbol{z}}^{(k)} = \boldsymbol{z}_k \bowtie \ldots \bowtie \boldsymbol{z}_N$, and the first core is updated as $\boldsymbol{y}_k = [\boldsymbol{v}_k, \boldsymbol{z}_k]$. The procedure continues to cycle over all TT-cores, and the entire sweep repeats until a convergence criteria is satisfied. The convergence rate of one sweep is given by the following proposition.

Proposition 3.49 (Theorem 2 in Dolgov & Savostyanov (2013b))**.** *If* $\boldsymbol{M}$ *in* (3.75) *is symmetric positive definite, the convergence rate of the AMEn method described in Algorithm 3 is given by*

$$\frac{\left\|\boldsymbol{x}^{(N)} - \hat{\boldsymbol{x}}\right\|_{\boldsymbol{M}}^2}{\left\|\boldsymbol{x}^{(0)} - \hat{\boldsymbol{x}}\right\|_{\boldsymbol{M}}^2} = \sum_{r=1}^{N-1} \omega_r^2 \prod_{j=1}^{r-1} \left(1 - \omega_j^2\right) \prod_{j=1}^{r} \mu_j^2 = \phi_N^2, \tag{3.79}$$

where

$$\mu_j^2 = \frac{\left\|\hat{\boldsymbol{x}}_{\geq j} - \boldsymbol{u}^{(j)}\right\|_{\boldsymbol{M}_{\geq j}}^2}{\left\|\hat{\boldsymbol{x}}_{\geq j} - \boldsymbol{x}_{\geq j}^{(0)}\right\|_{\boldsymbol{M}_{\geq j}}^2} \quad \textit{and} \quad \omega_j^2 = \frac{\left\|\hat{\boldsymbol{x}}_{\geq j} - \boldsymbol{x}_{\geq j}^{(j)}\right\|_{\boldsymbol{M}_{\geq j}}^2}{\left\|\hat{\boldsymbol{x}}_{\geq j} - \boldsymbol{u}^{(j)}\right\|_{\boldsymbol{M}_{\geq j}}^2}. \tag{3.80}$$

The global convergence can be achieved conditioned on $\phi_N < 1$, and this is guaranteed by making $\tilde{\boldsymbol{z}}^{(k)}$ sufficiently close to $\boldsymbol{z}^{(k)}$, see Remark 2 in Dolgov & Savostyanov

[‡]The approximation could be performed by several different strategies, including the SVD-based approximation, Cholesky-based approximation, and the ALS-based approximation. These methods have been hard-coded in the TT-toolbox, and We refer the readers to Dolgov & Savostyanov (2013b) for the details.

(2013b).

The positive definite linear system. In the inverse iteration (Algorithm 2), however, the matrix $\boldsymbol{M}$ is given by

$$\boldsymbol{M} = -\left(\boldsymbol{A}_{\boldsymbol{h},\hat{t}} - \sigma_{\text{shift}}\boldsymbol{I}\right), \tag{3.81}$$

and $\boldsymbol{M}$ is not symmetric while still positive definite (PD). Consequently, the previous analysis and the global convergence in Proposition 3.49 do not apply in the PD system. Instead, the authors in Dolgov & Savostyanov (2013b) related the AMEn to the full orthogonalization method and provided the following estimate.

Proposition 3.50. *Consider* $\boldsymbol{M}$ *in* (3.75) *as a positive definite matrix. The convergence of the AMEn method in Algorithm 3 in solving the tensor linear system* (3.75) *is given by*

$$\frac{\left\|\boldsymbol{b} - \boldsymbol{M}\boldsymbol{x}^{(N)}\right\|}{\left\|\boldsymbol{b} - \boldsymbol{M}\boldsymbol{x}^{(0)}\right\|} = \sum_{k=1}^{N-1} \omega_k \mu_k \prod_{m=1}^{k-1} \frac{\mu_m}{\sqrt{1-\omega_m^2}} = \phi_N, \tag{3.82}$$

and the coefficients are defined as

$$\mu_k = \frac{\left\|\boldsymbol{b}_{\geq k} - \boldsymbol{M}_{\geq k}\boldsymbol{u}^{(k)}\right\|}{\left\|\boldsymbol{b}_{\geq k} - \boldsymbol{M}_{\geq k}\boldsymbol{x}^{(0)}_{\geq k}\right\|} \quad \textit{and} \quad \omega_k = \frac{\left\|\boldsymbol{b}_{\geq k} - \boldsymbol{M}_{\geq k}\boldsymbol{x}^{(j)}_{\geq j}\right\|}{\left\|\boldsymbol{b}_{\geq k} - \boldsymbol{M}_{\geq k}\boldsymbol{u}^{(j)}\right\|} \leq \omega_{\boldsymbol{V}_k},$$

where

$$\omega_{\boldsymbol{V}_k} = 1 - \frac{\lambda_{\min}\left(\boldsymbol{V}_k^T\boldsymbol{M}_{\geq k}\boldsymbol{V}_k + \boldsymbol{V}_k^T\boldsymbol{M}_{\geq k}^T\boldsymbol{V}_k\right)}{2\left\|\boldsymbol{M}_{\geq k}\boldsymbol{V}_k\right\|}, \quad \textit{and} \quad \boldsymbol{V}_k = \boldsymbol{y}_k \otimes \boldsymbol{I}_{n_{k+1}\cdots n_N}. \tag{3.83}$$

The proof of Lemma 3.50 combines Lemmas 1 and 2 in Dolgov & Savostyanov

(2013b). Given that

$$\mu_j \leq 1, \quad \text{and} \quad \omega_k < 1, \quad k = 1, 2, \ldots, N-1,$$

the previous described SPD case (Proposition 3.49) converges globally in any case, whereas the convergence rate ϕ_N for the PD system in Proposition 3.50 can be greater than 1 in certain situations. In what follows, we derive a sufficient condition (a well-conditioned system) for the AMEn algorithm attain the global convergence for the PD linear systems.

Lemma 3.51. *Define* $\bar{\omega} \in (0,1)$ *to be*

$$\bar{\omega} := \frac{\lambda_{\max}(\boldsymbol{M}) - \lambda_{\min}(\boldsymbol{M})}{\lambda_{\max}(\boldsymbol{M})}, \tag{3.84}$$

and define the threshold value $\hat{\omega} \in (0,1)$ *such that*

$$f_\omega(\hat{\omega}) = 0, \tag{3.85}$$

where

$$f_\omega(\omega) = \omega^{2N-2} + (N-1)^2\omega^2 - (N-1)^2 = 0.$$

Then, a sufficient condition for the global convergence rate ϕ_N *in* (3.82) *to be strictly smaller than 1 is*

$$\bar{\omega} < \sqrt{1 - \hat{\omega}^2}. \tag{3.86}$$

Proof. We first bound the convergence rate ϕ_N in terms of $\omega_{\boldsymbol{V}_k}$ as

$$\phi_N \leq \sum_{k=1}^{N} \omega_{\boldsymbol{V}_k} \mu_k \prod_{m=1}^{k-1} \frac{\mu_m}{\sqrt{1-\omega_{\boldsymbol{V}_m}^2}} \leq \sum_{k=1}^{N} \omega_{\boldsymbol{V}_k} \prod_{m=1}^{k-1} \frac{1}{\sqrt{1-\omega_{\boldsymbol{V}_m}^2}}, \tag{3.87}$$

where we have use the fact that the optimisation step provides the residual decrease, i.e., $\mu_k \leq 1$. It can be shown that if all $\boldsymbol{x}_{<k}$ are orthogonal, we have (Zhang, 2011)

$$\begin{aligned}
&\frac{1}{2}\lambda_{\min}\left(\boldsymbol{V}_k^T \boldsymbol{M}_{\geq k} \boldsymbol{V}_k + \boldsymbol{V}_k^T \boldsymbol{M}_{\geq k}^T \boldsymbol{V}_k\right) \geq \lambda_{\min}\left(\boldsymbol{V}_k^T \boldsymbol{M}_{\geq k} \boldsymbol{V}_k\right) \geq \lambda_{\min}(\boldsymbol{M}_{\geq k}) \geq \lambda_{\min}(\boldsymbol{M}), \\
&\|\boldsymbol{M}_{\geq k} \boldsymbol{V}_k\| = \lambda_{\max}^{\frac{1}{2}}\left(\boldsymbol{V}_k^T \boldsymbol{M}_{\geq k}^T \boldsymbol{M}_{\geq k} \boldsymbol{V}_k\right) \leq \lambda_{\max}^{\frac{1}{2}}\left(\boldsymbol{M}_{\geq k}^T \boldsymbol{M}_{\geq k}\right) \leq \lambda_{\max}(\boldsymbol{M}_{\geq k}) \leq \lambda_{\max}(\boldsymbol{M}).
\end{aligned}$$

Substituting the above inequalities into (3.83) yields

$$\omega_{\boldsymbol{V}_k} \leq \bar{\omega} = \frac{\lambda_{\max}(\boldsymbol{M}) - \lambda_{\min}(\boldsymbol{M})}{\lambda_{\max}(\boldsymbol{M})} < 1.$$

and we can further bound (3.87) by

$$\phi_N \leq \sum_{k=1}^{N} \bar{\omega} \prod_{m=1}^{k-1} \frac{1}{\sqrt{1-\bar{\omega}^2}} \leq (N-1)\frac{\bar{\omega}}{\left(1-\bar{\omega}^2\right)^{(N-1)/2}}. \tag{3.88}$$

It is easy to verify that, let $f(\omega) = \omega^{2N-2} + (N-1)^2\omega^2 - (N-1)^2$, function $f(\cdot)$ is monotonically increasing within the interval $\omega \in (0,1)$, and there exists one roots $\hat{\omega}$ such that $f(\hat{\omega}) = 0$. Then, substituting the condition (3.85) into above formula, one can directly derive $\phi_N < 1$. □

Remark 3.52. Condition (3.86) in Lemma 3.51 is equivalent to requiring the tensor matrix $\boldsymbol{M}$ to be sufficiently well-conditioned. Let $\kappa(\boldsymbol{M})$ be the conditioning number of $\boldsymbol{M}$ defined as

$$\kappa^2(\boldsymbol{M}) = \frac{\lambda_{\max}(\boldsymbol{M})}{\lambda_{\min}(\boldsymbol{M})},$$

and one can rewrite condition (3.86) as

$$\kappa(\boldsymbol{M}) < \frac{1}{\sqrt{1-\sqrt{1-\hat{\omega}^2}}}, \tag{3.89}$$

where $\hat{\omega} \in (0,1)$ satisfies (3.85).

In our case, the PD matrix $\boldsymbol{M}$ is given in (3.81), and the corresponding extreme eigenvalues are

$$\lambda_{\min}(\boldsymbol{M}) = \sigma_{\text{shift}} + \lambda_{\min}\left(-\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\right), \quad \text{and} \quad \lambda_{\max}(\boldsymbol{M}) = \sigma_{\text{shift}} + \lambda_{\max}\left(-\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\right), \tag{3.90}$$

Hence, given a threshold value $\hat{\omega}$, one can always make the underlying conditioning number,

$$\kappa(\boldsymbol{M}) = \frac{\sqrt{\sigma_{\text{shift}} + \lambda_{\max}\left(-\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\right)}}{\sqrt{\sigma_{\text{shift}} + \lambda_{\min}\left(-\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\right)}}, \tag{3.91}$$

sufficiently small by choosing a sufficiently large σ_{shift} value such that condition (3.86) is met. We summarise this result in the following theorem.

Theorem 3.53. *Consider solving the inverse iteration* (3.69) *using the AMEn method (Algorithm 2). There exists a threshold value* $\hat{\sigma}_{\text{shift}} < \infty$ *such that* $\forall \sigma_{\text{shift}} > \hat{\sigma}_{\text{shift}}$ *we have* $\phi_N^2 < 1$*, where* ϕ_N *is defined in* (3.82)*, i.e., the AMEn converges globally.*

Proof. To show there exist a threshold value, we substitute (3.91) into 3.89 and obtain

$$\sigma_{\text{shift}} > \hat{\sigma}_{\text{shift}} = \frac{\lambda_{\max}\left(-\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\right)\sqrt{1-\sqrt{1-\hat{\omega}^2}} - \lambda_{\min}\left(\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\right)}{1-\sqrt{1-\sqrt{1-\hat{\omega}^2}}},$$

Then, the argument directly follows Lemma 3.51. □

Remark 3.54. We see from (3.88) that the convergence rate ϕ_N decreases monotonically with decreasing value of $\bar{\omega}$ defined in (3.84). Substituting (3.90) into (3.84) yields

$$\bar{\omega} = \frac{\lambda_{\max}\left(-\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\right) - \lambda_{\min}\left(-\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\right)}{\lambda_{\max}\left(-\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\right) + \sigma_{\text{shift}}},$$

which suggests the monotonic decrease of $\bar{\omega}$ with respect to the increasing value of σ_{shift}. Hence one could expect better convergence rate of the AMEn algorithm by increasing σ_{shift} value.

Remark 3.55. Another implication of (3.88) is that, if the number of dimensions N increases, the AMEn algorithm converges slower for a fixed σ_{shift} value.

3.4.3.2 An adaptive scheme

Comparing the results in Theorems 3.47 and 3.53, we find that there is a trade-off between the efficiencies of the outer iterations and the inner iterations of the inverse power method (Algorithm 2):

- the outer iterations converge faster for smaller σ_{shift} values, and may not converge if σ_{shift} is too large;
- the inner iterations converge faster for larger σ_{shift} values, and may not converge if σ_{shift} is too small.

In Algorithm 2, σ_{shift} value needs to be predetermined and treated as a constant throughout the execution, and as a result, the efficiency of the simulation can be significantly decreased and uncontrollable, or even the algorithm does not converge.

It is possible to derive an upper bound for $\hat{\sigma}_{\text{shift}}$ in Theorem 3.53. Consider imposing a constraint on the AMEn convergence rate $\phi_N \leq \hat{\phi}_N$ for a given threshold value $\hat{\phi}_N < 1$. One could reversely work out an indicator $\bar{\sigma}_{\text{shift}}$ of the threshold value $\bar{\sigma}_{\text{shift}}$

Algorithm 4 Adaptive inverse power method in tensor format.

Require: initial guess $\boldsymbol{p}_0$; shift value $\sigma_{\text{shift}} = \sigma_0$; a stopping criterion $\varepsilon_{\text{stop}}$ and a convergence threshold $\varepsilon_{\text{AMEn}}$; maximum number of AMEN sweeps in each inverse iteration $N_{\max}$; thresholds to increase (N_{in}) and decrease (N_{de}) the shift value, set $k = 1$.

while $\|\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\boldsymbol{p}_{k-1}\|/\|\boldsymbol{p}_{k-1}\| > \varepsilon_{\text{stop}}$ **do**
  Solve the k-th tensor-structured inverse iteration (3.69) up to $N_{\max}$ sweeps.
  Check the number of sweeps, N_{comp}, for the AMEN solver to converge.
  if $N_{\text{comp}} > N_{\text{in}}$ **then**
    $\sigma_{\text{shift}} \leftarrow 2\sigma_{\text{shift}}$.
  else if $N_{\text{de}} < N_{\text{comp}} \le N_{\text{in}}$ **then**
    $\sigma_{\text{shift}} \leftarrow \sigma_{\text{shift}}/2$, $k \leftarrow k+1$.
  else
    $k \leftarrow k+1$.
  end if
end while

using (3.84) and (3.88), and obtain

$$\bar{\sigma}_{\text{shift}} = \frac{\lambda_{\max}\left(-\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\right) - \lambda_{\min}\left(-\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\right)}{\sqrt{1-\hat{w}^2_{\hat{\phi}_N}}} - \lambda_{\max}\left(-\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\right) \approx \frac{1-\sqrt{1-\hat{w}^2_{\hat{\phi}_N}}}{\sqrt{1-\hat{w}^2_{\hat{\phi}_N}}}\lambda_{\max}\left(-\boldsymbol{A}_{\boldsymbol{h},\hat{t}}\right), \tag{3.92}$$

where $\hat{w}_{\hat{\phi}_N} \in (0,1)$ is the roof of the polynomial

$$f_{\hat{w}_{\hat{\phi}_N}} = \hat{\phi}_N \hat{w}^{2N-2}_{\hat{\phi}_N} + (N+1)^2 \hat{w}^2_{\hat{\phi}_N} - (N-1)^2.$$

In practice, however, the right-hand side of (3.92) often requires additional computations to estimate the maximal eigenvalue of $\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$. Also, the indicator $\bar{\sigma}_{\text{shift}}$ may be very different from the true threshold value $\hat{\sigma}_{\text{shift}}$, especially if $\sum_{r=1}^{N-1}\prod_{j=1}^{r}\mu_j^2 \ll N-1$. Thus, instead of using a priori estimate, we make use of the posteriori results and dynamically update the σ_{shift} value in each iteration based on the number of the AMEn sweeps.

The adaptive inverse power method (Algorithm 4) proceeds as follows. Given an initial guess $\boldsymbol{p}_0$ and an initial shift value $\sigma_{\text{shift}} = \sigma_0$, the algorithm seeks to solve the tensor linear system (3.69) using the maximum number $N_{\max}$ of the AMEn sweeps (Algorithm 3). If the AMEn does not achieve the prescribed tolerance $\varepsilon_{\text{AMEn}}$ within

Algorithm 5 *The multigrid adaptive inverse power method in the (Q)TT format.*

Require: initial grid size $\boldsymbol{h} = (h_1, h_2, \ldots, h_N)^T$, an initial guess $\boldsymbol{p}_{0,1}$ and a stopping criterion ε_1 on the coarsest grid; the total number of grid levels L_{max}.

for $\ell = 1, 2, \ldots, L_{max}$ **do**
  Construct the finite difference approximation $\boldsymbol{A}_{\boldsymbol{h},t,\ell}$ of the CFPE operator on grid $\omega_{\boldsymbol{h},\ell}$ in the QTT format, and apply rank truncation procedure to obtain a low-rank approximation $\boldsymbol{A}_{\boldsymbol{h},\hat{t},\ell}$;
  Apply Algorithm 4 to solve the principle eigenvalue problem with the initial guess $\boldsymbol{p}_{0,\ell}$ and the stopping criterion ε_ℓ, and obtain the approximated solution $\boldsymbol{p}_{k,\ell}$;
  if $\ell < L_{max}$ **then**
    Interpolate $\boldsymbol{p}_{k,\ell}$ to the finer grid $\omega_{\boldsymbol{h},\ell+1}$ using (3.98), and obtain $\boldsymbol{p}_{0,\ell+1}$;
    Let $\varepsilon_{\ell+1} = \varepsilon_\ell / 2$;
  end if
end for
return $\boldsymbol{p}_{k,L_{\max}}$

$N_{\max}$, the σ_{shift} is doubled and restart the AMEn solver. The whole process is repeated until the convergence criteria is met within $N_{\max}$ sweeps. If the tolerance $\varepsilon_{\text{AMEn}}$ is achieved with less than a prescribed number N_{de} of AMEn sweeps, the shift value σ_{shift} is then halved.

3.4.4 The multi-grid accelerated inverse power iterations

Since the convergence rate tends to be slower as the number of dimensions increases (Remark 3.55), (Q)TT-formatted acceleration techniques are necessary, and here we introduce the multi-grid (MG) acceleration technique. The MG method is often used as a pre-conditioner for solving the elliptic equations (Hackbusch, 2013), and has been applied to solve the tensor-based problems (Khoromskij & Khoromskaia, 2009; Khoromskaia, 2010; Khoromskaia et al., 2011).

Here, the main idea is to construct a hierarchy of grids, and the system (3.61) is first solved on a coarser grid, and the approximation is then interpolated to a finer grid and used as an initial guess. Let

$$\omega_{\boldsymbol{h},\ell} = \omega_{h_1,\ell} \times \omega_{h_1,\ell} \times \cdots \times \omega_{h_N,\ell} \tag{3.93}$$

be a hierarchy of grids with grid sizes $\boldsymbol{h} = (h_1, \ldots, h_N)^T$, where $\omega_{h_i,\ell}$ is a hierarchy of

grid with grid size h_i in the i-th direction. On each grid level, a finite difference approximation of the Fokker-Planck operator can be constructed following the method introduced in Sections § 3.1.2 and § 3.3, which yields a hierarchy of discrete problems:

$$\boldsymbol{A}_{\boldsymbol{h},t,\ell}\boldsymbol{p}_{\boldsymbol{h},t,\ell} = \lambda_{\boldsymbol{h},t,\ell}\boldsymbol{p}_{\boldsymbol{h},t,\ell}, \quad \boldsymbol{A}_{\boldsymbol{h},t,\ell} \in \mathbb{R}^{(n_1\times\cdots\times n_N)\times(n_1\times\cdots\times n_N)}, \quad \ell = 1,\ldots,L_{\max}, \tag{3.94}$$

where matrix $\boldsymbol{A}_{\boldsymbol{h},t,\ell}$ has the same structure as $\boldsymbol{A}_{\boldsymbol{h},t}$ in (3.55) and the additional subscription is used to refer the grid level, and $\lambda_{\boldsymbol{h},t,\ell}$ denotes the corresponding principle eigenvalue. Once (3.94) is formulated in the QTT format, matrix $\boldsymbol{A}_{\boldsymbol{h},t,\ell}$ can be approximated by $\boldsymbol{A}_{\boldsymbol{h},\hat{t},\ell}$ with lower TT-ranks (see Section § 3.3.2). Then, starting from the coarsest level $\ell = 1$, we solve the compressed problem

$$\boldsymbol{A}_{\boldsymbol{h},\hat{t},\ell}\boldsymbol{p}_{\boldsymbol{h},\hat{t},\ell} = \lambda_{\boldsymbol{h},\hat{t},\ell}\boldsymbol{p}_{\boldsymbol{h},\hat{t},\ell}, \quad \boldsymbol{A}_{\boldsymbol{h},\hat{t},\ell} \in \mathbb{R}^{(n_1\times\cdots\times n_N)\times(n_1\times\cdots\times n_N)}, \quad \ell = 1,\ldots,L_{\max}, \tag{3.95}$$

using the adaptive inverse power method (Algorithm 4).

For the grid level $\ell > 1$, the initial guess $\boldsymbol{p}_{0,\ell}$ of the adaptive inverse power method is an interpolation of the solution on the grid level $\ell - 1$, and the interpolation is performed through the so-called prolongation operator. Consider a one-dimensional problem for example, a hierarchy of grids can be constructed in a way where, on grid level $\ell+1$, extra mid-points are allocated between the nodes on grid level ℓ. Thus, the grid size is halved by going up a level. The prolongation operator for such purpose is

given by the matrix $\boldsymbol{P}_n^{2n} \in \mathbb{R}^{2n\times n}$ defined as

$$\boldsymbol{P}_n^{2n} = \frac{1}{2}\begin{pmatrix} 2 & & & & & \\ 1 & 1 & & & & \\ & 2 & & & & \\ & 1 & 1 & & & \\ & & \ddots & & & \\ & & & 1 & & \\ & & & 2 & & \\ & & & 1 & 1 & \\ & & & & 2 \end{pmatrix}, \tag{3.96}$$

and the value on each new mid-points is the average of its surrounding nodal values, while the values on the existing nodes remain the same. Let $n = 2^l$, the prolongation matrix $\boldsymbol{P}_n^{2n}$ admits explicit QTT representation of rank-$(2,\ldots,2)$ given by (Kazeev & Khoromskij, 2012)

$$\boldsymbol{P}_n^{2n} = \frac{1}{2}\begin{bmatrix} \Psi_I & \Psi'_J \end{bmatrix} \bowtie \begin{bmatrix} \Psi_I & \Psi'_J \\ & \Psi_J \end{bmatrix}^{\bowtie l-2} \bowtie \begin{bmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \\ \begin{pmatrix} 1 \\ 0 \end{pmatrix} \end{bmatrix},$$

where the 2×2 matrices Ψ_I and Ψ_J are defined in (3.43).

Extension of the one-dimensional prolongation matrix (3.96) to the multi-dimensional case is straightforward. Due to the product structure of grid $\omega_{\boldsymbol{h},\ell}$ in (3.93), the prolongation matrix to N-dimensions has the rank-one form

$$\boldsymbol{P}_{\boldsymbol{n}}^{2\boldsymbol{n}} = \boldsymbol{P}_{n_1}^{2n_1} \otimes \boldsymbol{P}_{n_2}^{2n_2} \otimes \cdots \otimes \underbrace{\boldsymbol{P}_{n_d}^{2n_d}}_{d\text{-th mode}} \otimes \cdots \otimes \boldsymbol{P}_{n_N}^{2n_N}, \tag{3.97}$$

where n_i, $1 \le i \le N$, refers to the number of nodal points in the i-th direction on the ℓ-th grid level. Then, the interpolation is performed through

$$\boldsymbol{p}_{0,\ell} = \boldsymbol{P}_{\boldsymbol{n}}^{2\boldsymbol{n}} \boldsymbol{p}_{t,\ell-1}, \quad \ell = 2,3,\ldots,L_{\max}. \tag{3.98}$$

The resulting tensor $\boldsymbol{p}_{0,\ell}$ is of size $(2n_1 \times \cdots \times 2n_N) \times (2n_1 \times \cdots \times 2n_N)$ with TT ranks (see property (v) of Proposition 3.23):

$$\left(2R_{1,1}^{\boldsymbol{p}_{t,\ell-1}},\ldots,2R_{1,l_1}^{\boldsymbol{p}_{t,\ell-1}}\right), R_1^{\boldsymbol{p}_{t,\ell-1}},\ldots,R_{N-1}^{\boldsymbol{p}_{t,\ell-1}}, \left(2R_{N,1}^{\boldsymbol{p}_{t,\ell-1}},\ldots,2R_{N,l_N}^{\boldsymbol{p}_{t,\ell-1}}\right),$$

where $R_{d,v_d}^{\boldsymbol{p}_{t,\ell-1}}$, $1 \le v_d \le l_d$, $1 \le d \le N$, are QTT ranks of $\boldsymbol{p}_{t,\ell-1}$. The complexity of performing (3.97) in the QTT format is of order $\mathcal{O}(lNR^2)$, and we summarise the multi-grid accelerated adaptive inverse power method in Algorithm 5.

3.4.4.1 Numerical experiment

We next simulate a reaction chain of 20 molecular species interacting through 42 reversible isomerization reactions:

$$\emptyset \underset{k_2}{\overset{k_1}{\rightleftarrows}} S_1 \underset{k_4}{\overset{k_3}{\rightleftarrows}} S_2 \underset{k_6}{\overset{k_5}{\rightleftarrows}} \cdots \underset{k_{38}}{\overset{k_{37}}{\rightleftarrows}} S_{19} \underset{k_{40}}{\overset{k_{39}}{\rightleftarrows}} X_{20} \underset{k_{42}}{\overset{k_{41}}{\rightleftarrows}} \emptyset, \tag{3.99}$$

where S_i, $i = 1,2,\ldots,20$, represent the i-th isometric form of species S, and reaction constants $k_j = 1$, $j = 1,2,\ldots,42$. The system volume is set to $V = 120$. The steady state distribution of the CME for system (3.99) is a product Poisson distribution as

$$\pi_{\mathrm{m}}(\boldsymbol{n}) = \prod_{i=1}^{20} \pi_{\mathrm{m}}^i(n_i) = \prod_{i=1}^{20} \frac{\phi_i^{n_i}}{n_i!} e^{-\phi_i}, \tag{3.100}$$

where π_{m}^i denotes the i-th single-dimensional Poisson distribution whose mean value ϕ_i equals to its deterministic mean value.

Instead of solving the corresponding Fokker-Planck equation over the finest grid,

we apply the multi-level acceleration (Algorithm 5) with adaptive shift values (Algorithm 4). The grid level details are listed in Table 3.1 and the simulation results are plotted in Figure 3.10. Comparing Figures 3.9d and 3.10c, the convergence rate at the initial stage of the simulation is sufficiently improved by applying the coarse grid approximation.

In general, the operators of both the CME and CFPE are non-symmetric and ill-conditioned, and challenging to handle using the existing QTT-structured linear solvers (see Remark 3.52). Although it can be improved by increase the shift value in the inverse iterations (Remark 3.54), the CFPE has a distinctive advantage over the CME for its flexibility in choosing the grid size, which enables to control the accuracy and use of acceleration strategies, such as the presented multi-level approach.

3.5 Tensor decomposition of the parametric CFPE

In practical applications, simulating the stationary distributions of stochastic networks often needs to be performed over ranges of parameter values, especially in parametric analysis (see Chapter 4). This requires solving the parametrised stochastic model equations (CME/CFPE), and in this section we extend the previous introduced tensor formalism to solve the parametric stationary CFPEs.

Table 3.1: *Multi-level discretisation for the* 20*-dimensional reaction chain* (3.99).

Level	1	2	3	4	5	6	7
No. of nodes n	8	16	32	64	128	256	512
Grid size h	20	10	5	2.5	1.25	0.625	0.3125
CPU time† ($\times 10^3$ sec)	2.82	7.02	2.84	2.3	2.25	0.08	0.10

† The computational time that the simulation spent on each grid level. The numbers correspond to the segments of the time axis in Figure 3.10c separated by the dashed lines.

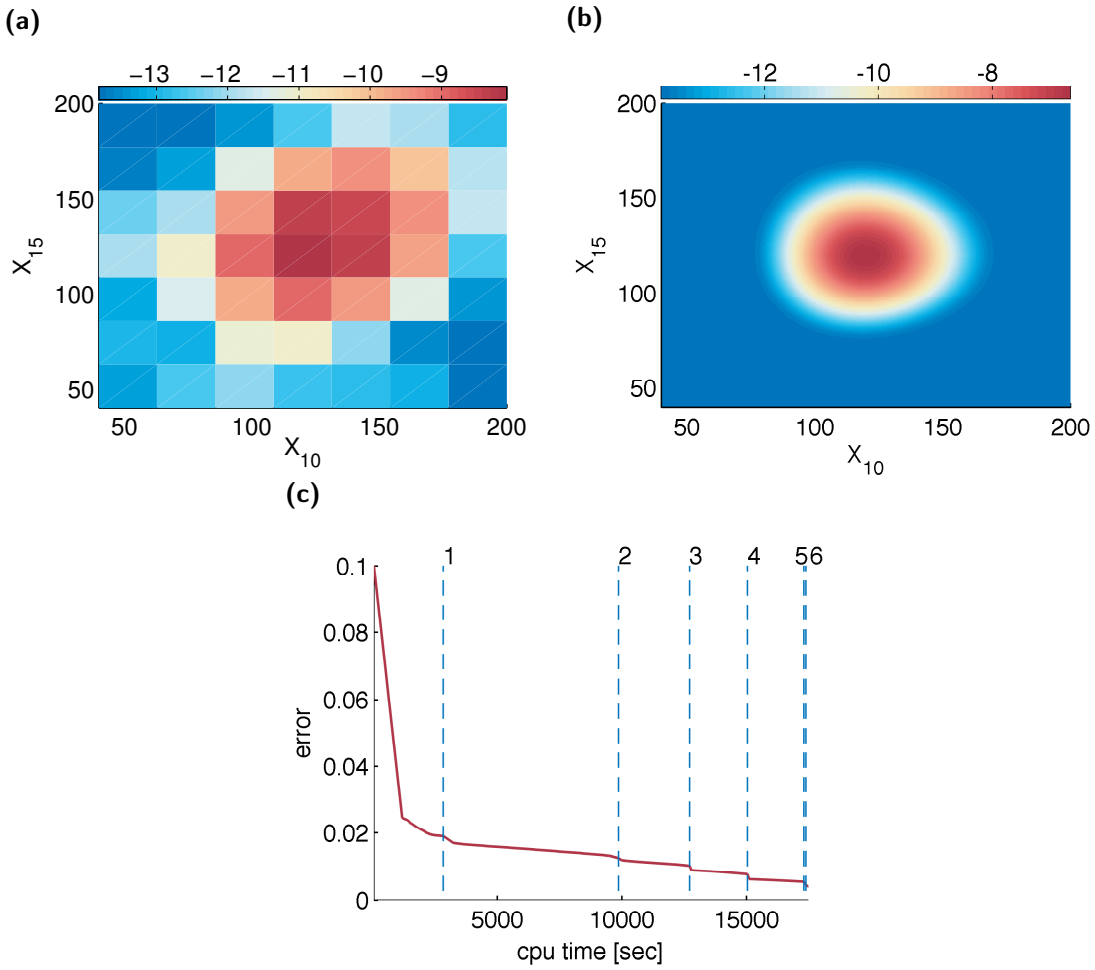

Figure 3.10: *Tensor-structured stochastic simulation of 20-dimensional reaction chain using the adaptive inverse power method with multi-level acceleration. The CFPE is successively solved on seven grid levels with an increasing number of nodal points. The logarithm of the marginal stationary distribution in the X_{10}–X_{15} plane computed on* **(a)** *the initial coarsest level; and* **(b)** *the finest grid level.* **(c)** *The convergence of the total error versus the computational time. The error is obtained by comparing the marginal distribution of the computed steady state distribution with the exact solution of the corresponding CME. The vertical dashed lines correspond to the grid levels. The grid size details on each grid level are given in Table 3.1.*

3.5.1 The parametric CFPE

We first formulate the parametric CFPE. We denote $\boldsymbol{k} = (k_1, \ldots, k_K)^T$ as K out of M reaction rate constants of the stochastic networks (2.1). We will treat rate constants $\boldsymbol{k}$ as auxiliary variables, and the remaining $M-K$ rate constants are fixed. In principle, we may also study the models where $K > M$, i.e., when we consider additional parameters, such as system volume V. Here, the parametric problem involves con-

sidering both N-dimensional state space $\boldsymbol{x} \in \Omega_{\mathrm{f}}$ and K-dimensional parameter space $\boldsymbol{k} \in \Omega_{\boldsymbol{k}}$.

We write $P_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k};t)$ be the probability distribution predicted by the CFPE that that the system will reach the state $\boldsymbol{x}$ at time t given the parameter values $\boldsymbol{k}$. Then the parametrised chemical Fokker-Planck equation can be written as

$$\begin{cases} \frac{\partial P_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k};t)}{\partial t} = \mathscr{L}_{\mathrm{f}}^{\boldsymbol{k}} P_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k};t), \quad \boldsymbol{x} \in \Omega_{\mathrm{f}}, \ \ \boldsymbol{k} \in \Omega_{\boldsymbol{k}}, \ \ t > 0 \\ P_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k};t) \geq 0, \quad \int_{\Omega_{\mathrm{f}}} P_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k};t)\mathrm{d}\boldsymbol{x} = 1, \quad \forall \boldsymbol{k} \in \Omega_{\boldsymbol{k}}, \end{cases} \tag{3.101}$$

with initial condition $P_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k};0)$. The parametric CFPE operator $\mathscr{L}_{\mathrm{f}}^{\boldsymbol{k}}$ has the similar form as the CFPE operator $\mathscr{L}_{\mathrm{f}}$ in (2.10) but with explicit representation of $\boldsymbol{k}$, written as

$$\begin{aligned} \mathscr{L}_{\mathrm{f}}^{\boldsymbol{k}} P_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k};t) = &- \sum_{i=1}^{N} \frac{\partial}{\partial x_i} \left(\sum_{j=1}^{M} \nu_{j,i} k_j \tilde{\alpha}_j(\boldsymbol{x}) P_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k};t) \right) \\ &+ \frac{1}{2} \sum_{i,i'=1}^{N} \frac{\partial^2}{\partial x_i \partial x_{i'}} \left(\sum_{j=1}^{M} \nu_{j,i} \nu_{j,i'} k_j \tilde{\alpha}_j(\boldsymbol{x}) P_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k};t) \right), \end{aligned} \tag{3.102}$$

where $\tilde{\alpha}_j$ refers to the non-parametric part of the propensity function α_j defined in (2.2), i.e., $\tilde{\alpha}_j = \alpha_j / k_j$, $j = 1, 2, \ldots, M$. If the system is observed for a sufficiently long time, the system (3.101) reaches its stationary state, and the stationary version of (3.101) is of the form

$$\begin{cases} \mathscr{L}_{\mathrm{f}}^{\boldsymbol{k}} \pi_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k}) = 0, \quad \boldsymbol{x} \in \Omega_{\mathrm{f}}, \ \ \boldsymbol{k} \in \Omega_{\boldsymbol{k}}, \\ \pi_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k}) \geq 0, \quad \int_{\Omega_{\mathrm{f}}} \pi_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k})\mathrm{d}\boldsymbol{x} = 1, \quad \forall \boldsymbol{k} \in \Omega_{\boldsymbol{k}}, \end{cases} \tag{3.103}$$

where $\pi_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k}) = \lim_{t \to \infty} P_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k};t)$.

We can rearrange Equation (3.102) and express the operator $\mathscr{L}_{\mathrm{f}}^{\boldsymbol{k}}$ as a sum of separated sub-operators in terms of state and parametric variables. We write it as

$$\mathscr{L}_{\mathrm{f}}^{\boldsymbol{k}} = k_1 \mathscr{L}_{\mathrm{f}}^{k_1} + k_2 \mathscr{L}_{\mathrm{f}}^{k_2} + \cdots + k_M \mathscr{L}_{\mathrm{f}}^{k_M}, \tag{3.104}$$

where the non-parametric operators, $\mathscr{L}_{\mathrm{f}}^{k_j}$, describes the normalised transition properties of the j-th reaction, and is defined by

$$\mathscr{L}_{\mathrm{f}}^{k_j}\pi_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k}) = -\sum_{i=1}^{N}\nu_{j,i}\frac{\partial}{\partial x_i}\left(\tilde{\alpha}_j(\boldsymbol{x})\pi_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k})\right) + \frac{1}{2}\sum_{i,i'=1}^{N}\nu_{j,i}\nu_{j,i'}\frac{\partial^2}{\partial x_i \partial x_{i'}}\left(\tilde{\alpha}_j(\boldsymbol{x})\pi_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k})\right), \tag{3.105}$$

It appears that the splitting in (3.104) is restricted to the mass-action reaction systems, where the propensity functions (2.2) has a simple first-order product in the reaction rate parameters. However, similar splitting can be obtained even for more general definitions (Wu et al., 2011). If the propensity functions depend non-linearly on the kinetic rates, as in (Barkai & Leibler, 2000) for example, the the splitting in (3.104) can still be obtained provided a multiplicative splitting (3.104) of the parametric Fokker-Planck operator is possible. Such splitting is always possible if the propensities can be written, or approximated by, as a product of two terms, where the first term depends (non-linearly) on kinetic rates, and the second term on state variables.

3.5.2 Tensor formalism

We now consider discretise the parametric operator $\mathscr{L}_{\mathrm{f}}^{\boldsymbol{k}}$ over an $(N+K)$-dimensional tensor grid, $\omega_{\boldsymbol{h}}^{\boldsymbol{x}} \times \omega_{\boldsymbol{h}}^{\boldsymbol{k}}$. The N-dimensional grid $\omega_{\boldsymbol{h}}^{\boldsymbol{x}}$ is an equidistant sampling of a bounded subspace (hypercube) within the state space Ω_{f}, and its formal definition is given in (3.6). The M-dimensional grid $\omega_{\boldsymbol{h}}^{\boldsymbol{k}}$ is discretise a K-dimensional parameter space, $\Omega_{\boldsymbol{k}} = \mathscr{I}_1^{\boldsymbol{k}} \times \cdots \times \mathscr{I}_K^{\boldsymbol{k}}$, $\mathscr{I}_d^{\boldsymbol{k}} = [a_d^{\boldsymbol{k}}, b_d^{\boldsymbol{k}}]$, $d = 1,2,\ldots,K$, and define the uniform grid $\omega_{\boldsymbol{h}}^{\boldsymbol{k}}$:

$$\omega_{\boldsymbol{h}}^{\boldsymbol{k}} = \left\{\boldsymbol{k}_{i_1,\ldots,i_N} = (k_{1,i_1},\ldots,k_{N,i_N}) : k_{d,i_d} = a_d^{\boldsymbol{k}} + i_d h_d^{\boldsymbol{k}}, i_d = 1,\ldots,m_d, d = 1,\ldots,K\right\},$$

with constant grid step $h_d^{\boldsymbol{k}} = \left(b_d^{\boldsymbol{k}} - a_d^{\boldsymbol{k}}\right)/(m_d+1)$.

Let $\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}} \in \mathbb{R}^{(n_1\cdots n_N m_1\cdots m_K)\times(n_1\cdots n_N m_1\cdots m_K)}$ be the (tensor) matrix from the finite difference approximation of operator $\mathscr{L}_{\mathrm{f}}^{\boldsymbol{k}}$. Based on the form of (3.104), one can show that matrix $\boldsymbol{p}_{\boldsymbol{h},t}^{\boldsymbol{k}}$ admits explicit rank-$(K+1)$ canonical representation of the form (Khoromskij & Schwab, 2011; Dolgov & Khoromskij, 2012):

$$\begin{aligned}\boldsymbol{A}_{\boldsymbol{h},\hat{t}}^{\boldsymbol{k}} =& \boldsymbol{A}_{\boldsymbol{h}}^{k_1}\otimes\boldsymbol{K}_1\otimes\boldsymbol{I}_2\otimes\cdots\otimes\boldsymbol{I}_K + \boldsymbol{A}_{\boldsymbol{h}}^{k_2}\otimes\boldsymbol{I}_1\otimes\boldsymbol{K}_2\otimes\cdots\otimes\boldsymbol{I}_K+\cdots \\ &+\boldsymbol{A}_{\boldsymbol{h}}^{k_K}\otimes\boldsymbol{I}_1\otimes\boldsymbol{I}_2\otimes\cdots\otimes\boldsymbol{K}_K + \left(\sum_{j=K+1}^{M} k_j \boldsymbol{A}_{\boldsymbol{h}}^{k_j}\right)\otimes\boldsymbol{I}_1\otimes\boldsymbol{I}_2\otimes\cdots\otimes\boldsymbol{I}_K,\end{aligned} \tag{3.106}$$

where $\boldsymbol{I}_d \in \mathbb{R}^{m_d\times m_d}$ denotes an identity matrix, $d = 1,2,\ldots,K$, and k_j are prescribed reaction rate constants $j = K+1,\ldots,M$. Matrix $\boldsymbol{K}_d \in \mathbb{R}^{m_d\times m_d}$ is diagonal with diagonal entries corresponding to the grid nodes of the d-th parameter, i.e.,

$$K_j(\ell_1,\ell_2) = \begin{cases} k_{j,\ell} & \ell_1 = \ell_2, \\ 0 & \ell_1 \neq \ell_2, \end{cases}$$

for $\ell_1,\ell_2 = 1,2,\ldots,m_j$ and $j = 1,2,\ldots,K$.

The (tensor) matrices $\boldsymbol{A}_{\boldsymbol{h}}^{k_j} \in \mathbb{R}^{(n_1\cdots n_N)\times(n_1\cdots n_N)}$, $j = 1,2,\ldots,K$, correspond to the finite difference discretisation of the sub-operator $\mathscr{L}_{\mathrm{f}}^{k_j}$ in (3.105) on grid $\omega_{\boldsymbol{h}}^{\boldsymbol{x}}$. Similar to the analysis in Section § 3.3.1, we construct $\boldsymbol{A}_{\boldsymbol{h},t}^{k_j}$ explicitly as

$$\boldsymbol{A}_{\boldsymbol{h}}^{k_j} \equiv \sum_{d=1}^{N}\left(\boldsymbol{\Lambda}_d^{k_j} + \boldsymbol{\Lambda}_{d,d}^{k_j}\right) + \sum_{\substack{d_1,d_2=1\\ d_1\neq d_2}}^{N}\left(\boldsymbol{\Lambda}_{d_1,d_2}^{k_j+} + \boldsymbol{\Lambda}_{d_1,d_2}^{k_j-}\right), \tag{3.107}$$

where

$$\Lambda^{k_j+}_{d_1,d_2} = \frac{\nu_{j,d_1}\nu_{j,d_2}}{2k_j}\left(\boldsymbol{Q}_{d_1,d_2} + \boldsymbol{Q}_{\hat{d}_1,\hat{d}_2}\right)\boldsymbol{H}_j \mathbb{1}_{\nu_{j,d_1}\nu_{j,d_2}>0},$$
$$\Lambda^{k_j-}_{d_1,d_2} = \frac{\nu_{j,d_1}\nu_{j,d_2}}{2k_j}\left(\boldsymbol{Q}_{d_1,\hat{d}_2} + \boldsymbol{Q}_{\hat{d}_1,d_2}\right)\boldsymbol{H}_j \mathbb{1}_{\nu_{j,d_1}\nu_{j,d_2}<0},$$
$$\Lambda^{k_j}_{d} = \frac{\nu_{j,d}}{k_j}\boldsymbol{Q}_{d+\hat{d}}\boldsymbol{H}_j \quad \text{and} \quad \Lambda^{k_j}_{d,d} = \frac{\nu_{j,d}}{k_j}\boldsymbol{Q}_{d,d}\boldsymbol{H}_j.$$

Here, the QTT structures of the tensor matrices $\boldsymbol{Q}_{d+\hat{d}}$, $\boldsymbol{Q}_{d,d}$, $\boldsymbol{Q}_{d_1,\hat{d}_2}$, $\boldsymbol{Q}_{\hat{d}_1,d_2}$, $\boldsymbol{Q}_{d_1,d_2}$, $\boldsymbol{Q}_{\hat{d}_1,\hat{d}_2}$ and $\boldsymbol{H}_j$ are explicitly given in Section § 3.3.1. A direct inspection of (3.106) and (3.107) leads to the following canonical rank estimate.

Proposition 3.56. *Consider (tensor) matrix* $\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t} \in \mathbb{R}^{(n_1\cdots n_N m_1\cdots m_K)\times(n_1\cdots n_N m_1\cdots m_K)}$ *in* (3.106)*. It has explicit canonical matrix representation (Definition 3.13) with canonical rank bounded by*

$$R^{\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t}} \leq 2MN + 2MN^2,$$

where M and N are the numbers of reactions and species, respectively.

The canonical representation of $\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t}$ can be converted to the corresponding TT representation with the same ranks (see Proposition 3.24). Further, since the matrices involved in (3.107) admit explicit QTT-structured construction, we derive the QTT representation of $\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t}$ in the following theorem.

Theorem 3.57. *Consider (tensor) matrix* $\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t} \in \mathbb{R}^{(n_1\cdots n_N m_1\cdots m_K)\times(n_1\cdots n_N m_1\cdots m_K)}$*,* $n_{d_1} = 2^{l^{x_{d_1}}_{d_1}}$*,* $m_{d_2} = 2^{l^{k_{d_2}}_{d_2}}$*,* $1 \leq d_1 \leq N$*,* $1 \leq d_2 \leq M$*, defined in* (3.106)*. It admits the QTT-structured representation of the form*

$$\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t} = \left(\boldsymbol{A}^{\boldsymbol{x}}_{\boldsymbol{h},t,1,1} \bowtie \cdots \bowtie \boldsymbol{A}^{\boldsymbol{x}}_{\boldsymbol{h},t,1,l^{x_1}_1}\right) \bowtie \cdots \bowtie \left(\boldsymbol{A}^{\boldsymbol{x}}_{\boldsymbol{h},t,N,1} \bowtie \cdots \bowtie \boldsymbol{A}^{\boldsymbol{x}}_{\boldsymbol{h},t,N,l^{x_N}_N}\right) \bowtie \cdots$$
$$\bowtie \left(\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t,1,1} \bowtie \cdots \bowtie \boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t,1,l^{k_1}_1}\right) \bowtie \cdots \bowtie \left(\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t,K,1} \bowtie \cdots \bowtie \boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t,K,l^{k_K}_k}\right),$$

where for $1 \le d \le N$

$$\begin{aligned}
\boldsymbol{A}^{\boldsymbol{x}}_{\boldsymbol{h},t,d,v_d} = & \bigoplus_{j=1}^{M} \Bigg\{ \Bigg[\bigoplus_{i=1}^{N} \left(\nu_{j,i} \left(\Psi_I \mathbb{1}_{i \neq d} + \Phi_{d+\hat{d},v_d} \mathbb{1}_{i=d} \right) \right) \oplus \left(\nu_{j,i}^2 \left(\Psi_I \mathbb{1}_{i \neq d} + \Phi_{d,d,v_d} \mathbb{1}_{i=d} \right) \right) \Bigg] \\
& \oplus \left[\bigoplus_{\substack{i_1,i_2=1 \\ i_1 \neq i_2}}^{N} \frac{\nu_{j,i_1} \nu_{j,i_2} \mathbb{1}_{\nu_{j,i_1} \nu_{j,i_2} > 0}}{2} \left(\left(\Phi_{d,v_d} \oplus \Phi_{\hat{d},v_d} \right) \mathbb{1}_{i_1=d \vee i_2=d} \right) \right] \\
& \oplus \left[\bigoplus_{\substack{i_1,i_2=1 \\ i_1 \neq i_2}}^{N} \frac{\nu_{j,i_1} \nu_{j,i_2} \mathbb{1}_{\nu_{j,i_1} \nu_{j,i_2} < 0}}{2} \left(\left(\Phi_{d,v_d} \oplus \Phi_{\hat{d},v_d} \right) \mathbb{1}_{i_1=d} + \left(\Phi_{\hat{d},v_d} \oplus \Phi_{d,v_d} \right) \mathbb{1}_{i_2=d} \right) \right] \\
& \oplus \left[\bigoplus_{\substack{i_1,i_2=1 \\ i_1 \neq i_2}}^{N} \frac{\nu_{j,i_1} \nu_{j,i_2} \mathbb{1}_{i_1 \neq d \wedge i_2 \neq d}}{2} \left(\Psi_I \oplus \Psi_I \right) \right] \Bigg\} \\
& \bullet \left\{ \left[\left((k_j - 1) \mathbb{1}_{j>K}) + 1 \right) V^{1 - \sum_{i=1}^{N} \nu_{j,i}^{-}} - 1 \right) \mathbb{1}_{d=N \wedge v_d = l_N} + 1 \right] \boldsymbol{H}_{j,d,v_d} \right\},
\end{aligned} \tag{3.108}$$

and for $1 \le d \le M$

$$\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t,d,v_d} = \Psi_I^{\oplus(j-1)} \oplus \Phi_{\boldsymbol{k},d,v_d} \oplus \Psi_I^{\oplus(N-j-1)}, \tag{3.109}$$

and $\Phi_{\boldsymbol{k},d,v_d}$ *is composed of the quantised block matrices:*

$$\Phi^{(1,j_d)}_{\boldsymbol{k},d,1} = C_1^{j_d - 1} \left(a_d^{\boldsymbol{k}} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + h_d^{\boldsymbol{k}} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix} \right)^{2 - j_d}, \quad \Phi^{(i_d,1)}_{j,d,l_d^{\boldsymbol{k}}} = 2^{(d-1)(i_d-1)} \left(h_d^{\boldsymbol{k}} \right)^{i_d - 1} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix},$$

$$\Phi^{(i_d,j_d)}_{j,d,v_d} = \begin{cases} C_{i_d-1}^{i_d - j_d} \left(2^{d-1} h_d^{\boldsymbol{k}} \right)^{i_d - j_d} \begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}, & i_d \ge j_d, \\ \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, & i_d < j_d, \end{cases}$$

for $i_d, j_d = 1, 2$, $v_d = 2, 3, \ldots, l_d^{k_d} - 1$, $1 \le d \le N$. *The summation operations of TT cores,*

$\oplus$ *and* $\bigoplus$*, are defined in Definition 3.22. The corresponding QTT ranks,*

$$\left(R_{1,1}^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}},\ldots,R_{1,l_1^{x_1}}^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}}\right),R_1^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}},\ldots,R_{N-1}^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}},\left(R_{N,1}^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}},\ldots,R_{N,l_N^{x_N}}^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}}\right),R_N^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}},$$
$$\left(R_{N+1,1}^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}},\ldots,R_{N+1,l_1^{k_1}}^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}}\right),R_{N+1}^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}},\ldots,R_{N+K-1}^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}},\left(R_{N+1,1}^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}},\ldots,R_{N+1,l_1^{k_1}}^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}}\right),$$

are bounded by: $R_d^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}} \leq 2MN + 2MN^2$*,* $1 \leq d \leq N-1$*, and*

$$R_{d,v_d}^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}} \leq \sum_{j=1}^{M}\left(\nu_{j,d}^{-}+1\right)\left[\sum_{i=1}^{N}\left(2\mathbb{1}_{i\neq d}+6\mathbb{1}_{i=d}\right)+\sum_{\substack{i_1,i_2=1\\ i_1\neq i_2}}^{N}\left(2\mathbb{1}_{i_1\neq d\wedge i_2\neq d}+6\mathbb{1}_{i_1=d\vee i_2=d}\right)\right], \tag{3.110}$$

for $1 \leq v_d \leq l_d^{x_d}-1$*,* $1 \leq d \leq N$*, and*

$$R_{d,v_d}^{\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}} \leq K+1,$$

for $1 \leq v_d \leq l_d^{k_d}-1$*,* $1 \leq d \leq K$*.*

Proof. The QTT structure of $\boldsymbol{A}_{\boldsymbol{h},t}^{\boldsymbol{k}}$ in (3.108) that corresponds to the first N dimensions (state variables) can be directly derived from (3.107) using the properties of TT operations (Proposition 3.23). The QTT structure (3.108) is similar to the QTT structure of $\boldsymbol{A}_{\boldsymbol{h},t}$ in (3.57) up to some difference in coefficient values, and thus the QTT rank bound in (3.110) is the same as the bound in (3.58). For the last K dimensions that corresponds to the parametric variables, the QTT structure is the result of (3.106), which suggests the QTT cores in the $(N+d)$-th dimension is the direct sum of the QTT core of $\boldsymbol{K}_d$ and the QTT cores of I_ℓ, $\ell \neq d$, which eventually yields (3.109). □

Remark 3.58. From Theorem 3.57, the leading order estimate of the storage re-

Table 3.2: *Comparison of the matrix-based and tensor-structured methodologies.*

Biochemical	Dimensionality			Matrix-based†		Tensor-based*			
system	N	K	$N+K$	$\mathrm{Mem}_{\mathrm{CME}}$	$\mathrm{Mem}_{\mathrm{CFPE}}$	$\mathrm{Mem}_{\boldsymbol{A}}$	$\mathrm{Mem}_{\boldsymbol{p}}$	$T_{\boldsymbol{A}}$	T_{tot}
Schlögl	1	4	5	2.68×10^{13}	2.74×10^{11}	4.01×10^{3}	2.07×10^{5}	1.2	30
Cell cycle	6	1	7	6.68×10^{17}	7.04×10^{13}	2.96×10^{4}	1.00×10^{7}	1.1	6433
F-N	2	4	6	6.38×10^{14}	1.75×10^{13}	7.65×10^{4}	4.02×10^{5}	0.7	37

† Estimated as the product of the number of discrete states and the number of parameter values.
* $\mathrm{Mem}_{\boldsymbol{A}}$ and $\mathrm{Men}_{\boldsymbol{p}}$ are the storage requirements for the QTT representation of the parametric CFPE operator $\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t}$ and the parametric solution $\boldsymbol{p}^{\boldsymbol{k}}_{\boldsymbol{h},t}$.

quirements and complexity of $\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t}$ in QTT format is

$$\mathcal{O}\left(M^2N^5l+K^3l\right),$$

in contrast to $\mathcal{O}\left(2^{2^{(N+K)l}}\right)$ in the full matrix format.

Now, the stationary parametric CFPE (3.103) is transformed into a principle eigenvalue problem in a tensor form

$$\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t}\boldsymbol{p}^{\boldsymbol{k}}_{\boldsymbol{h},t}=\lambda^{\boldsymbol{k}}_{\boldsymbol{h},t}\boldsymbol{p}^{\boldsymbol{k}}_{\boldsymbol{h},t}, \tag{3.111}$$

where $\lambda^{\boldsymbol{k}}_{\boldsymbol{h},t}$ and $\boldsymbol{p}^{\boldsymbol{k}}_{\boldsymbol{h},t}\in\mathbb{R}^{n_1\times\cdots\times n_N\times m_1\times\cdots\times m_K}$ are the corresponding principle eigenvalue and the principle (tensor) eigenvectors respectively. Since matrix $\boldsymbol{A}^{\boldsymbol{k}}_{\boldsymbol{h},t}$ is explicitly constructed in the QTT-structured representation as in Theorem 3.57, the system (3.111) may be solved by the numerical methods proposed in Section § 3.4.

3.5.3 Numerical experiments

We simulate the parametric CFPE of four examples of stochastic networks: a bistable switch in the 5-dimensional Schlögl model (Schlögl, 1972), oscillations in the 7-dimensional cell cycle model (Tyson, 1991), and neurons excitability in the 6-dimensional FitzHugh-Nagumo system (Tsai et al., 2008). More details of these models will be given in Chapter 4.

The simulation procedure starts with the QTT-structured finite difference discretisation of the parametric CFPE operator (Section § 3.5.1), and then the corresponding principal eigenvalue problem (3.57) is solved using the adaptive shift inverse power method (Algorithm 4).

Such tensor-based simulation performances are presented in Table 3.2, which also includes a comparison with the traditional matrix-based approach. Simulations are performed on a 64-bit Linux desktop equipped with Quad-Core AMD Opteron$^{\text{TM}}$ Processor 8356 x 16 and 63 GB RAM. The minimum memory requirements of solving the CME and the CFPE using matrix-based methods, Mem_{CME} and Mem_{CFPE}, are estimated as products of numbers of discrete states times the total number of parameter combinations. Although the CFPEs have less memory requirements via increasing the grid sizes, the memory requirements for the parametric CMEs and the parametric CFPEs vary in ranges 10^{13}–10^{17} and 10^{11}–10^{13}, respectively, which are beyond the limits of the available hardware. In contrary, the tensor (QTT-based) approach maintains affordable computational and memory requirements for all three problems considered, as we show in Table 3.2. The major memory requirements are $\text{Mem}_{\boldsymbol{A}}$ and $\text{Mem}_{\boldsymbol{p}}$ to store the QTT representations of the discretised parametric CFPE operator and the parametric steady state distributions $\pi_{\text{f}}(\boldsymbol{x}|\boldsymbol{k})$, respectively. Similarly, $T_{\boldsymbol{A}}$ is the computational time to assemble the operator and T_{tot} is the total computational time.

Table 3.2 shows that the tensor-based methods can outperform standard matrix-based methods. It can also be less computationally intensive than stochastic simulations in some cases. For example, the total computational time is around 30 minutes for the tensor approach to simulate 64^4 different parameter combinations within the 4-dimensional parameter space of the Schlögl chemical system (see Table 3.2). If we wanted to compute the same result using the Gillespie SSA, we would have to run 64^4 different stochastic simulations. If they had to be all performed on one proces-

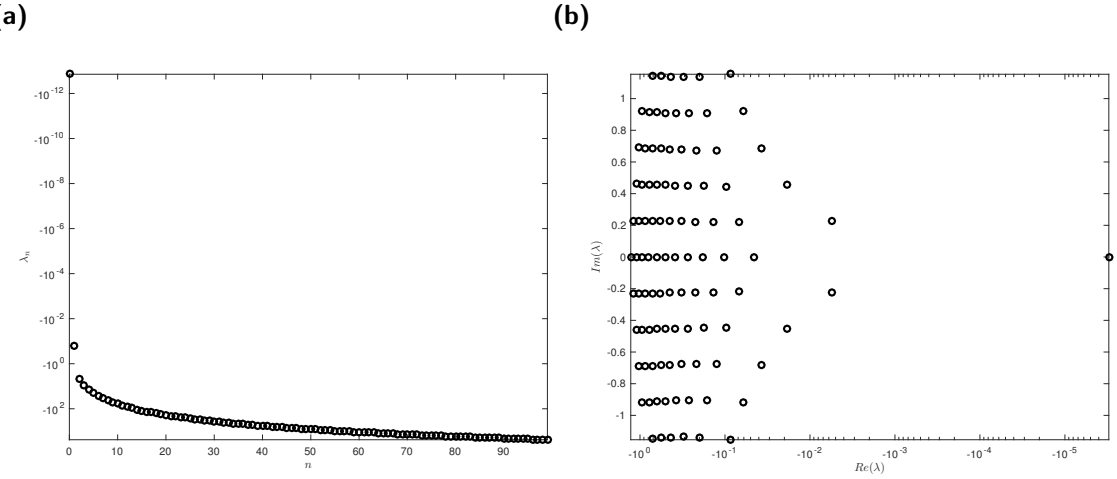

Figure 3.11: *The first 100 eigenvalues with smallest magnitudes of the finite difference operator of* **(a)** *the Schlögl model and* **(b)** *the F-N model.*

sor in 30 minutes, then we would only have 1.07×10^{-4} second per one stochastic simulation and it would not be possible to estimate the results with the same level of accuracy. In addition the tensor approach directly provides the steady state distribution $\pi_{\mathrm{f}}(\boldsymbol{x}|\boldsymbol{k})$, which would be computationally intensive to obtain by stochastic simulations (with the same level of accuracy) for larger values of $N+K$.

The computational time of the adaptive inverse power method in tensor format (Algorithm 4) mainly depends on three key factors: the dimensionality of the problem, the spectrum gap between the principle and the second eigenvalues, and the ranks of the (intermediate) solutions. The significant increase in the computational time of the cell cycle model (row 2 in Table 3.2) is mainly caused by first its oscillatory nature and its irregular shape of distribution. The oscillation time implies a long half-life and a very small spectrum gap between the principle and the second eigenvalues, and the irregular shape of distribution makes it more difficult to be approximated by a small set of low-rank function-related tensors (such as the Gaussian in Proposition 3.44).

Application of the (adaptive) inverse power scheme requires the principle eigenvalue of the finite difference operator to be non-positive and simple. The sufficient

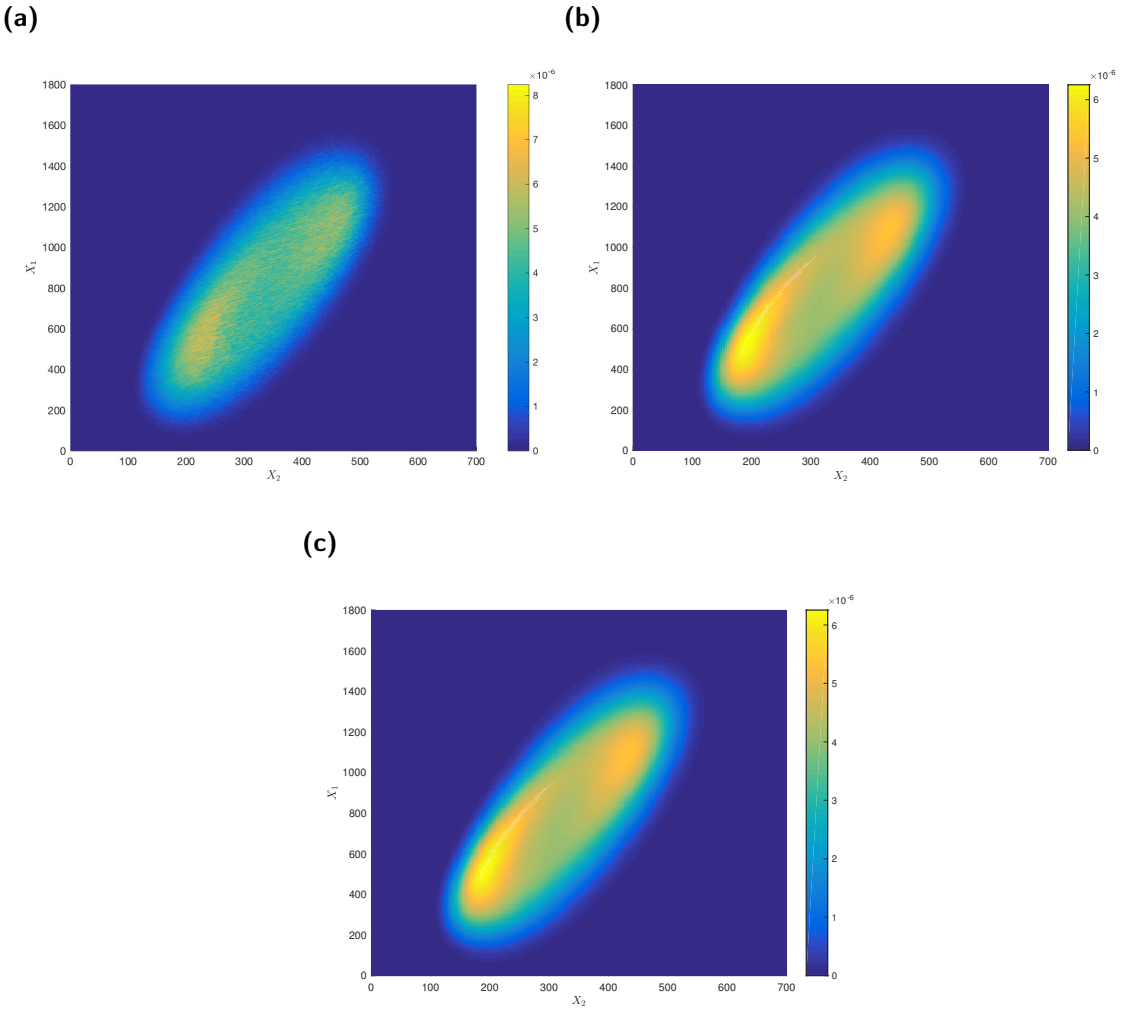

Figure 3.12: *Comparison of the numerical approximations of the stationary distribution of the F-N model, with model parameters:* $V = 2000$, $k_1 = 0.2005$, $k_2 = 0.1123$, $k_3 = 2.5059$ *and* $k_4 = 0.1052$. **(a)** *The CME distribution computed by using a long-time Gillespie simulation of* $T = 5 \times 10^6$. **(b)** *The principal eigenvector of the finite difference approximation of the corresponding CFPE computed using MATLAB eigs function in the full matrix format.* **(c)** *The CFPE distribution extracted from the tensor solution of the parametric CFPE. The four parameter ranges of the parametric CFPE are given in Table 4.4. The discretisation is equally spaced with 128 points. The plotted solution corresponds to the* 65*-th points along all four parameter direction.*

conditions for the simplicity of the principle eigenvalue are given in Lemma 3.4, Remark 3.5 and the subsequent discussions. The single-dimensional Schlögl model naturally satisfies all conditions, and its spectrum is plotted in Figure 3.11a. The spectrum distribution of the F-N model is also plotted in Figure 3.11a. Although the system does not fulfil the simplicity conditions proposed in section 3.1.2.2, the numerical data shows that the principle eigenvalue is simple with a spectrum gap of around 10^{-3}.

Figure 3.12 gives a visual comparison of the stationary distributions of the F-N model using different methods. The CFPE distribution (Figure 3.12b) gives a good approximation of the CME distribution (Figure 3.12a) with the given parameter values. Also, the tensor solution (Figure 3.12c) matches well with the solution obtained using the traditional matrix format (Figure 3.12b), and the error between the two solutions in ℓ^2-norm is 1.123×10^{-6}.

3.5.4 Computational implementation

The implementation of the proposed methods and tests is based on the MATLAB package Stochastic Bifurcation Analyser (StoBifAn), available at http://www.stobifan.org. The StoBifAn package includes the following functionalities:

- Assembling the (parametric) Fokker-Planck finite difference operator in the QTT format from a given stoichiometric matrix;
- Constructing the chemical master operator in the QTT format from a given stoichiometric matrix;
- The adaptive inverse power scheme (Algorithm 4 and 5) to solve the stationary (parametric) Fokker-Planck and CME;
- Post-processing the resulting tensor data to perform the parameter estimation using the moment-matching techniques.

The StoBifAn package relies on the MATLAB package TT-Toolbox to perform the ele-

mentary algebraic operations on the QTT tensors and solve the large linear systems in the QTT format. The TT-Toolbox is open access at http://github.-com/oseledets/TT-Toolbox.

"The purpose of computing is insight, not numbers."

Hamming (1962)

4

Tensor-structured parametric analysis of stochastic reaction networks

In this chapter, we will exemplify the tensor approach to conduct the parametric analysis of the stochastic networks, including parameter estimation, stochastic bifurcation analysis, and robustness analysis. The tensor-structured analysis framework can be divided into two main steps: a tensor-structured computation and a tensor-based analysis. First, the steady state distributions of stochastic models are simultaneously computed for all possible parameter combinations within a parameter space and stored in a tensor format, with smaller computational and memory requirements than in traditional approaches. The resulting tensor data are then analysed using algebraic operations with computational complexity which scales linearly with dimension (i.e., linearly with N and K). Major results in the current chapter have been published in the primarily authored paper, Liao et al. (2015). The

discussions were previously announced in a reduced form (due to the journal page limit), and are now further extended in this chapter. The numerical experiments are presented with no changes.

4.1 Parameter estimation

Small uncertainties in the reaction rate values of stochastic reaction networks (2.1) are common in applications. Some model parameters are difficult to measure directly, and instead are estimated by fitting to time-course data. If the reaction networks are modelled using deterministic ODEs, there is a wide variety of tools available for parameter estimation. Many simple approaches are non-statistical (Moles et al., 2003), and the procedure usually, although not necessarily (Jaqaman & Danuser, 2006), follows the algorithm presented in Algorithm 6. This approach seeks the set of those parameters that minimize the distance measure $d(\hat{\boldsymbol{x}}, \boldsymbol{x}^*)$, while the rules to generate candidate parameters $\boldsymbol{k}^*$ in step (a1) and the definition of distance function along with stopping criteria in step (a3) may vary in different methods. In optimization-based methods, $\boldsymbol{k}^*$ may follow the gradient on the surface of the distance function (Moles et al., 2003). In statistical methods, such distance measure is provided in the concept of likelihood, $L(\boldsymbol{k}^*|\hat{\boldsymbol{x}}) = \pi(\hat{\boldsymbol{x}}|\boldsymbol{k}^*)$ (Pedersen, 1995). In Bayesian methods, the candidate parameters $\boldsymbol{k}^*$ are generated from some prior information regarding uncertain parameters, $\pi(\boldsymbol{k})$, and form a posterior distribution rather than a single point estimate (Toni et al., 2009).

To extend the algorithm in Algorithm 6 from deterministic ODEs to stochastic models requires substantial modifications (Toni et al., 2009). One main obstacle is the step (a2) which requires repeatedly generating the likelihood function $L(\boldsymbol{k}^*|\hat{\boldsymbol{x}})$, as the outcome of stochastic models. In this case, a modeller must either apply statistical analysis to approximate the likelihood (Reinker et al., 2006), or use the Gillespie

Algorithm 6 *Parameter estimation for ODEs.*

1: **repeat**
2: generate a candidate parameter vector $\boldsymbol{k}^* \in \Omega_{\boldsymbol{k}}$;
3: compute model prediction $\boldsymbol{x}^*$ using the parameter vector $\boldsymbol{k}^*$;
4: compare the simulated data $\boldsymbol{x}^*$ with the experimental evidence $\hat{\boldsymbol{x}}$, using distance function $d(\hat{\boldsymbol{x}},\boldsymbol{x}^*)$ and tolerance ε (the tolerance $\varepsilon > 0$ is the desired level of agreement between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}^*$);
5: **if** $d(\hat{\boldsymbol{x}},\boldsymbol{x}^*) < \varepsilon$ **then**
6: Accept $\boldsymbol{k}^*$;
7: **else**
8: Reject $\boldsymbol{k}^*$;
9: **end if**
10: **until** some stopping criteria is reached.

SSA to estimate it (Tian et al., 2007). Consequently, the algorithms are computationally intensive and do not scale well to problems of realistic size and complexity. To avoid this problem, the TPA uses the tensor formalism to separate the simulation part from the parameter inference. The parameter estimation is performed on the tensor data obtained by simulation methods described in the previous chapter (see Table 3.2 in Section § 3.5). The algorithm used for the tensor-based parameter estimation is given in Algorithm 7. The distance function $d(\hat{\boldsymbol{x}},\boldsymbol{x}^*)$ in Algorithm 6 is replaced with a distance between summary statistics, $\hat{S}$ and S^*, which describe the average behaviour and the characteristics of the system noise. The steps 2–11 in Algorithm 7 are similar to steps 1–10 in Algorithm 6 under the ODE settings, and a variety of existing methods can be extended directly to stochastic settings. The newly introduced step 1 in Algorithm 7 is executed only once during the parameter estimation. Steps 3–10 are then repeated until convergence. Step 4 only requires manipulation of tensor data, of which the computational overhead is comparable to solving an ODE.

Algorithm 7 *An algorithm for the tensor-structured parameter estimation.*

compute the stationary distribution $\pi(\boldsymbol{x}|\boldsymbol{k})$ for all considered combinations of $\boldsymbol{x} \in \Omega$ and $\boldsymbol{k} \in \Omega_{\boldsymbol{k}}$; and store $\pi(\boldsymbol{x}|\boldsymbol{k})$ in a tensor format.
repeat
  generate a candidate parameter vector $\boldsymbol{k}^* \in \Omega_{\boldsymbol{k}}$;
  extract the stationary distribution $\pi(\boldsymbol{x}|\boldsymbol{k}^*)$ from the tensor-structured data $\pi(\boldsymbol{x}|\boldsymbol{k})$, and compute the summary statistics $S^* \equiv S^*(\pi)$;
  compare the model prediction S^* with the statistics $\hat{S}$ obtained from experimental data, using distance function $J(\hat{S}, S^*)$ and tolerance ε. (the tolerance $\varepsilon > 0$ is the desired level of agreement between $\hat{S}$ and S^*);
  if $J(\hat{S}, S^*) < \varepsilon$ **then**
    accept $\boldsymbol{k}^*$;
  else
    reject $\boldsymbol{k}^*$;
  end if
until some stopping criteria is reached.

4.1.1 Moment-based inference

We consider that the distance measure $J_{\text{dist}}(\hat{S}, S^*)$ in Algorithm 7 is defined using a moment matching procedure (Lillacci & Khammash, 2010; Zechner et al., 2012):

$$J_{\text{dist}}(\hat{S}, S^*) = \sum_{i_1,\ldots,i_N=1}^{L} \beta_{i_1,\ldots,i_N} \left(\frac{\hat{\mu}_{[i_1,\ldots,i_N]} - \mu_{[i_1,\ldots,i_N]}\left(\boldsymbol{k}^*\right)}{\hat{\mu}_{[i_1,\ldots,i_N]}} \right)^2, \tag{4.1}$$

where $\hat{\mu}_{[i_1,\ldots,i_N]}$ is the $(i_1,\ldots,i_N)$-th order empirical raw moment, $\mu_{[i_1,\ldots,i_N]}(\boldsymbol{k}^*)$ is the corresponding moment derived from $\pi(\boldsymbol{x}|\boldsymbol{k}^*)$ and L denotes the upper bound for the moment order. The weights, $\beta_{i_1,\ldots,i_N}$, can be chosen by modellers to attribute different relative importances to moments. Empirical moments are estimated from samples $\hat{x}_{d,\ell}$, $d = 1,2,\ldots,N$, $\ell = 1,2,\ldots,n_{\hat{\mu}}$, by

$$\hat{\mu}_{[i_1,\ldots,i_N]} = \frac{1}{n_{\hat{\mu}}} \sum_{\ell=1}^{n_{\hat{\mu}}} \hat{x}_{1,\ell}^{i_1} \cdots \hat{x}_{N,\ell}^{i_N}, \tag{4.2}$$

where $n_{\hat{\mu}}$ is the number of samples. Moments of the model output are computed as

$$\mu_{[i_1,\ldots,i_N]}\left(\boldsymbol{k}^*\right) = \int_{\Omega_{\boldsymbol{x}}} x_1^{i_1} \cdots x_N^{i_N} \, \pi\left(\boldsymbol{x}|\boldsymbol{k}^*\right) \mathrm{d}\boldsymbol{x}. \tag{4.3}$$

When the parametric distribution $\pi(\boldsymbol{x}|\boldsymbol{k}^*)$ is computed and stored in a tensor form, the integral (4.3) can be simultaneously approximated for all parameter sets $\boldsymbol{k}^* \in \Omega_{\boldsymbol{k}}$ through successive application of the mode product (see Section § 3.2.1), i.e.,

$$\mu_{[i_1,\ldots,i_N]}(\boldsymbol{k}^*) \approx \boldsymbol{\mu}_{[i_1,\ldots,i_N]} = h_1^{\boldsymbol{x}} h_2^{\boldsymbol{x}} \cdots h_N^{\boldsymbol{x}} \left(\boldsymbol{p}_{\boldsymbol{h},t}^{\boldsymbol{k}} \times_1 \boldsymbol{x}_1^{i_1} \times_2 \boldsymbol{x}_2^{i_2} \times_3 \cdots \times_N \boldsymbol{x}_N^{i_N} \right), \tag{4.4}$$

where $\boldsymbol{x}_d^{i_d} = \left(x_{d,1}^{i_d}, x_{d,2}^{i_d}, \ldots, x_{d,n_d}^{i_d} \right)^T$ and $h_d^{\boldsymbol{x}}$ is the grid size defined in (3.6). Note that tensor $\boldsymbol{x}_d^{i_d}$ refers to the equidistant samplings of a polynomial of order i_d, and its QTT ranks are bounded by $i_d + 1$ (see Proposition 3.28). Thus, the computational cost of performing (4.4) in the QTT format is of order $\mathcal{O}\left(\log_2(n) N R^2\right)$, where R is the maximum QTT ranks of $\boldsymbol{p}_{\boldsymbol{h},t}^{\boldsymbol{k}}$. One can execute (4.4) to obtain all required moments. Then the distance function $J_{\text{dist}}(\hat{S}, S^*)$ in (4.1) for all parameters within the parameter space can be directly computed in the tensor format via

$$\boldsymbol{J}_{\text{dist}} = \sum_{i_1,\ldots,i_N=1}^{L} \beta_{i_1,\ldots,i_N} \left(\boldsymbol{e} - \frac{1}{\hat{\mu}_{[i_1,\ldots,i_N]}} \boldsymbol{\mu}_{[i_1,\ldots,i_N]} \right)^2, \quad \boldsymbol{J}_{\text{dist}} \in \mathbb{R}^{m_1 \times \cdots \times m_K}, \tag{4.5}$$

where $\boldsymbol{e} \in \mathbb{R}^{n_1 \times \cdots \times n_N \times m_1 \times \cdots \times m_K}$ is an $(N+K)$-dimensional tensor of all ones. The resulting data $\boldsymbol{J}_{\text{dist}}$ are stored in the tensor format which enables efficient manipulations and post-processing.

4.1.2 An example of parameter estimation: the Schlögl reaction system

We illustrate the tensor-structured parameter estimation using the Schlögl chemical system (Schlögl, 1972), which is written for $N = 1$ molecular species and has $M = 4$ reaction rate constants k_i, $i = 1,2,3,4$. A detailed description of this system is given

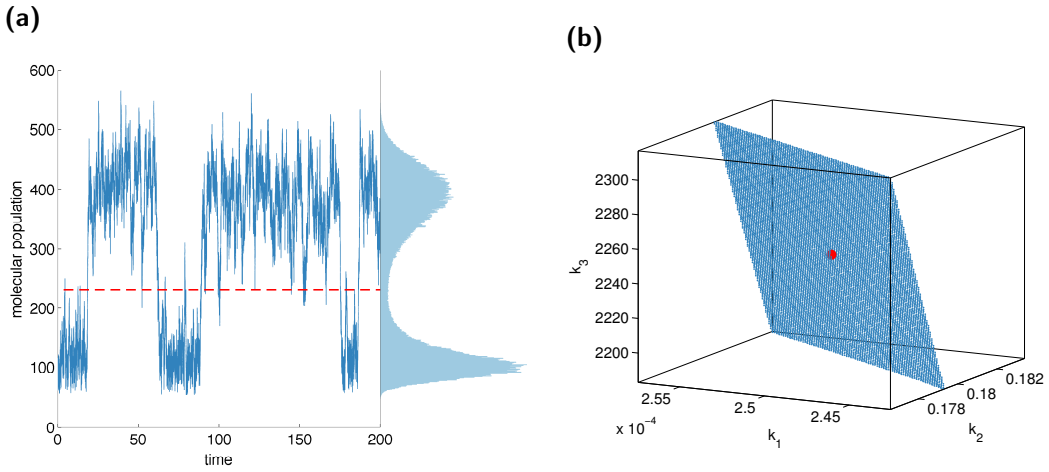

Figure 4.1: *Visualisation and parameter inference of the Schlögl reaction system.* **(a)** *A short segment of the time series data and the histogram for the Schlögl reaction system generated by a long-time stochastic simulation. The dashed line corresponds to the threshold* 230 *which is used to separate the two macroscopic states of this bistable system.* **(b)** *The triples of parameters* $[k_1, k_2, k_3]$ *for which the splitting probability* (4.9) *is equal to* $\hat{S} = 47.61\% \pm 5\%$ *form a plane with a very thin thickness (in blue) within the 3D parameter space. The value of* k_4 *is fixed at its true value, and the true value of* $[k_1, k_2, k_3]$ *is marked with the red dot.*

by

$$\emptyset \underset{k_2}{\overset{k_1}{\rightleftarrows}} S \quad \text{and} \quad 2S \underset{k_4}{\overset{k_3}{\rightleftarrows}} 3S, \tag{4.6}$$

where $k_j > 0$, $j = 1,2,3,4$, are reaction constants. We prescribe true parameter values as $k_1 = 2.5 \times 10^{-4}$, $k_2 = 0.18$, $k_3 = 2250$ and $k_4 = 37.5$, and use a long-time stochastic simulation to generate a time series as pseudo-experimental data (for a short segment, see Figure 4.1a). These pseudo-experimental data are then used for estimating the first three empirical moments $\hat{\mu}_i$, $i = 1,2,3$, using (4.2), and the values are given in Table 4.2.

Let n be the molecular number, and the propensities are given by

$$\alpha_1(n) = k_1 V, \quad \alpha_2(n) = k_2 n, \quad \alpha_3(n) = k_3 n(n-1)/V, \quad \text{and} \quad \alpha_4(n) = k_4 n(n-1)(n-2)/V^2.$$

We write the underlying stationary CME as

$$\beta_{n-1}\pi_{\mathrm{m}}(n-1)+\gamma_{n+1}\pi_{\mathrm{m}}(n+1)-(\beta_n+\gamma_n)\pi_{\mathrm{m}}(n)=0, \quad n=0,1,2,\ldots,\infty, \tag{4.7}$$

and the corresponding Fokker-Planck approximation is of the form

$$\frac{\mathrm{d}^2}{\mathrm{d}x^2}\left(D(x)\pi_{\mathrm{f}}(x)\right)-\frac{\mathrm{d}}{\mathrm{d}x}\left(\mu(x)\pi_{\mathrm{f}}(x)\right)=0, \quad x\in\Omega_{\mathrm{f}}, \tag{4.8}$$

where x is a continuous variable in domain $\Omega_{\mathrm{f}}=\{x\in\mathbb{R}:D(x)>0\}$, and the diffusion and drift coefficients are respectively given by

$$D(x)=\frac{1}{2}\left(\alpha_1(x)+\alpha_2(x)+\alpha_3(x)+\alpha_4(x)\right), \quad \mu(x)=\alpha_1(x)-\alpha_2(x)+\alpha_3(x)-\alpha_4(x).$$

Here, we consider all four parameters in (4.6) as axillary variables, and the parameter space $\Omega_{\boldsymbol{k}}$ is chosen such that it is centred at the prescribed true parameter values, as shown in Table 4.1. Then, the corresponding parametric CFPE is solved with all sampling values of parameters within Ω_{f} using the method described in Section § 3.5.1, and the simulation performance is given in the first row of Table 3.2. It then follows that the moments of the model output, $\boldsymbol{\mu}_i$, $i=1,2,3$, can be directly derived by applying the formula (4.4) to the stored tensor data.

Moment matching (4.1) can be sensitive to the choice of weights (Zechner et al., 2012). However, for the sake of simplicity we choose the weights β_i, $i=1,2,3$, in a way that the contributions of the different orders of moments are of similar magni-

Table 4.1: *Properties of molecular and rate variables in the Schlögl model.*

Type	Notation	Range	No. of nodes
Species	X	$[0,1000]$	1024
Rate	k_1	$[2.43\times10^{-4},2.58\times10^{-4}]$	128
Rate	k_2	$[0.17,0.19]$	128
Rate	k_3	$[2134,2266]$	128
Rate	k_4	$[36.08,38.63]$	128

tude within the parameter space (Table 4.2).

Having the moment order $\boldsymbol{\mu}_i$, $i = 1,2,3$, computed and stored in the QTT format, we can then compute the distant function using (4.5), and efficiently iterate steps (b1)–(b3) in Algorithm 7 to search for parameter values that produce adequate fit to the samples using the measure given in equation (4.1). We consider $\varepsilon = 0.25\%$ and visualise in Figure 4.2 the admissible parameter values satisfying $J_{\text{dist}}(\hat{S}, S^*) < \varepsilon$.

Note that all four values of parameters, k_1, k_2, k_3 and k_4, cannot be estimated solely from the steady state distributions, because it does not inform us how fast the system reaches the steady state. This can be also inferred from the splitting of the parametric Fokker-Planck operator (3.104) as a perfect collinear relationship between the kinetic rate parameters. In the context of parameter estimation this means that a single constraint (like mean and variance) restricts the original K-dimensional parameter space to an $(K-1)$-dimensional subspace of parameter values that comply with the constraint. Therefore, if a direct comparison between a model and a sample is not possible, a necessary condition to statistically infer $K \leq M$ parameters of a stochastic system is to define at least K constraints. In particular, if $\{k_i\}_{i=1,2,3,4}$ fit the pseudo-experimental data, then $\{Ck_i\}_{i=1,2,3,4}$ for any $C > 0$ fit these data as well. Therefore, in Figures 4.1b and 4.2, we fix one of the parameters at its true value and estimate values of the other three.

The summary statistics $\hat{S}$ are not restricted to moment orders. The TPA can efficiently evaluate different choices of the summary statistics, because of the simplicity and generality of the separable representation. For example, if one can experimentally measure the probability that the system stays in each of the two states of the

Table 4.2: *Moments estimated from stochastic simulation of the Schlögl model.*

Moment order	Value	Weight
1	$\hat{\mu}_1 = 261.32$	$\beta_1 = 1$
2	$\hat{\mu}_2 = 2.03 \times 10^4$	$\beta_2 = 100$
3	$\hat{\mu}_3 = -2.04 \times 10^5$	$\beta_3 = 0.001$

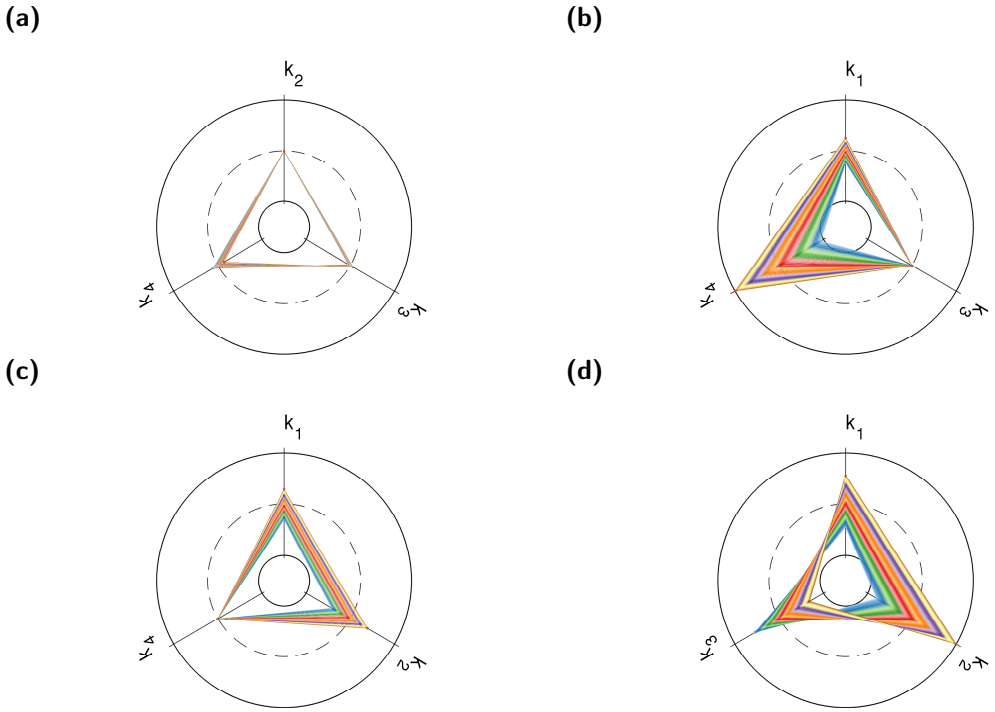

Figure 4.2: *Circular representation* (Von Dassow et al., 2000) *of estimated parameter combinations for the Schlögl model. Each spoke represents the corresponding parameter ranges:* $k_1 \in [2.43 \times 10^{-4}, 2.58 \times 10^{-4}]$, $k_2 \in [0.17, 0.19]$, $k_3 \in [2134, 2266]$, *and* $k_4 \in [36.08, 38.63]$. *The true parameter values are specified by the intersection points between the spokes and the dashed circle. Each triangle (or polygon in general) of a fixed colour corresponds to one admissible parameter set with* $\varepsilon = 0.25\%$. *Each panel* **(a)–(d)** *shows the situation with one parameter fixed at its true value, namely:* **(a)** k_1 *is fixed*; **(b)** k_2 *is fixed*; **(c)** k_3 *is fixed*; **(d)** k_4 *is fixed*.

bistable system, then distance measure $J_{\text{dist}}(\hat{S}, S^*)$ can be based on the probability of finding the system within a particular part of the state space $\Omega_{\boldsymbol{x}}$. Considering the Schlögl model, we estimate the probability that the system stays in the state with fewer molecules by

$$S(\boldsymbol{k}) = \mathbb{P}(x \leq 230 | \boldsymbol{k}), \tag{4.9}$$

where $\mathbb{P}$ denotes the probability and the threshold $x = 230$ separates the two macroscopic states of the Schlögl system, see the dashed line in Figure 4.1a. The splitting probability (4.9) can be estimated using long time simulation of the Schlögl system as the fraction of states which are less or equal than 230 and is equal to $\hat{S} = 47.61\%$ for our true parameter values. We can compute the model prediction of (4.9) using

the computed parametric distribution $\pi(x|\boldsymbol{k})$ via the integration

$$\mathbb{P}(x \leq 230|\boldsymbol{k}) = \int_0^{230} \pi(x|\boldsymbol{k})\mathrm{d}x. \tag{4.10}$$

The above integral can be also achieved by mode-1 product by a vector whose entries equal to 1 if it corresponds to $x < 230$ or 0 otherwise. The operation has $\mathcal{O}(n)$ complexity.

Once (4.10) is evaluated for all parameter values in the tensor format, one can search for the parameter combinations that satisfy certain tolerance. We show in Figure 4.1b the set of admissible parameters within the parameter space $\Omega_{\boldsymbol{k}}$ whose values provide desired agreement on the splitting probability (4.9) with tolerance $\varepsilon = 5\%$, i.e. we use

$$J_{\text{dist}}(\hat{S}, S^*) = |\hat{S} - S^*|$$

in Algorithm 7, where S^* is computed using (4.10).

4.1.3 Identifiability

One challenge of mathematical modelling of reaction networks is whether unique parameter values can be determined from available data. This is known as the problem of *identifiability*. Inappropriate choice of the distance measure may yield ranges of parameter values with equally good fit, i.e. the parameters being not identifiable (Gutenkunst et al., 2007). Here, we illustrate the tensor-structured identifiability analysis of the deterministic and stochastic models of the Schlögl chemical system. We plot the distance function against two parameter pairs, rate constants k_1-k_3 and k_2-k_4, in Figure 4.3. From the colour map, we see that the distance function (4.1) possesses a well distinguishable global minimum at the true values ($k_1 = 0.00025$, $k_2 = 0.18$, $k_3 = 2250$ and $k_4 = 37.5$). This indicates that the stochastic model is identifiable in both cases. In the deterministic scenario, the Schlögl system loses its

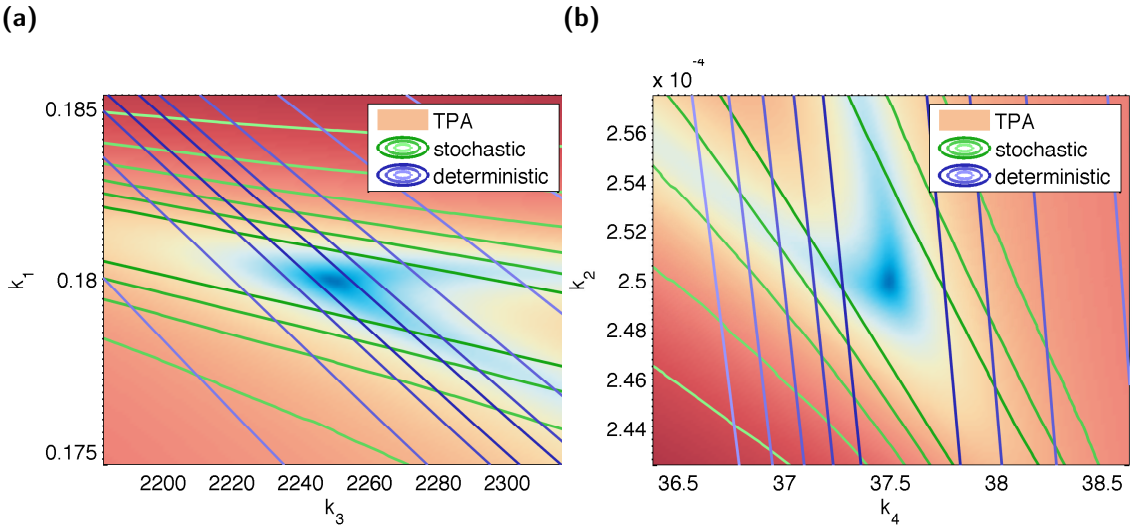

Figure 4.3: *Parameter identifiability analysis of the Schlögl reaction system.* **(a)** *Estimation and identifiability of parameters k_1 and k_3 with k_2 and k_4 fixed at their true values. The colour scale corresponds to values of the distance function* (4.1) *with $L = 3$, i.e., the first three moments are compared. The green and blue contour lines indicate the distance functional* (4.1) *with the first moment (mean value) only. Green corresponds to the stochastic model and blue to the deterministic model.* **(b)** *Estimation and identifiability of parameters k_2 and k_4 with k_1 and k_3 fixed at their true values. The same quantities as in panel* **(a)** *are plotted.*

identifiability. When the distance function (4.1) only fits the mean concentration, the minimal values are attained on a curve in the 2D parameter space (the distance function is indicated by blue contour lines in Figure 4.3). Stochastic models are advantageous in model identifiability, because they can be parametrized using a wider class of statistical properties (typically, K quantities are needed to estimate K reaction rate constants for mass-action reaction systems). The TPA enables efficient and direct evaluation of $J_{\text{dist}}\left(\hat{S}, S^*\right)$ all over the parameter space in a single computation by (4.5).

Figure 4.3 also reveals the differences between the model responses to parameter perturbations. The green contour lines show the landscape of $J_{\text{dist}}(\hat{S}, S^*)$ for the stochastic model using only the mean values, i.e., $L = 1$ in (4.1). The minimum is attained on a straight line, representing another non-identifiable situation. This line (green) has a different direction than the line obtained for the deterministic model (blue). In particular, this example illustrates that the parameter values estimated

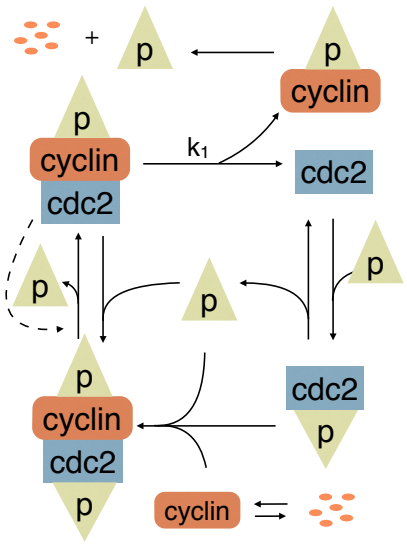

Figure 4.4: *Schematic description of the cyclin-cdc2 interactions.*

from deterministic models do not give good approximation of both average behaviour and the noise level when they are used in stochastic models (Wilkinson, 2009).

4.2 Stochastic bifurcation

Bifurcation is defined as a qualitative transformation in the behaviour of the system as a result of the continuous change in model parameters. Bifurcation analysis of ODE systems has been used to understand the properties of deterministic models of biological systems, including models of cell cycle (Tyson et al., 2002) and circadian rhythms (olde Scheper et al., 1999). Software packages, implementing numerical bifurcation methods for ODE systems, have also been presented in the literature (Doedel et al., 1997), but computational methods for bifurcation analysis of corresponding stochastic models are still in development (Erban et al., 2009).

4.2.1 A stochastic model of the cell cycle progression

In the current section, we exemplify stochastic bifurcation analysis based on tensors to a more complex genetic network: an early model of fission yeast cell cycle control by Tyson (Tyson, 1991). The cell cycle model consists of six chemical species which are subject to the following chemical reactions

$$\begin{aligned}
p-cyclin-cdc2 &\xrightarrow{k_1} cdc2+p-cyclin,\\
\emptyset &\xrightarrow{k_2 V} cyclin,\\
cdc2-p+cyclin &\xrightarrow{k_3/V} p-cyclin-cdc2-p,\\
p-cyclin-cdc2-p &\xrightarrow{k_4'+k_4[p-cyclin-cdc2]} p-cyclin-cdc2,\\
p-cyclin-cdc2 &\xrightarrow{k_5 P_{\text{tot}}} p-cyclin-cdc2-p,\\
cyclin &\xrightarrow{k_6} \emptyset,\\
p-cyclin &\xrightarrow{k_7} \emptyset,\\
cdc2 &\xrightarrow{k_8 P_{\text{tot}}} cdc2-p,\\
cdc2-p &\xrightarrow{k_9} cdc2.
\end{aligned} \tag{4.11}$$

The network structure of cyclin-cdc2 is also visualised in Figure 4.4. Free cyclin molecules combine rapidly with phosphorylated cdc2, to form the dimer MPF (cdc2-cyclin-p), which is immediately inactivated by phosphorylation process. The inactive MPF (p-cdc2-cyclin-p) can be converted to active MPF by autocatalytic dephosphorylation. The active MPF in excess breaks down into cdc2 molecules and phosphorylated cyclin, which is later subject to proteolysis. Finally, cdc2 is phosphorylated to repeat the cycle.

Table 4.3: *Properties of molecular and rate variables in the cell cycle model.*

Type	Name	Notation	Range	No. of nodes
Species	cdc2	C2	[2230, 4990]	N/A[a]
Species	cdc2-P	CP	[10, 70]	256
Species	p-cyclin-cdc2-p	pM	[0, 1500]	256
Species	p-cyclin-cdc2	M	[0, 1200]	256
Species	cyclin	Y	[20, 70]	256
Species	p-cyclin	YP	[0, 700]	256
Rate	degradation rate of active MPF	k_1	[0.25, 0.4]	64

[a] Discretisation of cdc2 is not applicable here, since this variable is eliminated by the conservation law of cdc2 assumed by the original author.

We adopt the parameter values from the original authors (Tyson, 1991) as

$$k_2 = 0.015, \quad k_3 = 200, \quad k_4 = 180, \quad k_4' = 0.018, \quad k_5 = 0, \quad k_6 = 0,$$
$$k_7 = 0.6, \quad k_8 = 1000, \quad k_9 = 1000 \quad P_{\text{tot}} = 0.001 \quad V = 1.$$

The parameter k_1, indicating the breakdown of the active M-phase-promoting factor (MPF), is chosen as the bifurcation parameter. The analysis of the corresponding ODE model reveals that the system displays a stable steady state when k_1 is at its low values, which describes the metaphase arrest of unfertilised eggs (Hunt, 1989). On the other hand, the ODE model is driven into rapid cell cycling exhibiting oscillations when k_1 increases (Tyson, 1991). The ODE cell cycle model has a bifurcation point at $k_1 = 0.2694$, where a limit cycle should appear (Tyson, 1991). The dynamical behaviours of the ODE model for different k_2 values are plotted in Figures 4.5a, 4.5c, 4.5e, and 4.5g.

For the stochastic model, we simulate the steady state distributions for a range of k_1 values using the tensor methods described in Section § 3.5.3. Specifically, the distribution is sought as the stationary solution of the parametric CFPE (3.101), and the ranges of the state and parametric variables, as well as their discretisations, are shown in Table 4.3. If we using the traditional vector-based approaches, the size of the problem would be of order 10^{13}, and it is far beyond the available computational

resources. While using tensor approach and perform simulations in the QTT format, we are capable to maintain the whole simulation within affordable computational cost (see the second row of Table 3.2). The stored tensor data allows us to visualise marginal distributions of the stochastic model as plotted in Figures 4.5b, 4.5d, 4.5f and 4.5h.

4.2.2 Stochastic bifurcation structure revealed by tensor simulations

We study the behaviour of the stochastic model for the values of k_1 which are close to the deterministic bifurcation point. We observe that the steady state distribution changes from a uni-modal shape (Figure 4.5f) to a probability distribution with a "doughnut-shaped" region of high probability (Figure 4.5h) at $k_1 = 0.3032$. In particular, the stochastic bifurcation appears for higher values of k_1 than the deterministic bifurcation.

In Figure 4.6, we use the computed tensor-structured parametric probability distribution to visualise the stochastic bifurcation structure of the cell cycle model. As the bifurcation parameter k_1 increases, the expected oscillation tube is formed and amplified in the marginalised YP-pM-M state space (see panels (a)–(d) of Figure 4.6). In panels (e)–(h) of Figure 4.6, the marginal distribution in the Y-CP-pM subspace is plotted. We see that it changes from a uni-modal (Figure 4.6e) to a bimodal distribution (Figure 4.6f). Cell cycle models have been studied in the deterministic context either as oscillatory (Tyson, 1991) or bistable (Pomerening et al., 2003; Xiong & Ferrell, 2003) systems. In Figure 4.6, we see that the presented stochastic cell cycle model can appear to have both oscillations and bi-modality, when different subsets of observables are considered.

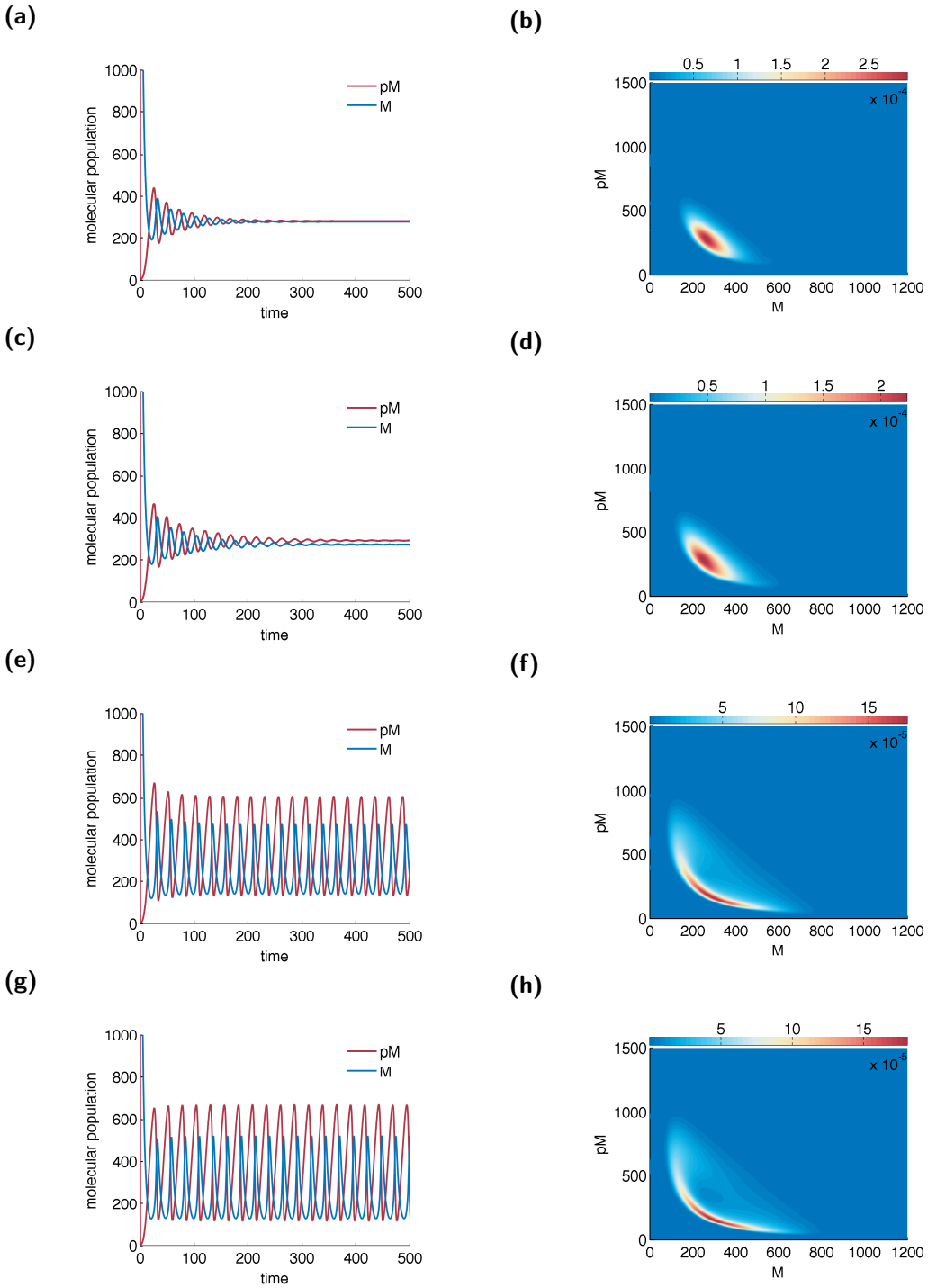

Figure 4.5: *Comparisons of deterministic and stochastic bifurcation of the cell cycle model. The first column of figures demonstrate the deterministic trajectories of the inactive MPF (pM) and the active MPF (M) with different value of kinetic rate k_1, and figures on the right demonstrate the corresponding steady state distributions computed using the tensor methods. The order of the four figure pairs are organised as follows:* **(a,b)** *at $k_1 = 0.2645$ right before the ODE bifurcation point;* **(c,d)** *at $k_1 = 0.2694$ after the ODE bifurcation point;* **(e,f)** *at $k_1 = 0.2984$ before the stochastic bifurcation point; and* **(g,h)** *at $k_1 = 0.3032$ after the stochastic bifurcation point.*

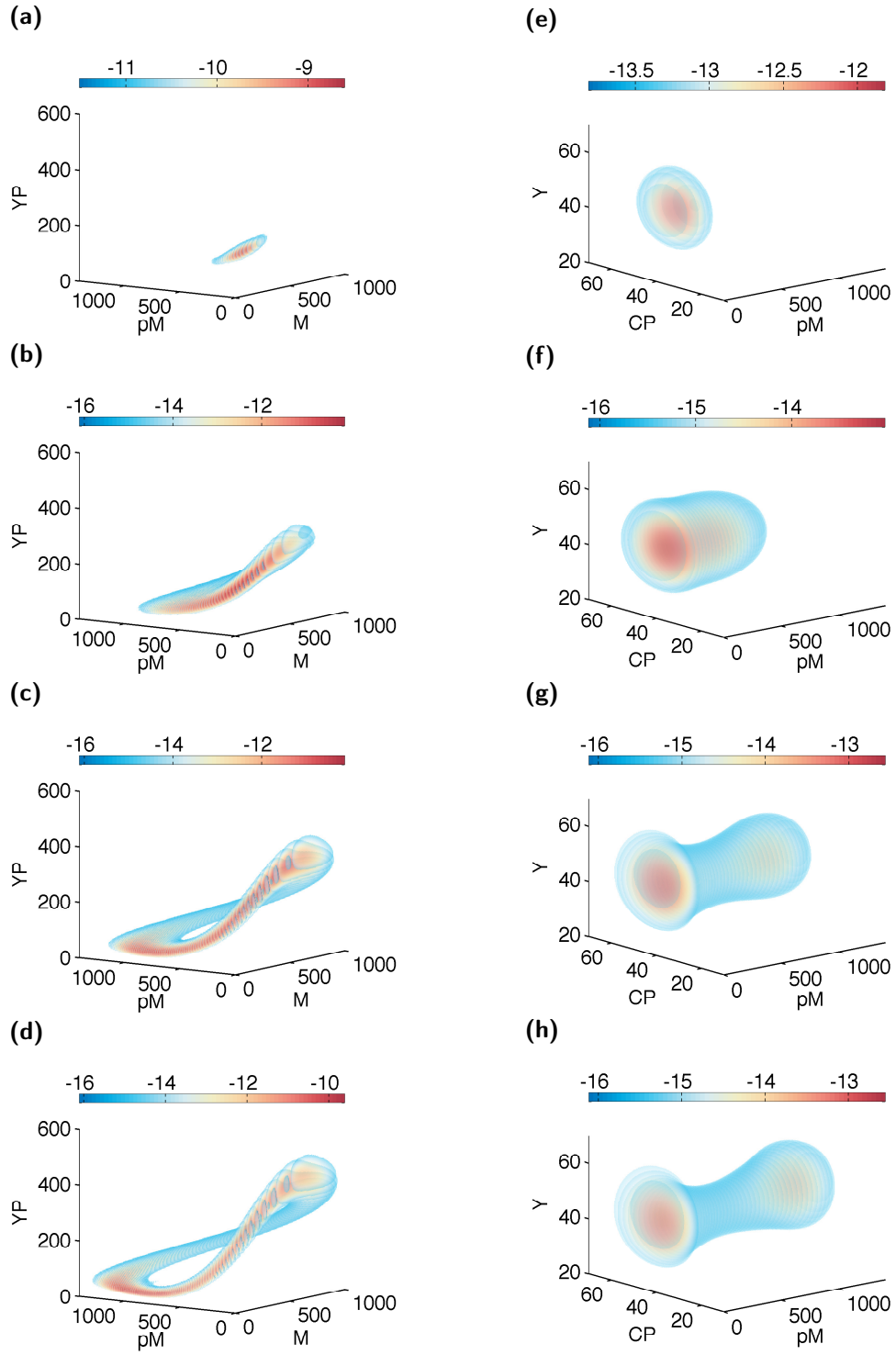

Figure 4.6: *Visualisation of the bifurcation structure of the stochastic cell cycle model.* **(a)–(d)** *Marginal steady state distributions of the phosphorylated cyclin (YP), the inactive MPF (pM) and the active MPF (M).* **(e)–(h)** *Marginal stationary distributions of the cyclin (Y), the phosphorylated cdc2 (CP) and the inactive MPF (pM). Each column corresponds to the same value of the bifurcation parameter k_1, which from the top to the bottom are* 0.25, 0.3, 0.35 *and* 0.4, *respectively. All figures show* $\log_2$ *of the marginal steady state distribution for better visualisation.*

4.2.3 Sensitivity analysis

Related to bifurcation analysis, stochastic sensitivity analysis quantifies the dependence of certain quantities of interest on continuous changes in model parameters. The sensitivity indicator for an observable Θ and an underlying parameter k is generally defined as (Savageau, 1971)

$$S(\Theta) = \frac{\mathrm{d}\Theta(k)}{\mathrm{d}k}\frac{k}{\Theta(k)}.$$

The indicator is often computed as a finite difference

$$S(\Theta) \approx \frac{\|\Theta(k+\Delta k) - \Theta(k)\|}{\Delta k}\frac{k}{\|\Theta(k)\|}, \tag{4.12}$$

where $\|\cdot\|$ represent a suitable norm, and Δk is a change in the value of k. The model is sensitive to the parameter Θ if $S(\Theta)$ has a large value. For deterministic models, the observable Θ is usually the steady-state mean concentration. In the stochastic setting, we have more options.

For example, let us consider the cell cycle model described in Figure 4.4. We will study the sensitivity with respect to the parameter k_1 for the following three observables Θ: mean concentration of the MPF (Θ_m), the oscillation amplitude (Θ_a) and the steady state distribution (Θ_p). In the case of the oscillation amplitude, we quantify Θ_a as the probability that the molecular population of the active MPF exceeds 400.

In the TPA framework, Θ_m and Θ_a are evaluated for all considered values of k_1, and the corresponding tensor operations are similar to (4.4), which admit linear complexity with N. More importantly, the tensor-structured data enable direct comparison of two steady state probabilities in the 6-dimensional state space. Namely, the norm $\|\Theta_p(k_1+\Delta k_1) - \Theta_p(k_1)\|_2$, needed in (4.12) with $\Theta_p(k_1) = \pi(\boldsymbol{x}|k_1)$, can be directly computed. The results are plotted in Figure 4.7. We observe that, within the con-

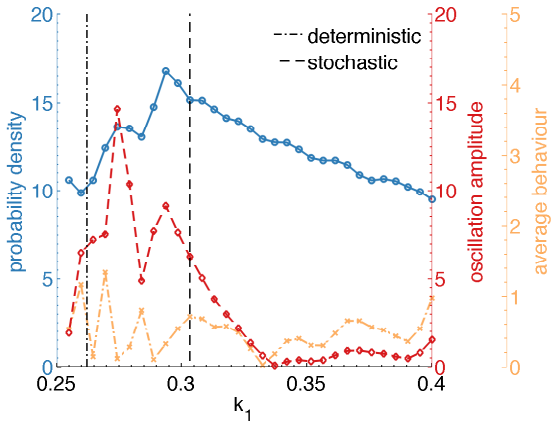

Figure 4.7: *Sensitivity indicators $S(\Theta)$ calculated by (4.12) for 32 equidistant nodes within the range $[0.25, 0.4]$ of parameter k_1. Three observables are considered: stationary distribution Θ_p (blue), oscillation amplitude Θ_a (red) and average number Θ_m of active MPF (orange). The dot-dashed and dashed lines indicate parameter values for which Figures 4.5f and 4.5h were computed, i.e. $k_1 = 0.2694$ (deterministic bifurcation point) and $k_1 = 0.3032$, respectively.*

sidered range of $k_1 \in [0.25, 0.4]$, the sensitivity in the steady state distribution (blue curve) dominates in magnitude over $S(\Theta_m)$ and $S(\Theta_a)$. The steady state distribution contains global information about the system and is more sensitive to parameter changes than derived quantities, like Θ_m and Θ_a.

4.3 Robustness analysis

Reaction networks are subject to extrinsic noise which is manifested by fluctuations of parameter values (Paulsson, 2004). This extrinsic noise originates from interactions of the modelled system with other stochastic processes in the cell or its surrounding environment. In this section, we apply the tensor-structured framework to facilitate the inclusion of extrinsic fluctuations into the robustness analysis of stochastic networks.

For a reaction system as in (2.1), we consider the copy numbers $x_1, x_2, \ldots, x_N$ as intrinsic variables and reaction rates $k_1, k_2, \ldots, k_K$ as extrinsic variables. Total stochasticity is quantified by the stationary distribution of the intrinsic variables, $\pi(\boldsymbol{x}|\boldsymbol{k})$,

$\boldsymbol{x} \in \Omega_{\boldsymbol{x}}$, $\boldsymbol{k} \in \Omega_{\boldsymbol{k}}$. We assume that the invariant probability density of extrinsic variables, $q(\boldsymbol{k})$, does not depend on the values of intrinsic variables $\boldsymbol{x}$. Then the stationary probability distribution of intrinsic variables may be implied from the law of total probability as

$$\pi(\boldsymbol{x}) = \int_{\Omega_{\boldsymbol{k}}} \pi(\boldsymbol{x}|\boldsymbol{k})\, q(\boldsymbol{k})\, \mathrm{d}\boldsymbol{k}, \tag{4.13}$$

where $\Omega_{\boldsymbol{k}}$ is the parameter space and $\pi(\boldsymbol{x}|\boldsymbol{k})$ represents the invariant density of intrinsic variables conditioned on constant values of kinetic parameters.

If distributions $q(\boldsymbol{k})$ of extrinsic variables can be determined from high quality experimental data then the stationary density can be computed directly by (4.13). If not, the TPA framework enables us to test the behaviour of GRNs for different hypothesis about the distribution of the extrinsic variables.

In the TPA framework, the high-dimensional integration in (4.13) can be performed efficiently using the tensor mode products (see Section § 3.2.1). Let $\boldsymbol{p}^{\boldsymbol{k}}_{\boldsymbol{h},t} \in \mathbb{R}^{n_1 \times n_N \times m_1 \times m_K}$ be the parametric steady state distribution stored in a tensor format computed using the method described in Section § 3.5. The parametric distributions are often determined independently (Scott et al., 2006), it is appropriate to assume that the extrinsic distribution has a separated form, $q(\boldsymbol{k}) = q_1(k_1) \cdots q_K(k_K)$. In this way, upon discretisation, the resulting tensor has the rank-one form

$$\boldsymbol{q} = \boldsymbol{q}_1 \otimes \boldsymbol{q}_2 \otimes \cdots \otimes \boldsymbol{q}_K,$$

where the K-dimensional tensor $\boldsymbol{q} \in \mathbb{R}^{m_1 \times \cdots \times m_K}$ is the nodal evaluation of function $q(\boldsymbol{k})$, and $\boldsymbol{q}_j \in \mathbb{R}^{m_j}$, $j = 1, 2, \ldots, K$, are one-dimensional tensors (vectors) that discretise functions $q_j(k_j)$. Then, the integral (4.13) can be performed via successive tensor mode products as

$$\boldsymbol{p}^{\boldsymbol{x}}_{\boldsymbol{h},t} = \boldsymbol{p}^{\boldsymbol{k}}_{\boldsymbol{h},t} \times_{N+1} \boldsymbol{q}_1 \times_{N+2} \boldsymbol{q}_2 \times_{N+3} \cdots \times_{N+K} \boldsymbol{q}_K, \tag{4.14}$$

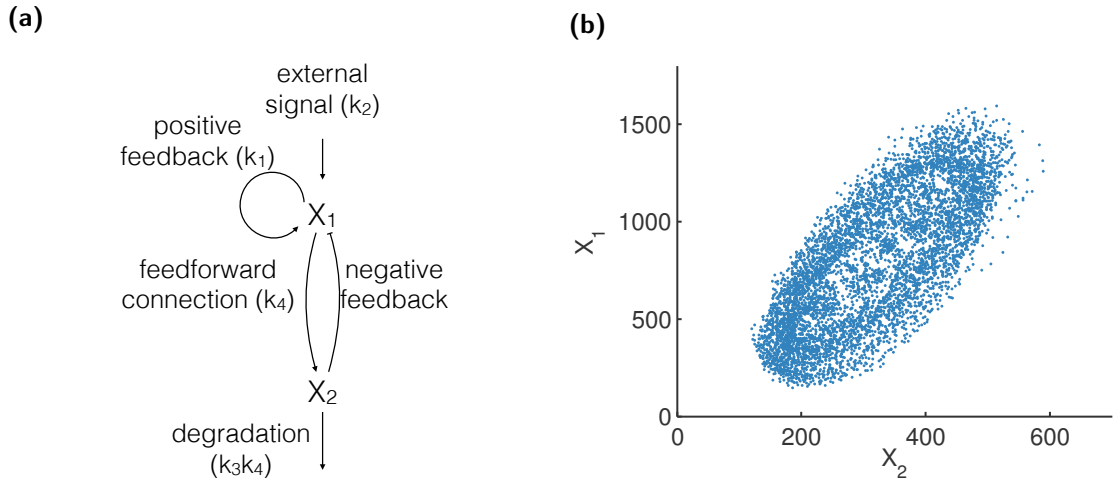

Figure 4.8: *Illustration of the FitzHugh-Nagumo reaction system.* **(a)** *Schematic of the model.* **(b)** *Gillespie simulation samples under a strict parameter set in the original reference (Tsai et al., 2008).*

where the N-dimensional tensor $\boldsymbol{p}^{\boldsymbol{x}}_{\boldsymbol{h},t} \in \mathbb{R}^{n_1 \times \cdots \times n_N}$ is the tensorised version of function $\pi(\boldsymbol{x})$ defined in (4.13).

If the parametric distribution $\boldsymbol{p}^{\boldsymbol{k}}_{\boldsymbol{h},t}$ is given as a full (not decomposed) tensor, the computational complexity of performing operations (4.14) is of order $\mathcal{O}(n^{N+K}K)$, since each mode product requires $\mathcal{O}(n^{N+K})$ operations. But following the tensor approach described in Section § 3.5, the parametric distribution $\boldsymbol{p}^{\boldsymbol{k}}_{\boldsymbol{h},t}$ is computed and stored in the QTT format. Under such format, the complexity of performing each mode product, from Proposition 4.14(v), is $\mathcal{O}(n^2R)$, and the total computational cost of (4.14) is $\mathcal{O}(n^2RK)$. In what follows, we exemplify the above tensor robustness analysis framework to a reaction network for neuron excitations.

4.3.1 Extrinsic noise in FitzHugh-Nagumo model

We consider the effect of extrinsic fluctuations on an activator-inhibitor oscillator with simple negative feedback loop: the FitzHugh-Nagumo neuron model which is presented in Figure 4.8a. Self-autocatalytic positive feedback loop activates the X_1 molecules, which are further triggered by the external signal. The species X_2 is

enhanced by the feedforward connection and it acts as an inhibitor that turns off the signalling (Tsai et al., 2008). The elementary reactions are given by

$$\begin{aligned} X_1 &\xrightarrow{(X_1-k_1V)(V-X_1)} 2X_1, \\ X_1+X_2 &\xrightarrow{1} X_2, \\ \emptyset &\xrightarrow{k_2V} X_1, \\ X_2 &\xrightarrow{k_3k_4} \emptyset, \\ X_1 &\xrightarrow{k_4} X_1+X_2, \end{aligned} \tag{4.15}$$

where k_j, $j = 1,2,3,4$, are (extrinsic) reaction parameters. Let x_i be the molecular number of X_i, $i = 1,2$, and the propensity functions are given by

$$\alpha_1 = x_1(x_1 - k_1V)(V - x_1), \quad \alpha_2 = x_1x_2, \quad \alpha_3 = k_2V, \quad \alpha_4 = k_3k_4x_2, \quad \alpha_5 = k_4x_1.$$

We performed the tensor-structured simulations of steady state distributions that vary all four parameters within a prescribed parameter space. The parameter ranges are given in Table 4.4. The parametric distributions were approximated by the stationary solution of the parametric CFPE (3.101). The simulation method and performance for solving the 6-dimensional equation are described in Section § 3.5.3 (summarised on the last row of Table 3.2). The total number of the parameter samples is 2.7×10^8 and the total simulation time is 37 minutes. Thus the CPU timer per parameter is 8.3×10^{-6} second, why is far more efficient than the existing stochastic simulation method for fixed parameter values.

Once the tensor-structured solution is computed and stored, we can perform necessary tensor operations to perform robustness analysis. In our computational examples, we assume that $\pi(\boldsymbol{k}) = q_1(k_1)q_2(k_1)\dots q_M(k_M)$, i.e. the invariant distributions

of rate constants $k_1, k_2, \ldots, k_M$ are independent. Then (4.13) reads as follows

$$\pi(\boldsymbol{x}) = \int_{\Omega_{\boldsymbol{k}}} \pi(\boldsymbol{x}|\boldsymbol{k}) q_1(k_1) \cdots q_M(k_M) \mathrm{d}\boldsymbol{k}. \tag{4.16}$$

Extrinsic variability in the FitzHugh-Nagumo system is studied in four prototypical cases of q_i, $i = 1,2,\ldots,M$: (i) Dirac delta, (ii) normal, (iii) uniform, and (iv) bimodal distributions, as shown in Figure 4.9a. Since these distributions for perturbation from the mean have zero mean, the extrinsic noise is not biased. We can then use this information about extrinsic noise to simulate the stationary probability distribution of intrinsic variables by (4.16).

When the extrinsic noise is omitted, the inhibited and excited states are linked by a volcano-shaped oscillatory probability distribution (Figure 4.9b). At the inhibited state, X_1 molecules first get activated from the positive feedback loop, and then excite X_2 molecules by feedforward control. The delay between the excitability of the two molecular species gives rise to the path (solid line) describing switching from the inhibited state to the excited state (Figure 4.9b). If the normal or uniform noise are introduced to the extrinsic variables, then the path becomes straighter (Figure 4.9c and Figure 4.9d). This suggests that, once X_1 molecules get excited or inhibited, X_2 molecules require less time to respond.

Reaction networks with stronger negative feedback regulation gain higher potential to reduce the stochasticity. This argument has been both theoretically anal-

Table 4.4: *Properties of molecular and rate variables in the FitzHugh-Nagumo model.*

Type	Notation	Range	No. of nodes
Species	X_1	$[0,1800]$	256
Species	X_2	$[0,700]$	256
Rate	k_1	$[0.17,0.23]$	128
Rate	k_2	$[0.0952,0.1288]$	128
Rate	k_3	$[2.125,2.875]$	128
Rate	k_4	$[0.0892,0.1207]$	128

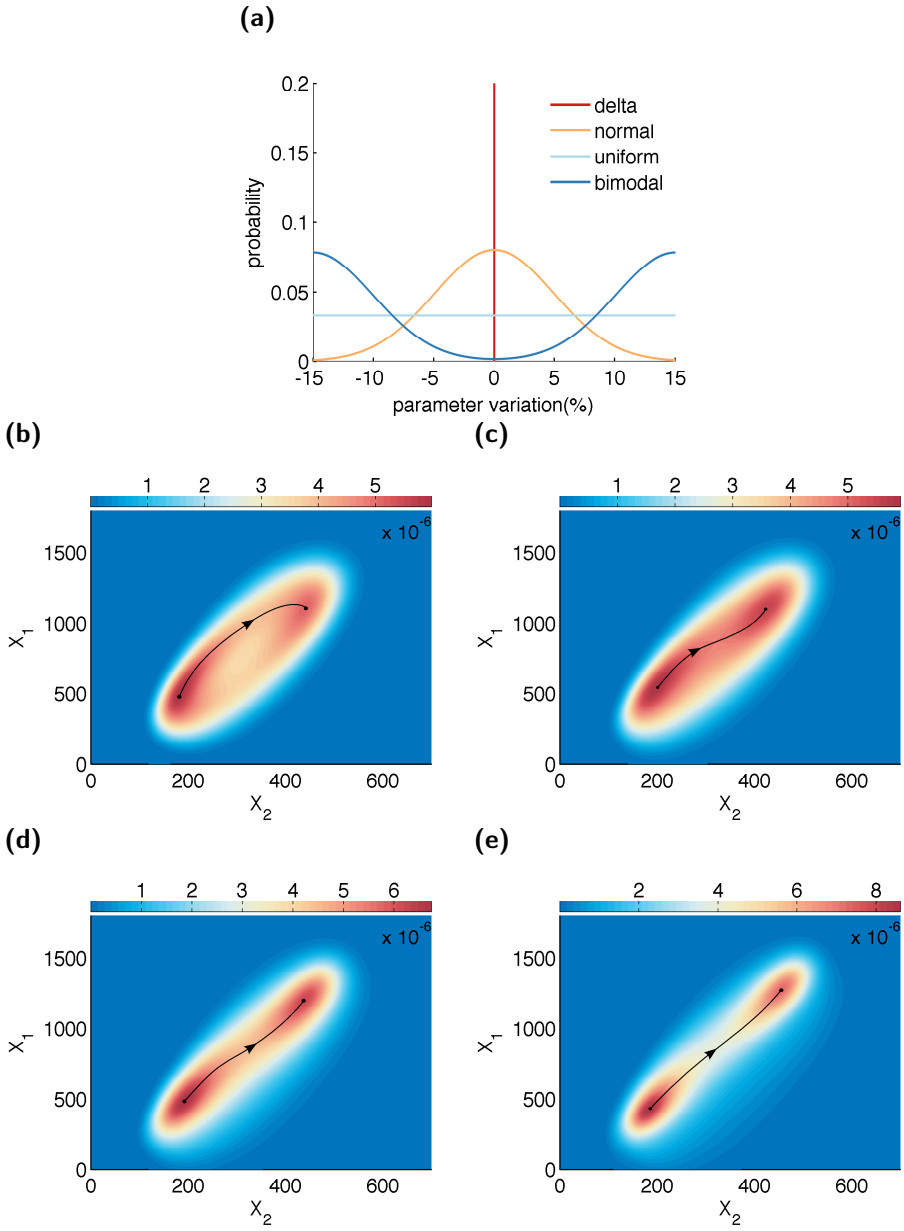

Figure 4.9: *Robustness analysis of the FitzHugh-Nagumo reaction system.* **(a)** *Four types of distributions of the extrinsic noise applied to model parameters.* **(b) – (e)** *Steady state behaviour of the FitzHugh-Nagumo model with* **(b)** *constant reaction rates,* **(c)** *normal,* **(d)** *uniform, and* **(e)** *bimodal distribution of the model parameters. The distributions were computed following formula* (4.16)*. The directed black curves track the ridges of the stationary distribution, and depict the "most-likely" transition paths from the inhibited state to the excited state.*

ysed (Thattai & Van Oudenaarden, 2001; Swain, 2004), and experimentally tested for a plasmid-borne system (Becskei & Serrano, 2000). We have shown that the extrinsic noise reduces the delay caused by the feedback loop (Figure 4.9c). If we

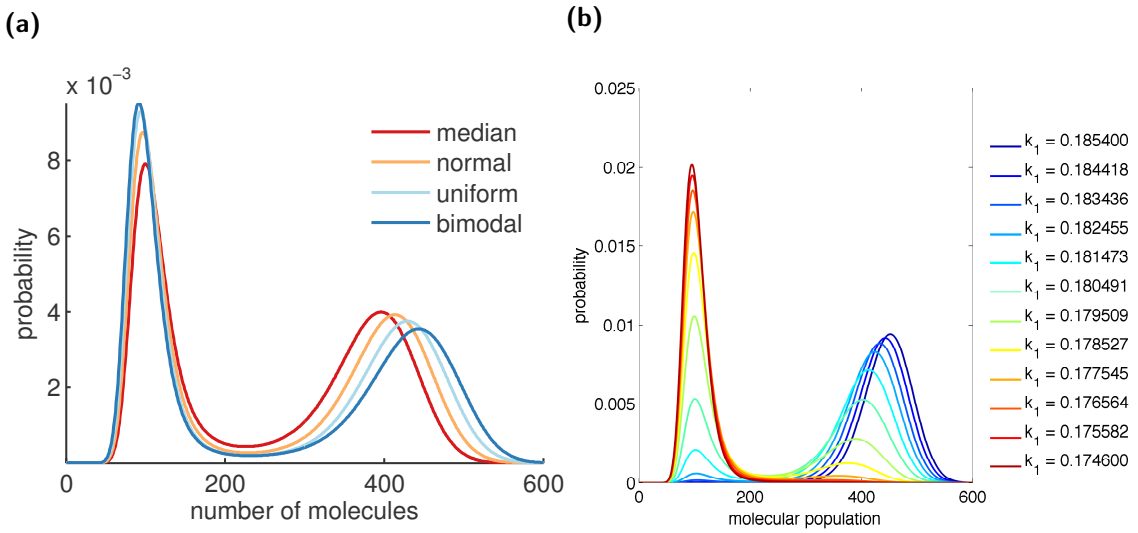

Figure 4.10: *Robustness analysis of the Schlögl model.* **(a)** *Schematic of the model.* **(b)** *Gillespie simulation samples under a strict parameter set in the original reference (Tsai et al., 2008).*

further increase the variability of the extrinsic noise, then the delay caused by the feedback loop is further reduced (Figure 4.9d). In the case of the bimodal distribution of extrinsic fluctuations, the most-likely path linking the inhibited and excited states even shrinks into an almost straight line (Figure 4.9e). This means that, for the same level of the inhibitor X_2, the number of the activator X_1 is lower, i.e. the presented robustness analysis shows that the behaviour of stochastic GRNs with negative feedback regulation can benefit from the extrinsic noise.

4.3.2 Constructive effect of extrinsic noise

The previous section showed one case where the extrinsic noise enhance the negative feedback regulation, and in the current section we present the case where the extrinsic noise can have constructive effect in the stochastic models.

We will re-consider the Schlögl reaction system described in Section § 4.1.2, and use the parametric steady state distributions computed in the tensor format in Section § 3.5.3. The extrinsic stochasticity is incorporated using Equation (4.16), and we adopt the four prototypical types of parametric distributions in Figure 4.9a. The

steady state distributions in presence of extrinsic noise are plotted in Figure 4.10a, and in Figure 4.10b, we plot the distributions for different parameter (k_1) values but without the parametric variations.

We observe that the distribution of the intrinsic variable is sensitive to the change of the fixed parameter values (Figure 4.10b), but the bistable distribution does not get attenuated with increasing variation level of the mode parameters (Figure 4.10a). This implies that, although the characteristics of intrinsic noise can be sensitive to the parameter values, the extrinsic noise in parameters can be constructive.

The reality of extrinsic noise might be far more complex than what Figure 4.9a shows, but our analysis suggests intrinsic stochasticity enhances the robustness of the cellular function. Deterministic models in comparison would be less robust, as they completely neglect the intrinsic variability that buffers the external noise. As a result, parameter fluctuations always act destructively, and we are not able to speculate how regulation networks cope with, or exploit, the environmental uncertainties with deterministic models. On the other hand, stochastic systems modelled with constant kinetic rates may explain one possibility, but are not sufficient to prove the characterized variability is intrinsically driven, because both the intrinsic and extrinsic sources of noise may combine constructively (Figure 4.10a).

"A theory has only the alternative of being right or wrong. A model has a third possibility: it may be right, but irrelevant."

Eigen (1973)

5 Conclusion

5.1 Key contributions

In contrast to the CME, the CFPE admits more established analytical and numerical manipulations for various types of the reaction networks. But the validity of the CFPE-based analysis requires a concrete error analysis of the CFPE. Thus in Section § 2.2, we

C1. Derived the convergence rate between the CME distribution and the CFPE distribution in the thermodynamic limit for systems without and with the detailed balance condition.

This was proved by a perturbation analysis of the system-size expansion and Hermite polynomial expansion of the two equations.

For systems partially in the thermodynamic limit, we showed a type of noise-induced multi-stability in Section § 2.3, where the CFPE distribution may not cap-

ture the multi-modality of the CME distribution. We proposed two alternative approaches to address this:

C2. Proposed the quasi-stationary approximation (QSA) of the CFPE to recovery the noise-induced multi-stability;

C3. Proposed a hybrid scheme which partially approximate the CME by the CFPE to recover the noise-induced multi-stability, and developed its Monte Carlo simulation algorithm.

The results of **C2** have been published in a co-authored paper, Duncan et al. (2015), and my contribution includes creating the idea of the QSA, and performing Finite element and Monte-Carlo simulation of the model equations for the example systems, as well as editing the manuscript.

Numerical solution of the CFPE requires its domain of definition to be truncated into a bounded domain, and thus we

C4. Applied the theories for the generalised principle eigenvalue problems to show the convergence of the principle eigen-pair of the CFPE operator in a truncated domain to the original domain.

After domain truncation, we

C5. Discretised the CFPE using a monotone difference scheme where the stiffness matrix admits explicit tensor constructions, and proved the monotonicity of the scheme.

Comparing to the classic monotone difference scheme, the proposed scheme is tensor-friendly, i.e., the stiffness matrix can be constructed from the tensor-formatted components. In our case, the QTT format is mainly used, and thus the work is followed by

C6. Derived the explicit QTT representations of the CFPE operator and the parametric CFPE operator; and derived rank estimate of the CFPE operator to be $\mathcal{O}(M^2N^5l)$ and its parametric operator to be $\mathcal{O}(M^2N^5l + K^3l)$.

And in Section § 3.4.1, we

C7. Proved the existence of low-rank QTT representation for a class of the stationary distributions that can be approximated by a small sum of Gaussian functions centred away from the boundary.

As mentioned in Section § 3.4.1.2, other authors have derived rank estimates for various analytical functions, and can be used for a priori estimate of a wider class of the stationary distributions.

The existence of low-rank approximations of the final solutions validates the use of algorithm to search for QTT-formatted solutions. Our contribution involves

C8. Proposed an adaptive inverse power methods to solve the stationary (parametric) CFPEs in the tensor format.

In comparison, the classic inverse power method with constant shift may not converge because the tensor linear solver is less robust, and global convergence has been proved for SPD systems only. In Section § 3.4.3.1, we

C9. Proved the global convergence of the AMEn algorithm for positive definite system with a sufficiently large shift.

This provides a justification to use the adaptive scheme, where the conditioning of the tensor matrix (CFPE operator) is improved by the adaptively chosen shift values.

Another contribution is applying the tensor methods in the parametric analysis of the GRNs, and this includes

C10. Applied the tensor framework to the parameter inference of the GRNs.

C11. Applied the tensor framework to analyse the stochastic bifurcation and parameter sensitivity of the GRNs.

C12. Applied the tensor framework to the sensitivity analysis of the GRNs.

The work has been published in a primarily-authored paper, Liao et al. (2015).

5.2 Future directions

One key information revealed in deriving the error between the CME and the CFPE distributions, but not yet fully explored, is that the number (or summation index) of the Hermite polynomials required to express the coefficients on order $V^{-2/j}$ is $3j$. This implies that it is possible to combine the Hermite polynomials, whose summation indices are bounded by $3j$, to yield an order $V^{-2/j}$ approximation of the CME distribution. Finding such an approximation boils down to determine the (time-dependent) coefficients of these Hermite polynomials. That is the CME can be reduced to a problem of determining the values of $\binom{3j+N-1}{3j}$ coefficients, and the curse of dimensionality is avoided naturally.

Also it is worth noting that the hybrid CME-CFPE equation proposed in section 2.4.5 is tensor-friendly, i.e., the operator of the equation can be constructed explicitly using the tensor-formatted components. This can be used to simulate larger systems in presence of multiple population scales.

On the application side, we have shown that the TPA framework has capacity to analyse very different types of stochastic models of GRNs. The framework can also be easily equipped with a wider range of parametric tools from deterministic models. A future extension to a broader class of stochastic models, for example the spatial-temporal stochastic models, would be of interest.

References

Acar, E., Aykut-Bingol, C., Bingol, H., Bro, R., & Yener, B. (2007). Multiway analysis of epilepsy tensors. *Bioinformatics*, 23(13), i10–i18.

Adalsteinsson, D., McMillen, D., & Elston, T. C. (2004). Biochemical network stochastic simulator (bionets): software for stochastic modeling of biochemical networks. *BMC Bioinformatics*, 5(1), 24.

Affleck, I., Kennedy, T., Lieb, E. H., & Tasaki, H. (1987). Rigorous results on valence-bond ground states in antiferromagnets. *Physical Review Letters*, 59(7), 799.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (1997). *Molecular Biology of the Cell*. Garland Science, New York.

Alfa, A. S., Xue, J., & Ye, Q. (2002). Entrywise perturbation theory for diagonally dominant M–matrices with applications. *Numerische Mathematik*, 90(3), 401–414.

Alter, O. & Golub, G. H. (2004). Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. *Proceedings of the National Academy of Sciences of the United States of America*, 101(47), 16577–16582.

Alter, O. & Golub, G. H. (2005). Reconstructing the pathways of a cellular system from genome-scale signals by using matrix and tensor computations. *Proceedings of the National Academy of Sciences of the United States of America*, 102(49), 17559–17564.

Ammar, A., Cueto, E., & Chinesta, F. (2012). Reduction of the chemical master equation for gene regulatory networks using proper generalized decompositions. *International Journal for Numerical Methods in Biomedical Engineering*, 28(9), 960–973.

Balaban, N. Q., Merrin, J., Chait, R., Kowalik, L., & Leibler, S. (2004). Bacterial persistence as a phenotypic switch. *Science*, 305(5690), 1622–1625.

Balázsi, G., van Oudenaarden, A., & Collins, J. J. (2011). Cellular decision making and biological noise: from microbes to mammals. *Cell*, 144(6), 910–925.

Barkai, N. & Leibler, S. (2000). Biological rhythms: Circadian clocks limited by noise. *Nature*, 403(6767), 267–268.

Becskei, A. & Serrano, L. (2000). Engineering stability in gene networks by autoregulation. *Nature*, 405(6786), 590–593.

Bellman, R. E. (1961). *Adaptive control processes: a guided tour*, volume 4. Princeton University Press, Princeton.

Berestycki, H., Nirenberg, L., & Varadhan, S. S. (1994). The principal eigenvalue and maximum principle for second-order elliptic operators in general domains. *Communications on Pure and Applied Mathematics*, 47(1), 47–92.

Berestycki, H. & Rossi, L. (2015). Generalizations and properties of the principal eigenvalue of elliptic operators in unbounded domains. *Communications on Pure and Applied Mathematics*, 68(6), 1014–1065.

Beylkin, G. & Mohlenkamp, M. J. (2005). Algorithms for numerical analysis in high dimensions. *SIAM Journal on Scientific Computing*, 26(6), 2133–2159.

Biancalani, T., Dyson, L., & McKane, A. J. (2014). Noise-induced bistable states and their mean switching time in foraging colonies. *Physical Review Letters*, 112(3), 038101.

Bowsher, C. G. & Swain, P. S. (2012). Identifying sources of variation and the flow of information in biochemical networks. *Proceedings of the National Academy of Sciences*, 109(20), E1320–E1328.

Carasso, A. (1969). Finite-difference methods and the eigenvalue problem for non-selfadjoint Sturm-Liouville operators. *Mathematics of Computation*, 23(108), 717–729.

Carroll, J. D. & Chang, J.-J. (1970). Analysis of individual differences in multi-dimensional scaling via an n-way generalization of "Eckart-Young" decomposition. *Psychometrika*, 35(3), 283–319.

Cohen, J. E. (2004). Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS Biology*, 2(12), e439.

Crudu, A., Debussche, A., & Radulescu, O. (2009). Hybrid stochastic simplifications for multiscale gene networks. *BMC Systems Biology*, 3(1), 89.

De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000a). A multilinear singular value decomposition. *SIAM journal on Matrix Analysis and Applications*, 21(4), 1253–1278.

De Lathauwer, L., De Moor, B., & Vandewalle, J. (2000b). On the best rank-1 and rank-($r_1, r_2, ..., r_n$) approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4), 1324–1342.

De Silva, V. & Lim, L.-H. (2008). Tensor rank and the ill-posedness of the best low-rank approximation problem. *SIAM Journal on Matrix Analysis and Applications*, 30(3), 1084–1127.

Deif, A. (1995). Rigorous perturbation bounds for eigenvalues and eigenvectors of a matrix. *Journal of Computational and Applied Mathematics*, 57(3), 403–412.

Delbrück, M. (1940). Statistical fluctuations in autocatalytic reactions. *The Journal of Chemical Physics*, 8(1), 120–124.

Demmel, J. W. (1997). *Applied numerical linear algebra*. SIAM.

Doedel, E., Champneys, A., Fairgrieve, T., Kuznetsov, Y., Sandstede, B., & Wang, X. (1997). Auto 97: Continuation and bifurcation software for ordinary differential equations, user's manual. *Center for research on parallel computing, California Institute of Technology, Pasadena*.

Dolgov, S. & Khoromskij, B. (2013a). Simultaneous state-time approximation of the chemical master equation using tensor product formats. *arXiv preprint arXiv:1311.3143*.

Dolgov, S. & Khoromskij, B. (2013b). Two-level QTT-Tucker format for optimized tensor calculus. *SIAM Journal on Matrix Analysis and Applications*, 34(2), 593–623.

Dolgov, S. & Khoromskij, B. (2015). Simultaneous state-time approximation of the chemical master equation using tensor product formats. *Numerical Linear Algebra with Applications*, 22(2), 197–219.

Dolgov, S. & Khoromskij, B. N. (2012). *Tensor-product approach to global time-space-parametric discretization of chemical master equation*. Technical report, Max Planck Institute for Mathematics in the Sciences.

Dolgov, S., Khoromskij, B. N., & Oseledets, I. V. (2012). Fast solution of parabolic problems in the tensor train/quantized tensor train format with initial application to the Fokker–Planck equation. *SIAM Journal on Scientific Computing*, 34(6), A3016–A3038.

Dolgov, S. V. (2013). TT-GMRES: solution to a linear system in the structured tensor format. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 28(2), 149–172.

Dolgov, S. V., Khoromskij, B. N., Oseledets, I. V., & Savostyanov, D. V. (2014). Computation of extreme eigenvalues in higher dimensions using block tensor train format. *Computer Physics Communications*, 185(4), 1207–1216.

Dolgov, S. V. & Savostyanov, D. V. (2013a). Alternating minimal energy methods for linear systems in higher dimensions. Part I: SPD systems. *arXiv preprint arXiv:1301.6068*.

Dolgov, S. V. & Savostyanov, D. V. (2013b). Alternating minimal energy methods for linear systems in higher dimensions. Part II: Faster algorithm and application to nonsymmetric systems. *arXiv preprint arXiv:1304.1222*.

Douglass, J. K., Wilkens, L., Pantazelou, E., Moss, F., et al. (1993). Noise enhancement of information transfer in crayfish mechanoreceptors by stochastic resonance. *Nature*, 365(6444), 337–340.

Duncan, A., Liao, S., Vejchodskỳ, T., Erban, R., & Grima, R. (2015). Noise-induced multistability in chemical systems: Discrete versus continuum modeling. *Physical Review E*, 91(4), 042111.

Eckart, C. & Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3), 211–218.

Eigen, M. (1973). The origin of biological information. In *The Physicist's Conception of Nature* (pp. 594–632). Springer.

English, B. P., Min, W., van Oijen, A. M., Lee, K. T., Luo, G., Sun, H., Cherayil, B. J., Kou, S., & Xie, X. S. (2006). Ever-fluctuating single enzyme molecules: Michaelis-menten equation revisited. *Nature chemical biology*, 2(2), 87–94.

Erban, R. & Chapman, S. J. (2009). Stochastic modelling of reaction–diffusion processes: algorithms for bimolecular reactions. *Physical Biology*, 6(4), 046001.

Erban, R., Chapman, S. J., Kevrekidis, I. G., & Vejchodskỳ, T. (2009). 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.

Espig, M. & Hackbusch, W. (2012). A regularized newton method for the efficient approximation of tensors represented in the canonical tensor format. *Numerische Mathematik*, 122(3), 489–525.

Fange, D., Berg, O. G., Sjöberg, P., & Elf, J. (2010). Stochastic reaction-diffusion kinetics in the microscopic limit. *Proceedings of the National Academy of Sciences*, 107(46), 19820–19825.

Fannes, M., Nachtergaele, B., & Werner, R. F. (1992a). Finitely correlated states on quantum spin chains. *Communications in Mathematical Physics*, 144(3), 443–490.

Fannes, M., Nachtergaele, B., & Werner, R. F. (1992b). Ground states of VBS models on Cayley trees. *Journal of Statistical Physics*, 66(3-4), 939–973.

Feng, X., Glowinski, R., & Neilan, M. (2013). Recent developments in numerical methods for fully nonlinear second order partial differential equations. *SIAM Review*, 55(2), 205.

Flegg, M. B., Chapman, S. J., & Erban, R. (2011). The two-regime method for optimizing stochastic reaction–diffusion simulations. *Journal of the Royal Society Interface*, (pp. rsif20110574).

Gary, J. (1965). Computing eigenvalues of ordinary differential equations by finite differences. *Mathematics of Computation*, 19(91), 365–379.

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

Gillespie, D. (1977). Exact stochastic simulation of coupled chemical reactions. *Journal of Physical Chemistry*, 81, 2340–2361.

Gillespie, D. (1992). A rigorous derivation of the chemical master equation. *Physica A*, 188, 404–425.

Gillespie, D. (2000). The chemical Langevin equation. *Journal of Physical Chemistry*, 113, 297.

Gillespie, D. (2001). Approximate accelerated stochastic simulation of chemically reacting systems. *Journal of Physical Chemistry*, 115, 1716–1733.

Gillespie, D. (2007). Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58, 35–55.

Gillespie, D. T. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4), 403–434.

Gillespie, D. T., Hellander, A., & Petzold, L. R. (2013). Perspective: Stochastic algorithms for chemical kinetics. *The Journal of Chemical Physics*, 138(17), 170901.

Golub, G. H. & Ye, Q. (2000). Inexact inverse iteration for generalized eigenvalue problems. *BIT Numerical Mathematics*, 40(4), 671–684.

Grasedyck, L. (2010). Hierarchical singular value decomposition of tensors. *SIAM Journal on Matrix Analysis and Applications*, 31, 2029 – 2054.

Grima, R., Schmidt, D., & Newman, T. (2012). Steady-state fluctuations of a genetic feedback loop: An exact solution. *The Journal of Chemical Physics*, 137(3), 035104.

Grima, R., Thomas, P., & Straube, A. V. (2011). How accurate are the nonlinear chemical Fokker-Planck and chemical Langevin equations? *The Journal of Chemical Physics*, 135(8), 084103.

Gutenkunst, R. N., Waterfall, J. J., Casey, F. P., Brown, K. S., Myers, C. R., & Sethna, J. P. (2007). Universally sloppy parameter sensitivities in systems biology models. *PLoS Comput Biol*, 3(10), e189.

Hackbusch, W. (2013). *Multi-grid methods and applications*, volume 4. Springer Science.

Hackbusch, W., Khoromskij, B. N., & Tyrtyshnikov, E. E. (2005). Hierarchical kronecker tensor-product approximations. *Journal of Numerical Mathematics*, 13(2), 119–156.

Hackbusch, W. & Kühn, S. (2009). A new scheme for the tensor representation. *Journal of Fourier Analysis and Applications*, 15(5), 706–722.

Hamming, R. (1962). *Numerical methods for scientists and engineers*. New York: McGraw-Hill.

Hanggi, P., Grabert, H., Talkner, P., & Thomas, H. (1984). Bistable systems: Master equation versus Fokker-Planck modeling. *Physical Review A*, 29(1), 371.

Håstad, J. (1990). Tensor rank is NP-complete. *Journal of Algorithms*, 11(4), 644–654.

Hegland, M. & Garcke, J. (2011). On the numerical solution of the chemical master equation with sums of rank one tensors. *ANZIAM Journal*, 52, 628–643.

Hey, J. & Nielsen, R. (2007). Integration within the Felsenstein equation for improved Markov chain Monte Carlo methods in population genetics. *Proceedings of the National Academy of Sciences*, 104(8), 2785–2790.

Hillar, C. J. & Lim, L.-H. (2013). Most tensor problems are NP-hard. *Journal of the ACM (JACM)*, 60(6), 45.

Hitchcock, F. L. (1927). The expression of a tensor or a Polyadic as a sum of products. *Journal of Mathematics and Physics*, 6(1), 164–189.

Hitchcock, F. L. (1928). Multiple invariants and generalized rank of a P-Way matrix or tensor. *Journal of Mathematics and Physics*, 7(1), 39–79.

Holtz, S., Rohwedder, T., & Schneider, R. (2012). The alternating linear scheme for tensor optimization in the tensor train format. *SIAM Journal on Scientific Computing*, 34(2), A683–A713.

Hoops, S., Sahle, S., Gauges, R., Lee, C., Pahle, J., Simus, N., Singhal, M., Xu, L., Mendes, P., & Kummer, U. (2006). Copasi—a complex pathway simulator. *Bioinformatics*, 22(24), 3067–3074.

Horn, R. A. & Johnson, C. R. (2012). *Matrix analysis*. Cambridge University Press.

Horsthemke, W. & Lefever, R. (1984). Noise-induced transitions in physics, chemistry, and biology. *Noise-Induced Transitions: Theory and Applications in Physics, Chemistry, and Biology*, (pp. 164–200).

Hübener, R., Nebendahl, V., & Dür, W. (2010). Concatenated tensor network states. *New Journal of Physics*, 12(2), 025004.

Huckle, T., Waldherr, K., & Schulte-Herbrüggen, T. (2013). Computations in quantum tensor networks. *Linear Algebra and its Applications*, 438(2), 750–781.

Hunt, T. (1989). Embryology. Under arrest in the cell cycle. *Nature*, 342(6249), 483–484.

Jahnke, T. & Huisinga, W. (2007). Solving the chemical master equation for monomolecular reaction systems analytically. *Journal of Mathematical Biology*, 54(1), 1–26.

Jahnke, T. & Huisinga, W. (2008). A dynamical low-rank approach to the chemical master equation. *Bulletin of Mathematical Biology*, 70(8), 2283–2302.

Jaqaman, K. & Danuser, G. (2006). Linking data to models: data regression. *Nature Reviews Molecular Cell Biology*, 7(11), 813–819.

Jiang, Y., Pjesivac-Grbovic, J., Cantrell, C., & Freyer, J. P. (2005). A multiscale model for avascular tumor growth. *Biophysical Journal*, 89(6), 3884–3894.

Kazeev, V., Khammash, M., Nip, M., & Schwab, C. (2014). Direct solution of the chemical master equation using quantized tensor trains. *PLoS Computational Biology*, 10(3).

Kazeev, V., Reichmann, O., & Schwab, C. (2012). *hp-DG-QTT solution of high-dimensional degenerate diffusion equations*. ETH-Zürich.

Kazeev, V. A. & Khoromskij, B. N. (2012). Low-rank explicit QTT representation of the Laplace operator and its inverse. *SIAM Journal on Matrix Analysis and Applications*, 33(3), 742–758.

Kazeev, V. A., Khoromskij, B. N., & Tyrtyshnikov, E. E. (2013). Multilevel toeplitz matrices generated by tensor-structured vectors and convolution with logarithmic complexity. *SIAM Journal on Scientific Computing*, 35(3), A1511–A1536.

Keller, H. B. (1965). On the accuracy of finite difference approximations to the eigenvalues of differential and integral operators. *Numerische Mathematik*, 7(5), 412–419.

Kelvin, W. T. B. & Tait, P. G. (1873). *Elements of Natural Philosophy, Preface*. Clarendon Press.

Kepler, T. B. & Elston, T. C. (2001). Stochasticity in transcriptional regulation: origins, consequences, and mathematical representations. *Biophysical Journal*, 81(6), 3116–3136.

Khoromskaia, V. (2010). *Numerical solution of the Hartree-Fock equation by multilevel tensor-structured methods*. PhD thesis, TU Berlin.

Khoromskaia, V., Khoromskij, B., & Schneider, R. (2011). QTT representation of the hartree and exchange operators in electronic structure calculations. *Computational Methods in Applied Mathematics Comput. Methods Appl. Math.*, 11(3), 327–341.

Khoromskaia, V. & Khoromskij, B. N. (2014a). Grid-based lattice summation of electrostatic potentials by assembled rank-structured tensor approximation. *Computer Physics Communications*, 185(12), 3162–3174.

Khoromskaia, V. & Khoromskij, B. N. (2014b). Moller–Plesset (MP2) energy correction using tensor factorization of the grid-based two-electron integrals. *Computer Physics Communications*, 185(1), 2–10.

Khoromskaia, V. & Khoromskij, B. N. (2015). Tensor numerical methods in quantum chemistry: from Hartree–Fock to excitation energies. *Physical Chemistry Chemical Physics*, 17(47), 31491–31509.

Khoromskaia, V., Khoromskij, B. N., & Schneider, R. (2013). Tensor-structured factorized calculation of two-electron integrals in a general basis. *SIAM Journal on Scientific Computing*, 35(2), A987–A1010.

Khoromskij, B. & Oseledets, I. (2010a). Quantics-TT collocation approximation of parameter-dependent and stochastic elliptic PDEs. *Computational Methods in Applied Mathematics*, 10, 376 – 394.

Khoromskij, B. N. (2006). Structured rank-$(r_1,...,r_d)$ decomposition of function-related tensors in R^D. *Computational Methods in Applied Mathematics*, 6(2), 194–220.

Khoromskij, B. N. (2009a). *$O(d\log N)$-Quantics Approximation of N-d Tensors in High-Dimensional Numerical Modeling*. Technical report, Max Planck Institute for Mathematics in the Sciences.

Khoromskij, B. N. (2009b). Tensor-structured preconditioners and approximate inverse of elliptic operators in R^d. *Constructive Approximation*, 30(3), 599–620.

Khoromskij, B. N. (2011). $O(d\log N)$-quantics approximation of N-d tensors in high-dimensional numerical modeling. *Constructive Approximation*, 34(2), 257–280.

Khoromskij, B. N. (2012). Tensors-structured numerical methods in scientific computing: Survey on recent advances. *Chemometrics and Intelligent Laboratory Systems*, 110(1), 1–19.

Khoromskij, B. N. (2015). Tensor numerical methods for multidimensional PDEs: theoretical analysis and initial applications. *ESAIM: Proceedings and Surveys*, 48, 1–28.

Khoromskij, B. N. & Khoromskaia, V. (2009). Multigrid accelerated tensor approximation of function related multidimensional arrays. *SIAM Journal on Scientific Computing*, 31(4), 3002–3026.

Khoromskij, B. N., Khoromskaia, V., & Flad, H.-J. (2011). Numerical solution of the Hartree-Fock equation in multilevel tensor-structured format. *SIAM Journal on Scientific Computing*, 33(1), 45–65.

Khoromskij, B. N. & Oseledets, I. V. (2010b). *DMRG+ QTT approach to computation of the ground state for the molecular Schrödinger operator*. Technical report, Max Planck Institute for Mathematics in the Sciences.

Khoromskij, B. N. & Repin, S. I. (2015). A fast iteration method for solving elliptic problems with quasiperiodic coefficients. *Russian Journal of Numerical Analysis and Mathematical Modelling*, 30(6), 329–344.

Khoromskij, B. N. & Schwab, C. (2011). Tensor-structured Galerkin approximation of parametric and stochastic elliptic PDEs. *SIAM Journal on Scientific Computing*, 33(1), 364–385.

Kierzek, A. M. (2002). STOCKS: Stochastic kinetic simulations of biochemical systems with Gillespie algorithm. *Bioinformatics*, 18(3), 470–481.

Klümper, A., Schadschneider, A., & Zittartz, J. (1993). Matrix product ground states for one-dimensional spin-1 quantum antiferromagnets. *EPL (Europhysics Letters)*, 24(4), 293.

Kolda, T. & Bader, B. (2009). Tensor decompositions and applications. *SIAM Review*, 51, 455 – 500.

Komorowski, M., Costa, M. J., Rand, D. A., & Stumpf, M. P. (2011). Sensitivity, robustness, and identifiability in stochastic chemical kinetics models. *Proceedings of the National Academy of Sciences*, 108(21), 8645–8650.

Kramers, H. A. (1940). Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4), 284–304.

Krein, M. G. & Rutman, M. A. (1948). Linear operators leaving invariant a cone in a banach space. *Uspekhi Matematicheskikh Nauk*, 3(1), 3–95.

Kurtz, T. G. (1972). The relationship between stochastic and deterministic models for chemical reactions. *The Journal of Chemical Physics*, 57(7), 2976–2978.

Kurtz, T. G. (1978). Strong approximation theorems for density dependent Markov chains. *Stochastic Processes and Their Applications*, 6(3), 223–240.

Kuttler, J. R. (1970). Finite difference approximations for eigenvalues of uniformly elliptic operators. *SIAM Journal on Numerical Analysis*, 7(2), 206–232.

Larsson, S. & Thomée, V. (2008). *Partial differential equations with numerical methods*, volume 45. Springer Science & Business Media.

Layton, W. & Morley, T. (1987). On central difference approximations to general second order elliptic equations. *Linear Algebra and its Applications*, 97, 65–75.

Lee, C. H., Alpert, B. O., Sankaranarayanan, P., & Alter, O. (2012). GSVD comparison of patient-matched normal and tumor aCGH profiles reveals global copy-number alterations predicting glioblastoma multiforme survival. *PLoS One*, 7(1), e30098.

Liao, S., Vejchodskỳ, T., & Erban, R. (2015). Tensor methods for parameter estimation and bifurcation analysis of stochastic reaction networks. *Journal of Royal Society Interface*, 12(20150233).

Liberzon, D. & Brockett, R. W. (2000). Spectral analysis of Fokker–Planck and related operators arising from linear stochastic differential equations. *SIAM Journal on Control and Optimization*, 38(5), 1453–1467.

Lillacci, G. & Khammash, M. (2010). Parameter estimation and model selection in computational biology. *PLoS Comput Biol*, 6(3), e1000696.

Liu, Z., Pu, Y., Li, F., Shaffer, C. A., Hoops, S., Tyson, J. J., & Cao, Y. (2012). Hybrid modeling and simulation of stochastic effects on progression through the eukaryotic cell cycle. *The Journal of Chemical Physics*, 136(3), 034105.

Matus, P. & Rybak, I. (2004). Difference schemes for elliptic equations with mixed derivatives. *Computational Methods in Applied Mathematics*, 4(4), 494–505.

McQuarrie, D. (1967). Stochastic approach to chemical kinetics. *Journal of Applied Probability*, 4, 413 – 478.

McQuarrie, D. A. (1963). Kinetics of small systems. I. *The Journal of Chemical Physics*, 38(2), 433–436.

Moles, C. G., Mendes, P., & Banga, J. R. (2003). Parameter estimation in biochemical pathways: a comparison of global optimization methods. *Genome Research*, 13(11), 2467–2474.

Moyal, J. (1949a). The distribution of wars in time. *Journal of the Royal Statistical Society. Series A (General)*, 112(4), 446–449.

Moyal, J. (1949b). Stochastic processes and statistical physics. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(2), 150–210.

Munsky, B. & Khammash, M. (2006). The finite state projection algorithm for the solution of the chemical master equation. *The Journal of Chemical Physics*, 124(4), 044104.

Murray, J. D. (2001). *Mathematical Biology. II Spatial Models and Biomedical Applications*. Springer-Verlag New York Incorporated.

Narasimhan, M. N. (1993). *Principles of continuum mechanics*. Wiley-Interscience.

olde Scheper, T., Klinkenberg, D., Pennartz, C., & Van Pelt, J. (1999). A mathematical model for the intracellular circadian rhythm generator. *The Journal of Neuroscience*, 19(1), 40–47.

Omberg, L., Golub, G. H., & Alter, O. (2007). A tensor higher-order singular value decomposition for integrative analysis of DNA microarray data from different studies. *Proceedings of the National Academy of Sciences*, 104(47), 18371–18376.

Oseledets, I. (2009). Approximation of matrices with logarithmic number of parameters. In *Doklady Mathematics*, volume 80 (pp. 653–654).: Springer.

Oseledets, I. (2013). Constructive representation of functions in low-rank tensor formats. *Constructive Approximation*, 37(1), 1–18.

Oseledets, I. & Tyrtyshnikov, E. (2010). TT-cross approximation for multidimensional arrays. *Linear Algebra and its Applications*, 432(1), 70–88.

Oseledets, I. V. (2010). Approximation of $2^d \times 2^d$ matrices using tensor decomposition. *SIAM Journal on Matrix Analysis and Applications*, 31(4), 2130–2145.

Oseledets, I. V. (2011). Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5), 2295–2317.

Oseledets, I. V. & Dolgov, S. (2012). Solution of linear systems and matrix inversion in the TT-format. *SIAM Journal on Scientific Computing*, 34(5), A2718–A2739.

Oseledets, I. V. & Tyrtyshnikov, E. E. (2009). Breaking the curse of dimensionality, or how to use SVD in many dimensions. *SIAM Journal on Scientific Computing*, 31(5), 3744–3759.

Ostrowski, A. (1937). Über die determinanten mit überwiegender hauptdiagonale. *Commentarii Mathematici Helvetici*, 10(1), 69–96.

Pahle, J. (2009). Biochemical simulations: stochastic, approximate stochastic and hybrid approaches. *Briefings in Bioinformatics*, 10(1), 53–64.

Paulsson, J. (2004). Summing up the noise in gene networks. *Nature*, 427(6973), 415–418.

Paulsson, J., Berg, O. G., & Ehrenberg, M. (2000). Stochastic focusing: Fluctuation-enhanced sensitivity of intracellular regulation. *Proceedings of the National Academy of Sciences of the United States of America*, 97, 7148–7153.

Pawula, R. (1967). Approximation of the linear Boltzmann equation by the Fokker–Planck equation. *Physical review*, 162(1), 186.

Pedersen, A. R. (1995). A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scandinavian Journal of Statistics*, (pp. 55–71).

Pennington, J. M. & Rosenberg, S. M. (2007). Spontaneous DNA breakage in single living escherichia coli cells. *Nature Genetics*, 39(6), 797–802.

Perez-García, D., Verstraete, F., Wolf, M. M., & Cirac, J. I. (2007). Matrix product state representations. *Quantum Information and Computation*, 7(5-6), 401–430.

Pomerening, J. R., Sontag, E. D., & Ferrell, J. E. (2003). Building a cell cycle oscillator: hysteresis and bistability in the activation of cdc2. *Nature Cell Biology*, 5(4), 346–351.

Qian, H., Shi, P.-Z., & Xing, J. (2009). Stochastic bifurcation, slow fluctuations, and bistability as an origin of biochemical complexity. *Physical Chemistry Chemical Physics*, 11(24), 4861–4870.

Raj, A. & van Oudenaarden, A. (2008). Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2), 216–226.

Raser, J. M. & O'Shea, E. K. (2005). Noise in gene expression: origins, consequences, and control. *Science*, 309(5743), 2010–2013.

Reinker, S., Altman, R., Timmer, J., et al. (2006). Parameter estimation in stochastic biochemical reactions. *Systems Biology*, 153(4), 168.

Risken, H. (1984). *Fokker-Planck Equation*. Springer.

Risken, H. (1989). *The Fokker-Planck Equation, methods of solution and applications*. Springer-Verlag.

Rohwedder, T. & Uschmajew, A. (2013). On local convergence of alternating schemes for optimization of convex problems in the tensor train format. *SIAM Journal on Numerical Analysis*, 51(2), 1134–1162.

Rybak, I. (2004). Monotone and conservative difference schemes for elliptic equations with mixed derivatives 1. *Mathematical Modelling and Analysis*, 9(2), 169–178.

Salis, H. & Kaznessis, Y. (2005). Accurate hybrid stochastic simulation of a system of coupled chemical or biochemical reactions. *The Journal of chemical physics*, 122(5), 054103.

Samarskii, A., Matus, P., Mazhukin, V., & Mozolevski, I. (2002). Monotone difference schemes for equations with mixed derivatives. *Computers & Mathematics with Applications*, 44(3), 501–510.

Sanft, K. R., Wu, S., Roh, M., Fu, J., Lim, R. K., & Petzold, L. R. (2011). Stochkit2: software for discrete stochastic simulation of biochemical systems with events. *Bioinformatics*, 27(17), 2457–2458.

Sankaranarayanan, P., Schomay, T. E., Aiello, K. A., & Alter, O. (2015). Tensor GSVD of patient-and platform-matched tumor and normal DNA copy-number profiles uncovers chromosome arm-wide patterns of tumor-exclusive platform-consistent alterations encoding for cell transformation and predicting ovarian cancer survival. *PloS One*.

Savageau, M. A. (1971). Parameter sensitivity as a criterion for evaluating and comparing the performance of biochemical systems. *Nature*, 229, 542–544.

Schlögl, F. (1972). Chemical reaction models for non-equilibrium phase transitions. *Zeitschrift für Physik*, 253(2), 147–161.

Schnoerr, D., Sanguinetti, G., & Grima, R. (2014). The complex chemical Langevin equation. *The Journal of Chemical Physics*, 141(2), 024103.

Schollwöck, U. (2011a). The density-matrix renormalization group: a short introduction. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 369(1946), 2643–2661.

Schollwöck, U. (2011b). The density-matrix renormalization group in the age of matrix product states. *Annals of Physics*, 326(1), 96–192.

Schrodinger, E. (1944). *What Is Life? With Mind and Matter and Autobiographical Sketches*. Cambridge University Press.

Scott, M., Ingalls, B., & Kaern, M. (2006). Estimations of intrinsic and extrinsic noise in models of nonlinear genetic networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 16(2), 026107.

Shahrezaei, V., Ollivier, J. F., & Swain, P. S. (2008). Colored extrinsic fluctuations and stochastic gene expression. *Molecular Systems Biology*, 4(1), 196.

Shahrezaei, V. & Swain, P. S. (2008). Analytical distributions for stochastic gene expression. *Proceedings of the National Academy of Sciences*, 105(45), 17256–17261.

Sharafutdinov, V. A. (1994). *Integral geometry of tensor fields*, volume 1. Walter de Gruyter.

Sidiropoulos, N. D. & Bro, R. (2000). On the uniqueness of multilinear decomposition of n-way arrays. *Journal of Chemometrics*, 14(3), 229–239.

Silverman, R. A. et al. (1972). *Special functions and their applications*. Courier Corporation.

Simons, K. & Toomre, D. (2000). Lipid rafts and signal transduction. *Nature Reviews Molecular Cell Biology*, 1(1), 31–39.

Sjöberg, P. (2007). Partial approximation of the master equation by the Fokker-Planck equation. In *Applied Parallel Computing. State of the Art in Scientific Computing* (pp. 637–646). Springer.

Soto, R. (2001). A sufficient condition for the existence of an M-matrix with prescribed spectrum. *Computers & Mathematics with Applications*, 42(6), 849–859.

Swain, P. S. (2004). Efficient attenuation of stochasticity in gene expression through post-transcriptional control. *Journal of Molecular Biology*, 344(4), 965–976.

Takahashi, K., Tănase-Nicola, S., & Ten Wolde, P. R. (2010). Spatio-temporal correlations can drastically change the response of a mapk pathway. *Proceedings of the National Academy of Sciences*, 107(6), 2473–2478.

Thattai, M. & Van Oudenaarden, A. (2001). Intrinsic noise in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 98(15), 8614–8619.

Thomas, P. & Grima, R. (2015). Approximate probability distributions of the master equation. *Physical Review E*, 92(1), 012120.

Thomas, P., Matuschek, H., & Grima, R. (2012). Intrinsic noise analyzer: a software package for the exploration of stochastic biochemical kinetics using the system size expansion. *PloS One*, 7(6), e38518–e38518.

Thomas, P., Popović, N., & Grima, R. (2014). Phenotypic switching in gene regulatory networks. *Proceedings of the National Academy of Sciences*, 111(19), 6994–6999.

Thomas, P., Straube, A. V., Timmer, J., Fleck, C., & Grima, R. (2013). Signatures of nonlinearity in single cell noise-induced oscillations. *Journal of Theoretical Biology*, 335, 222–234.

Tian, T., Xu, S., Gao, J., & Burrage, K. (2007). Simulated maximum likelihood method for estimating kinetic rates in gene expression. *Bioinformatics*, 23(1), 84–91.

Toni, T., Welch, D., Strelkowa, N., Ipsen, A., & Stumpf, M. P. (2009). Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. *Journal of the Royal Society Interface*, 6(31), 187–202.

Truskey, G. A., Yuan, F., & Katz, D. F. (2004). *Transport phenomena in biological systems*. Upper Saddle River NJ:: Pearson/Prentice Hall.

Tsai, T. Y.-C., Choi, Y. S., Ma, W., Pomerening, J. R., Tang, C., & Ferrell, J. E. (2008). Robust, tunable biological oscillations from interlinked positive and negative feedback loops. *Science*, 321(5885), 126–129.

Tucker, L. (1966). Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31, 279–311.

Tyson, J. J. (1991). Modeling the cell division cycle: cdc2 and cyclin interactions. *Proceedings of the National Academy of Sciences*, 88(16), 7328–7332.

Tyson, J. J., Csikasz-Nagy, A., & Novak, B. (2002). The dynamics of cell cycle regulation. *Bioessays*, 24(12), 1095–1109.

van Kampen, N. (1961). A power series expansion of the master equation. *Canadian Journal of Physics*, 39(4), 551–567.

Van Kampen, N. (2007). Stochastic processes in physics and chemistry. Elsevier.

Van Kampen, N. G. (1992). *Stochastic processes in physics and chemistry*, volume 1. Elsevier.

Verstraete, F., Porras, D., & Cirac, J. I. (2004). Density matrix renormalization group and periodic boundary conditions: a quantum information perspective. *Physical review letters*, 93(22), 227205.

Vidal, G. (2003). Efficient classical simulation of slightly entangled quantum computations. *Physical Review Letters*, 91(14), 147902.

Volfson, D., Marciniak, J., Blake, W. J., Ostroff, N., Tsimring, L. S., & Hasty, J. (2006). Origins of extrinsic variability in eukaryotic gene expression. *Nature*, 439(7078), 861–864.

Von Dassow, G., Meir, E., Munro, E. M., & Odell, G. M. (2000). The segment polarity network is a robust developmental module. *Nature*, 406(6792), 188–192.

White, S. R. (1992). Density matrix formulation for quantum renormalization groups. *Physical Review Letters*, 69(19), 2863.

White, S. R. (1993). Density-matrix algorithms for quantum renormalization groups. *Physical Review B*, 48(14), 10345.

Wilkinson, D. J. (2009). Stochastic modelling for quantitative description of heterogeneous biological systems. *Nature Reviews Genetics*, 10(2), 122–133.

Wilkinson, D. J. (2011). *Stochastic modelling for systems biology*. CRC press.

Withers, C. S. (2000). A simple expression for the multivariate Hermite polynomials. *Statistics & Probability Letters*, 47(2), 165–169.

Wu, J., Vidakovic, B., & Voit, E. O. (2011). Constructing stochastic models from deterministic process equations by propensity adjustment. *BMC Systems Biology*, 5(1), 187.

Xiong, W. & Ferrell, J. E. (2003). A positive-feedback-based bistable 'memory module' that governs a cell fate decision. *Nature*, 426(6965), 460–465.

Yip, E. (1996). A necessary and sufficient condition for M-matrices and its relation to block LU factorization. *Linear Algebra and Its Applications*, 235, 261–274.

Zechner, C., Ruess, J., Krenn, P., Pelet, S., Peter, M., Lygeros, J., & Koeppl, H. (2012). Moment-based inference predicts bimodality in transient gene expression. *Proceedings of the National Academy of Sciences*, 109(21), 8340–8345.

Zhang, F. (2011). *Matrix theory: basic results and techniques*. Springer Science & Business Media.

A

A perturbative analysis of the CME

In this section, we perform an asymptotic analysis of the CME (2.4). We will start with introducing the system size expansion transform the distribution of molecular population into the distribution of the fluctuation around the mean concentration in Section § A.1. Then, the distribution is expanded into a series expansion in the inverse square root of the system volume V in Section § A.2. To solve for the expanded coefficients, we first perform the multivariate Hermite expansion to the coefficients in Section § A.3, and then transform the coordinate in Section § A.4 such that the covariance matrix is a scaled identity matrix. In Section § A.5, we introduce the eigenfunction expansion and show the equivalence between the eigenfunction expansion and the transformed Hermite expansion. We show in Section § A.6 that the two types of the expansions can be applied to obtain the equations for the expanded coefficients, and the analytical derivation of the asymptotic solution is obtained.

A.1 System size expansion of the CME

In this section, we present a system-size expansion (SSE) of the chemical master equation (2.4) developed by van Kampen (1961). The validity of the expansion is implicitly based on the assumption that the steady-state of the chemical system is asymptotically stable.

The starting point of the SSE is performing the change of variables, or ansatz,

$$\frac{n_i}{V} = \phi_i + \sqrt{V}\epsilon_i, \quad i = 1,2,\ldots,N, \tag{A.1}$$

where the instantaneous molecular population $\boldsymbol{x}/V$ is decomposed into a deterministic part $\boldsymbol{\phi} = [\phi_1,\phi_2,\ldots,\phi_N]^T$ and a fluctuation part $\boldsymbol{\epsilon} = [\epsilon_1,\epsilon_2,\ldots,\epsilon_N]^T$. The evolution of $\boldsymbol{\phi}$ is full determined by the deterministic reaction rate equation, and the probability distribution of molecular populations, $P_{\mathrm{m}}(\boldsymbol{n},t)$, is transformed into the probabilistic distribution of the fluctuation, written as $\Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t)$.

In the large volume limit, the distribution $\Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t)$ is assumed to be continuous, and the change of variables (A.1) implies

$$\Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t) = V^{N/2} P_{\mathrm{m}}(\boldsymbol{n},t). \tag{A.2}$$

Here, as an analogue of the Riemann integral, the integral $P_{\mathrm{m}}(\boldsymbol{n},t) = \int_{\boldsymbol{\epsilon}_0}^{\boldsymbol{\epsilon}_0+1/\sqrt{V}} \Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t)\mathrm{d}\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}_0 = \frac{1}{V\sqrt{V}}\boldsymbol{n} - \frac{1}{\sqrt{V}}\boldsymbol{\phi}$, is approximated by the product, $\frac{1}{\sqrt{V}}\Pi_{\mathrm{m}}(\boldsymbol{\epsilon}_0,t)$, for sufficiently large volume V. To derive the CME in the new variables, the step operator, upon the change of variable, is expanded by the Taylor series as

$$\prod_{i=1}^{N} \mathscr{E}^{-\nu_{j,i}} - 1 = \sum_{i=1}^{\infty} \left(-\frac{1}{V}\right)^{\frac{i}{2}} \sum_{\substack{c_1,c_2,\ldots,c_N=0 \\ c_1+c_2\cdots+c_N=i}}^{i} \frac{\boldsymbol{\nu}_j^{\boldsymbol{c}}}{\boldsymbol{c}!} \left(\frac{\partial}{\partial \boldsymbol{\epsilon}}\right)^{\boldsymbol{c}},$$

for $j = 1,2,\ldots,M$. In what follows, we rewrite the above expansion in a simplified

form as

$$\prod_{i=1}^{N} \mathscr{E}^{-\nu_{j,i}} - 1 = \sum_{i=1}^{\infty} (-1)^i V^{-i/2} a_j^{(i)}, \tag{A.3}$$

where

$$a_j^{(s)} = \frac{1}{s!} \left(\sum_{i=1}^{N} \nu_{j,i} \frac{\partial}{\partial \epsilon_i} \right)^s.$$

Next, we substitute the ansatz (A.1) into propensity function, $\alpha_j(\boldsymbol{n})$, and apply the Taylor expansion, which yields

$$\alpha_j(V\boldsymbol{\phi} + \sqrt{V}\boldsymbol{\epsilon}) = \sum_{i=0}^{\infty} \left(\frac{1}{V} \right)^{\frac{i}{2}} \sum_{\substack{c_1, c_2, \ldots, c_N = 0 \\ c_1 + c_2 \cdots + c_N = i}}^{i} \frac{\boldsymbol{\epsilon}^{\boldsymbol{c}}}{\boldsymbol{c}!} \left(\frac{\partial}{\partial \boldsymbol{\phi}} \right)^{\boldsymbol{c}} \alpha_j(V\boldsymbol{\phi}). \tag{A.4}$$

We also expand the propensity function, $\alpha_j(V\boldsymbol{\phi})$, in inverse powers of V via

$$\alpha_j(V\boldsymbol{\phi}) = V \sum_{s=0}^{\infty} V^{-s} f_j^{(2s)}(\boldsymbol{\phi}). \tag{A.5}$$

For the mass-action kinetics, the propensity function α_j has the general form (2.2), and hence by collecting terms of different inverse orders of V, we have the leading order term

$$f_j^{(0)}(\boldsymbol{\phi}) = f_j(\boldsymbol{\phi}) = k_j \prod_i \phi_i^{\nu_{j,i}^-}, \tag{A.6}$$

where f_j is the macroscopic rate function of j-th reaction, and the higher order terms are given by

$$f_j^{(2s)} = k_j \sum_{\substack{s_1, s_2, \ldots, s_N = 0 \\ s_1 + s_2 \cdots + s_N = s}}^{s} \prod_{i=1}^{N} \phi_i^{\nu_{j,i}^- - s_i} \begin{bmatrix} \nu_{j,i}^- \\ s_i \end{bmatrix} \mathbb{1}_{s_i < \nu_{j,i}^-}, \quad s = 1, 2, 3, \ldots,$$

where notation $\left[\cdot \right]$ denotes the Stirling number of the first kind. Then substituting

(A.5) into (A.4) yields the series expansion of the microscopic propensity, written as

$$\alpha_j(\boldsymbol{n}) = V\left[b_j^{(0,0)} + V^{-1/2}b_j^{(1,0)} + V^{-1}b_j^{(2,0)} + V^{-1}b_j^{(0,2)} + V^{-3/2}b_j^{(1,2)} + \mathcal{O}\left(V^{-2}\right)\right], \quad \text{(A.7)}$$

where

$$b_j^{(m,n)} = \frac{1}{m!}\left(\sum_{i=1}^{N}\epsilon_i\frac{\partial}{\partial\phi_i}\right)^m f_j^{(n)}(\boldsymbol{\phi}), \quad m = 0,1,2,\ldots, \text{ and } n = 0,2,4,\ldots.$$

Note that we do not impose any restriction on the order of the biochemical reactions.

Substituting now expansions (A.3) and (A.7) into the CME (2.4), and rearranging the equation in the inverse power of $\sqrt{V}$, we have

$$\begin{aligned}
&\left(\frac{\partial}{\partial t} - V^{1/2}\sum_{i=1}^{N}\frac{\mathrm{d}\phi_i}{\mathrm{d}t}\frac{\partial}{\partial\epsilon_i}\right)\Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t)\\
&= -V^{1/2}\sum_{j=1}^{M}a_j^{(1)}b_j^{(0,0)}\Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t) + V^0\sum_{j=1}^{N}\left(a_j^{(2)}b_j^{(0,0)} - a_j^{(1)}b_j^{(1,0)}\right)\Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t)\\
&+ V^{-1/2}\sum_{j=1}^{M}\left(-a_j^{(3)}b_j^{(0,0)} + a_j^{(2)}b_j^{(1,0)} - a_j^{(1)}b_j^{(2,0)} - a_j^{(1)}b_j^{(0,2)}\right)\Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t)\\
&+ V^{-1}\sum_{j=1}^{M}\left(a_j^{(4)}b_j^{(0,0)} - a_j^{(3)}b_j^{(1,0)} + a_j^{(2)}b_j^{(2,0)} + a_j^{(2)}b_j^{(0,2)} - a_j^{(1)}b_j^{(1,2)}\right)\Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t)\\
&+ \mathcal{O}\left(V^{-3/2}\right).
\end{aligned}$$

Equating the order $V^{1/2}$ terms on both sides of the above equation gives

$$\sum_{i=1}^{N}\frac{\mathrm{d}\phi_i}{\mathrm{d}t}\frac{\partial\Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t)}{\partial\epsilon_i} = \sum_{i=1}^{N}\sum_{j=1}^{M}\nu_{j,i}f_j(\boldsymbol{\phi})\frac{\partial\Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t)}{\partial\epsilon_i}.$$

Using the fact that $\boldsymbol{\phi}$ are subject to the macroscopic reaction rate equation, these terms cancel regardless of distribution Π_{m}. Then, we can write the following new

form of the CME in variables $\boldsymbol{\epsilon}$ as

$$\begin{aligned}\frac{\partial \Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t)}{\partial t} =& V^0 \sum_{j=1}^{N} \left(a_j^{(2)} b_j^{(0,0)} - a_j^{(1)} b_j^{(1,0)}\right) \Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t) \\ &+ V^{-1/2} \sum_{j=1}^{M} \left(-a_j^{(3)} b_j^{(0,0)} + a_j^{(2)} b_j^{(1,0)} - a_j^{(1)} b_j^{(2,0)} - a_j^{(1)} b_j^{(0,2)}\right) \Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t) \\ &+ V^{-1} \sum_{j=1}^{M} \left(a_j^{(4)} b_j^{(0,0)} - a_j^{(3)} b_j^{(1,0)} + a_j^{(2)} b_j^{(2,0)} + a_j^{(2)} b_j^{(0,2)} - a_j^{(1)} b_j^{(1,2)}\right) \Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t) \\ &+ \mathscr{O}\left(V^{-3/2}\right).\end{aligned} \tag{A.8}$$

To simplify the notation, we define $\mathscr{L}_s$, $s = 0,1,2,\ldots$, by the terms on order $V^{-s/2}$ that operate on Π_{m} in (A.11), i.e.,

$$\mathscr{L}_s = \sum_{j=1}^{M} \left(\sum_{w=0}^{\lceil s/2 \rceil} \sum_{k=1}^{s-2(w-1)} (-1)^k a_j^{(k)} b_j^{(s-k-2(w-1),2w)} \right),$$

where $\lceil \cdot \rceil$ denotes the ceiling value. We will later use an equivalent explicit expression of $\mathscr{L}_s$ written as

$$\mathscr{L}_s = \sum_{w=0}^{\lceil s/2 \rceil} \sum_{k=1}^{s-2(w-1)} \sum_{\substack{i_1,\ldots,i_N=0 \\ i_1+\cdots+i_N=k}}^{k} \sum_{\substack{j_1,\ldots,j_N=0 \\ j_1+\cdots+j_N=s-k-2(w-1)}}^{s-k-2(w-1)} \mathscr{D}_{k,s-k-2(w-1),w}^{\boldsymbol{i},\boldsymbol{j}} \left(\frac{\partial}{\partial \boldsymbol{\epsilon}}\right)^{\boldsymbol{i}} \boldsymbol{\epsilon}^{\boldsymbol{j}}, \tag{A.9}$$

and the constant coefficients are given by

$$\mathscr{D}_{k,q,w}^{\boldsymbol{i},\boldsymbol{j}} = \sum_{r=1}^{M} \frac{1}{k!q!} \left[\prod_{n=1}^{N-1} \binom{q - \sum_{\ell=1}^{n-1} i_\ell}{i_n}\right] \left[\prod_{n=1}^{N-1} \binom{q - \sum_{\ell=1}^{n-1} j_\ell}{j_n}\right] \left(\prod_{n=1}^{N} \nu_{j,n}^{i_n}\right) \left(\frac{\partial}{\partial \boldsymbol{\phi}}\right)^{\boldsymbol{j}} f_r^{(w)}(\boldsymbol{\phi}). \tag{A.10}$$

Then (A.8) can be written in a simple form as

$$\frac{\partial \Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t)}{\partial t} = \sum_{s=0}^{\infty} V^{-s/2} \mathscr{L}_s \Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t). \tag{A.11}$$

A.2 Perturbative expansion of the CME distribution

We now express the CME distribution Π_{m} in terms of a perturbation series in inverse powers of square root of system volume as

$$\Pi_{\mathrm{m}}(\boldsymbol{\epsilon},t)=\sum_{j=0}^{\infty}\Pi_{\mathrm{m},j}(\boldsymbol{\epsilon},t)V^{-j/2}. \tag{A.12}$$

We proceed by substituting (A.12) into (A.11) and rearranging the result to derive the equations to order $V^{-j/2}$ with $j\geq 0$ as

$$\left(\frac{\partial}{\partial t}-\mathscr{L}_0\right)\Pi_{\mathrm{m},j}(\boldsymbol{\epsilon},t)=\sum_{i=0}^{j-1}\mathscr{L}_{j-i}\Pi_{\mathrm{m},i}(\boldsymbol{\epsilon},t), \tag{A.13}$$

where operators $\mathscr{L}_s$ are defined in (A.9). We use the normalisation conditions

$$\int_{\mathbb{R}^N}\Pi_{\mathrm{m},0}(\boldsymbol{\epsilon},t)\mathrm{d}\boldsymbol{\epsilon}=1 \quad \text{and} \quad \int_{\mathbb{R}^N}\Pi_{\mathrm{m},j}(\boldsymbol{\epsilon},t)\mathrm{d}\boldsymbol{\epsilon}=0, \quad j=1,2,\ldots. \tag{A.14}$$

It implies that $\Pi_{\mathrm{m},j}$ with $j\geq 1$ are negative in some region of the state space $\mathbb{R}^N$, and this is in accordance with Pawula's theorem (Pawula, 1967) that $\Pi_{\mathrm{m},j}$ cannot be a genuine probability density function if its evolution equation involves the higher order derivatives (Grima et al., 2011).

Of particular interest is the order V^0 solution, which satisfies a linear Fokker-Plank equation

$$\frac{\partial\Pi_{\mathrm{m},0}(\boldsymbol{\epsilon},t)}{\partial t}=\mathscr{L}_0\Pi_{\mathrm{m},0}(\boldsymbol{\epsilon},t), \tag{A.15}$$

and an explicit expression for operator $\mathscr{L}_0$ is given by

$$\mathscr{L}_0\Pi_{\mathrm{m},0}(\boldsymbol{\epsilon},t)=\frac{1}{2}\sum_{i,j=1}^{N}D_{i,j}(t)\frac{\partial^2\Pi_{\mathrm{m},0}(\boldsymbol{\epsilon},t)}{\partial\epsilon_i\partial\epsilon_j}-\sum_{i,j=1}^{N}J_{i,j}(t)\epsilon_j\frac{\partial\Pi_{\mathrm{m},0}(\boldsymbol{\epsilon},t)}{\partial\epsilon_i}-\mathrm{tr}\boldsymbol{J}(t)\cdot\Pi_{\mathrm{m},0}(\boldsymbol{\epsilon},t),$$

where the Jacobian matrix $\boldsymbol{J}(t) = [J_{i,j}(t)]_{i,j=1}^{N}$ and the diffusion matrix $\boldsymbol{D}(t) = [D_{i,j}(t)]_{i,j=1}^{N}$ are respectively defined by

$$\boldsymbol{J}(t) = \left.\frac{\mathrm{d}(\boldsymbol{\nu}\boldsymbol{f})}{\mathrm{d}\boldsymbol{\phi}}\right|_{\boldsymbol{\phi}=\boldsymbol{\phi}(t)} \quad \text{and} \quad \boldsymbol{D}(t) = \boldsymbol{B}(t)\boldsymbol{B}^{T}(t) = \sum_{j=1}^{M} f_j(\boldsymbol{\phi})\boldsymbol{\nu}_j \otimes \boldsymbol{\nu}_j, \tag{A.16}$$

with $\boldsymbol{f} = (f_1(\boldsymbol{\phi}), \ldots, f_M(\boldsymbol{\phi}))^T$ is a vector of the deterministic rate functions.

The above equation is also known as the *linear noise approximation* (LNA), and it describes the time-evolution of the probability distribution of the following linear stochastic differential equation:

$$\mathrm{d}\boldsymbol{\epsilon} = \boldsymbol{J}(t)\boldsymbol{\epsilon}\mathrm{d}\boldsymbol{\epsilon} + \boldsymbol{B}(t)\mathrm{d}\boldsymbol{w}, \quad \boldsymbol{\epsilon} \in \mathbb{R}^N, \tag{A.17}$$

where $\boldsymbol{w}$ is a standard N-dimensional Wiener process. Under the normalisation condition (A.14), the solution of the LNA (A.15) is known to be a multivariate Gaussian distribution (Van Kampen, 1992), written as

$$\Pi_{\mathrm{m},0}(\boldsymbol{\epsilon}, t) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{\Sigma}(t)|}} \exp\left(-\frac{1}{2}\boldsymbol{\epsilon}^T \boldsymbol{\Sigma}^{-1}(t)\boldsymbol{\epsilon}\right), \tag{A.18}$$

where the covariance matrix $\boldsymbol{\Sigma}(t) \in \mathbb{R}^{N\times N}$ satisfies the (Lyapunov) matrix equation

$$\frac{\mathrm{d}\boldsymbol{\Sigma}(t)}{\mathrm{d}t} = \boldsymbol{J}(t)\boldsymbol{\Sigma}(t) + \boldsymbol{\Sigma}(t)\boldsymbol{J}^T(t) + \boldsymbol{D}(t). \tag{A.19}$$

A.3 Hermite polynomial based expansion of the steady-state distribution of the CME

A.3.1 The steady-state distribution

In what follows, we focus on the stationary case. We assume that trajectories of the deterministic reaction rate equation would converge to an asymptotically stable fixed point. Under this assumption, one can derive the system size expansions of the stationary distribution by setting the time derivatives to zero.

We denote the stationary quantities by dropping its time dependence, i.e.,

$$\Pi_{\mathrm{m}}(\boldsymbol{\epsilon}) = \lim_{t\to\infty} \Pi_{\mathrm{m}}(\boldsymbol{\epsilon}, t) \quad \text{and} \quad \Pi_{\mathrm{m},j}(\boldsymbol{\epsilon}) = \lim_{t\to\infty} \Pi_{\mathrm{m},j}(\boldsymbol{\epsilon}, t), \quad j = 1, 2, \ldots.$$

Then, the stationary counterpart of the series expansion (A.12) is written as

$$\Pi_{\mathrm{m}}(\boldsymbol{\epsilon}) = \sum_{j=0}^{\infty} \Pi_{\mathrm{m},j}(\boldsymbol{\epsilon}) V^{-j/2}. \tag{A.20}$$

The coefficients $\Pi_{\mathrm{m},j}$ can be obtained by solving the following set of the equations in the increasing order:

$$\mathscr{L}_0 \Pi_{\mathrm{m},0}(\boldsymbol{\epsilon}) = 0, \quad \text{with} \quad \int_{\mathbb{R}^N} \Pi_{\mathrm{m},0}(\boldsymbol{\epsilon}) \mathrm{d}\boldsymbol{\epsilon} = 1, \tag{A.21}$$

and

$$\mathscr{L}_0 \Pi_{\mathrm{m},j}(\boldsymbol{\epsilon}) = \sum_{i=0}^{j-1} \mathscr{L}_{j-i} \Pi_{\mathrm{m},i}(\boldsymbol{\epsilon}), \quad \text{with} \quad \int_{\mathbb{R}^N} \Pi_{\mathrm{m},j}(\boldsymbol{\epsilon}) \mathrm{d}\boldsymbol{\epsilon} = 0, \quad j = 1, 2, 3, \ldots,$$

where operators $\mathscr{L}_j$ are defined in (A.9).

The coefficient $\Pi_{\mathrm{m},0}(\boldsymbol{\epsilon})$ on order V^0 is the steady state distribution of the LNA

(A.15), given by

$$\Pi_{\mathrm{m},0}(\boldsymbol{\epsilon}) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\boldsymbol{\epsilon}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\epsilon}\right), \tag{A.22}$$

where the stationary covariance matrix $\boldsymbol{\Sigma}$ satisfies the Lyapunov equation

$$\boldsymbol{J}\boldsymbol{\Sigma} + \boldsymbol{\Sigma}\boldsymbol{J}^T + \boldsymbol{D} = \boldsymbol{0}, \tag{A.23}$$

where the Jacobian matrix $\boldsymbol{J}$ and the diffusion matrix $\boldsymbol{D}$, defined in (A.16), are evaluated at the deterministic steady state.

A.3.2 The Hermite polynomial expansion

The higher order coefficients in (A.20) are solved via using the multivariate Hermite expansion, written as

$$\Pi_{\mathrm{m},j}(\boldsymbol{\epsilon}) = \sum_{s=0}^{N_j} \sum_{\substack{i_1, i_2, \ldots, i_N = 0 \\ i_1 + i_2 + \cdots + i_N = s}}^{s} c_{\boldsymbol{i}}^{(j)} H_{\boldsymbol{i}}(\boldsymbol{\epsilon}, \boldsymbol{\Sigma}) \Pi_0(\boldsymbol{\epsilon}, \boldsymbol{\Sigma}), \quad N_j \in \mathbb{N}, \quad c_{\boldsymbol{i}}^{(j)} \in \mathbb{R}, \tag{A.24}$$

where $\Pi_0(\boldsymbol{\epsilon}, \boldsymbol{\Sigma}) = \Pi_{\mathrm{m},0}(\boldsymbol{\epsilon})$ denotes the N-dimensional Gaussian distribution with mean 0 and covariance matrix $\boldsymbol{\Sigma}$. The $\boldsymbol{i}$-th (multivariate) Hermite polynomial function, $H_{\boldsymbol{i}}$, with respect to $\boldsymbol{\Sigma}$ is defined as

$$H_{\boldsymbol{i}}(\boldsymbol{\epsilon}, \boldsymbol{\Sigma}) = \Pi_0^{-1}(\boldsymbol{\epsilon}, \boldsymbol{\Sigma})(-\partial/\partial\boldsymbol{\epsilon})^{\boldsymbol{i}} \Pi_0(\boldsymbol{\epsilon}, \boldsymbol{\Sigma}). \tag{A.25}$$

It is known that the Hermite polynomials are orthogonal in the sense that

$$\int H_{\boldsymbol{j}}(\boldsymbol{\Sigma}^{-1}\boldsymbol{\epsilon}, \boldsymbol{\Sigma}^{-1}) H_{\boldsymbol{i}}(\boldsymbol{\epsilon}, \boldsymbol{\Sigma}) \Pi_0(\boldsymbol{\epsilon}) \mathrm{d}\boldsymbol{\epsilon} = \begin{cases} \boldsymbol{j}! & \text{if } \boldsymbol{i} = \boldsymbol{j} \\ 0 & \text{otherwise} \end{cases} \tag{A.26}$$

and they form an orthogonal basis for $L^2(\mathbb{R}^N)$. The Hermite polynomial function can be explicitly written in terms of monomials as

$$H_{\boldsymbol{i}}(\boldsymbol{\epsilon}, \boldsymbol{\Sigma}) = \sum_{s=0}^{\left\lfloor \frac{i_1+i_2+\cdots+i_N}{2} \right\rfloor} (-1)^s \sum_{\substack{j_1,j_2,\ldots,j_N=0 \\ i_1+i_2+\cdots+i_N=2s}}^{s} \mathbb{E}\left(\boldsymbol{\epsilon}^{\boldsymbol{j}}\right) \prod_{w=1}^{N} \binom{i_w}{j_w} \epsilon_w^{i_w-j_w}. \tag{A.27}$$

We refer the readers to Withers (2000) for detailed properties of the multivariate Hermite polynomials.

Note that N_j in (A.24) refers to the maximum summation index of Hermite polynomials that are necessary to express the function $\Pi_{\mathrm{m},0}$, and is not bounded in general case. But, later on, we will show that $N_j = 3j$ and elaborate how the coefficients $\left\{c_{\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\le N_j}$ may be evaluated.

A.4 Transform of coordinate

Solving the expanded coefficients $\left\{c_{\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\le N_j}$ in (A.24) is difficult, thus in this section, we introduce a transform of coordinate such that covariance matrix is a diagonal matrix.

Consider an orthogonal coordinate transformation given by

$$\boldsymbol{z} = \boldsymbol{T}\boldsymbol{\epsilon}, \tag{A.28}$$

and the transition matrix is defined as

$$\boldsymbol{T} = \frac{\sqrt{2}}{2}\boldsymbol{\Lambda}^{1/2}\boldsymbol{L}^*,$$

where matrices $\boldsymbol{L}$ and $\boldsymbol{\Lambda}$ correspond to the eigenvalue decomposition of the covariance matrix, i.e.,

$$\boldsymbol{\Sigma} = \boldsymbol{L}\Lambda\boldsymbol{L}^T = 2\boldsymbol{T}^*\boldsymbol{T}.$$

Using the above transformation, the steady state distribution (A.22) on order V^0, as the solution of (A.21), now reads

$$\tilde{\Pi}_0(\boldsymbol{z}) = \frac{1}{\pi^{N/2}} \exp\left(-\boldsymbol{z}^T \boldsymbol{z}\right), \tag{A.29}$$

with the stationary covariance matrix being $\tilde{\boldsymbol{\Sigma}} = \frac{1}{2}\boldsymbol{I}$. To see this, we multiply (A.23) by $\boldsymbol{T}$ from the left and by $\boldsymbol{T}^*$ from the right, and use the fact that $\boldsymbol{L}\boldsymbol{L}^* = \boldsymbol{I}$, which yields

$$\tilde{\boldsymbol{J}}\tilde{\boldsymbol{\Sigma}} + \tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{J}}^T + \tilde{\boldsymbol{D}} = \boldsymbol{0}, \quad \text{and} \quad \tilde{\boldsymbol{\Sigma}} = \frac{1}{2}\boldsymbol{I}, \tag{A.30}$$

where $\tilde{\boldsymbol{J}} = \boldsymbol{T}\boldsymbol{J}\boldsymbol{T}^{-1}$ and $\tilde{\boldsymbol{B}} = \boldsymbol{T}\boldsymbol{B}$.

We also apply the transformation (A.28) to the higher order coefficients in the series expansion (A.20) of the CME stationary distribution, and rewrite (A.20) in the new variables as

$$\tilde{\Pi}_{\mathrm{m}}(\boldsymbol{z}) = \sum_{j=0}^{\infty} \tilde{\Pi}_{\mathrm{m},j}(\boldsymbol{z}) V^{-j/2}. \tag{A.31}$$

The higher order coefficients $\tilde{\Pi}_{\mathrm{m},j}$, $j = 1, 2, \ldots$, are the solutions of the following set of the second-order partial differential equations

$$\tilde{\mathscr{L}}_0 \tilde{\Pi}_{\mathrm{m},j}(\boldsymbol{z}) = \sum_{i=0}^{j-1} \tilde{\mathscr{L}}_{j-i} \tilde{\Pi}_{\mathrm{m},i}(\boldsymbol{z}). \quad j = 1, 2, 3, \ldots. \tag{A.32}$$

We denote $\tilde{\mathscr{L}}_s$ by the operators $\mathscr{L}_s$ in (A.9) but with transformed variables $\boldsymbol{z}$, and the expression for $\tilde{\mathscr{L}}_s$ is of the form

$$\tilde{\mathscr{L}}_s = \sum_{w=0}^{\lceil s/2 \rceil} \sum_{k=1}^{s-2(w-1)} \sum_{\substack{i_1,\ldots,i_N=0 \\ i_1+\cdots+i_N=k}}^{k} \sum_{\substack{j_1,\ldots,j_N=0 \\ j_1+\cdots+j_N=s-k-2(w-1)}}^{s-k-2(w-1)} \mathscr{F}^{\boldsymbol{i},\boldsymbol{j}}_{k,s-k-2(w-1),w} \left(\frac{\partial}{\partial \boldsymbol{z}}\right)^{\boldsymbol{i}} \boldsymbol{z}^{\boldsymbol{j}}, \tag{A.33}$$

The explicit expression for the coefficients $\mathscr{F}_{k,q,w}^{\boldsymbol{i},\boldsymbol{j}}$ is given by

$$\mathscr{F}_{k,q,w}^{\boldsymbol{i},\boldsymbol{j}} = \sum_{\substack{c_1,\ldots,c_N=0 \\ c_1+\cdots+c_N=|\boldsymbol{i}|}}^{|\boldsymbol{i}|} \sum_{\substack{d_1,\ldots,d_N=0 \\ d_1+\cdots+d_N=|\boldsymbol{j}|}}^{|\boldsymbol{j}|} \mathscr{C}_{\boldsymbol{c}} \mathscr{C}_{\boldsymbol{d}} \mathscr{D}_{k,q,w}^{\boldsymbol{c},\boldsymbol{d}}, \tag{A.34}$$

where $\dot{\boldsymbol{J}} = \left(\dot{J}_{i,j}\right)_{i,j=1}^{N} = \partial \boldsymbol{z}/\partial \boldsymbol{\epsilon}$ is the Jacobian matrix, and

$$\mathscr{C}_{\boldsymbol{c}} = \sum_{\substack{k_{1,1},\ldots,k_{N,1}=0 \\ n_{1,1}+\cdots+n_{N,1}=i_1 \\ \boldsymbol{n}_{:,1}\leq \boldsymbol{c}}}^{c_1} \cdots \sum_{\substack{n_{1,N},\ldots,n_{N,N}=0 \\ n_{1,N}+\cdots+n_{N,N}=i_N \\ \boldsymbol{n}_{:,N}\leq \boldsymbol{c}}}^{c_N} \left(\prod_{h_1=1}^{N} \prod_{h_2=1}^{N-1} \binom{c_{h_1} - \sum_{r=1}^{h_2} n_{h_1,r}}{n_{h_1,h_2}} \right) \left(\prod_{\ell_1,\ell_2=1}^{N} \dot{J}_{\ell_1,\ell_2}^{n_{\ell_1,\ell_2}} \right).$$

The above expression shows that the coefficients $\mathscr{F}_{k,q,w}^{\boldsymbol{i},\boldsymbol{j}}$ in (A.33) is a linear combination of a finite number of coefficients $\left\{\mathscr{D}_{k,q,w}^{\boldsymbol{c},\boldsymbol{d}}\right\}_{|\boldsymbol{c}|\leq|\boldsymbol{i}|,|\boldsymbol{d}|\leq|\boldsymbol{j}|}$ given in (A.9).

Upon the change of coordinate, the Hermite expansions (A.24) in the new variables, $\boldsymbol{z}$, can be written as

$$\tilde{\Pi}_{\mathrm{m},j}(\boldsymbol{z}) = \sum_{s=0}^{N_j} \sum_{\substack{i_1,i_2,\ldots,i_N=0 \\ i_1+i_2+\cdots+i_N=s}}^{s} \tilde{c}_{\boldsymbol{i}}^{(j)} H_{\boldsymbol{i}}\left(\boldsymbol{z}, \frac{1}{2}\boldsymbol{I}\right) \tilde{\Pi}_0\left(\boldsymbol{z}, \frac{1}{2}\boldsymbol{I}\right), \tag{A.35}$$

where the coefficients $\left\{\tilde{c}_{\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\leq N_j}$ are related to $\left\{c_{\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\leq N_j}$ via

$$c_{\boldsymbol{i}}^{(j)} = \sum_{\substack{j_1,j_2,\ldots,j_N=0 \\ j_1+j_2+\cdots+j_N=s}}^{s} \tilde{c}_{\boldsymbol{j}}^{(j)} \dot{T}_{\boldsymbol{j},\boldsymbol{i}}^{N_j}. \tag{A.36}$$

Here, matrix $\dot{\boldsymbol{T}}^{N_j} \in \mathbb{R}^{N^{N_j} \times N^{N_j}}$ is a transition matrix from the polynomial basis $\left\{H_{\boldsymbol{i}}(\boldsymbol{z}, \frac{1}{2}\boldsymbol{I})\right\}_{|\boldsymbol{i}|\leq N_j}$ to the basis $\{H_{\boldsymbol{i}}(\boldsymbol{z}, \boldsymbol{\Sigma})\}_{|\boldsymbol{i}|\leq N_j}$, and the elements of $\dot{\boldsymbol{T}}^{N_j}$ are given by

$$\dot{T}_{\boldsymbol{i},\boldsymbol{j}}^{N_j} = \mathscr{C}_{\boldsymbol{j}}. \tag{A.37}$$

Since both $\left\{H_{\boldsymbol{i}}(\boldsymbol{z}, \frac{1}{2}\boldsymbol{I})\right\}_{|\boldsymbol{i}|\leq N_j}$ and $\{H_{\boldsymbol{i}}(\boldsymbol{z}, \boldsymbol{\Sigma})\}_{|\boldsymbol{i}|\leq N_j}$ are orthogonal basis functions, the transition matrix $\dot{\boldsymbol{T}}^{N_j}$ is of full rank (invertible). This means, given the values of $\tilde{c}_{\boldsymbol{i}}^{(j)}$,

we can invert the operator $\boldsymbol{T}^{N_j}$ to compute the values of $c_{\boldsymbol{i}}^{(j)}$. In the following section, we show how to derive the values of $\tilde{c}_{\boldsymbol{i}}^{(j)}$ using the eigenfunction expansions.

A.5 The eigenfunction expansion

Consider the following eigenfunction expansion

$$\tilde{\Pi}_{\mathrm{m},j}(\boldsymbol{z}) = \sum_{s=0}^{N_j^*} \sum_{\substack{i_1,i_2,\ldots,i_N=0 \\ i_1+i_2+\cdots+i_N=s}}^{s} \hat{c}_{\boldsymbol{i}}^{(j)} \rho_{\boldsymbol{i}}(\boldsymbol{z}), \tag{A.38}$$

with constant coefficients $\hat{c}_{\boldsymbol{i}}^{(j)}$. Here, we denote $\rho_{\boldsymbol{i}}$ by the $\boldsymbol{i}$-th eigenfunction corresponding to the eigenvalue $\lambda_{\boldsymbol{i}}$ of operator $\tilde{\mathscr{L}}_0$, i.e.,

$$\tilde{\mathscr{L}}_0 \rho_{\boldsymbol{i}}(\boldsymbol{z}) = \lambda_{\boldsymbol{i}} \rho_{\boldsymbol{i}}(\boldsymbol{z}) \quad \boldsymbol{z} \in \mathbb{R}^N.$$

Next, we will derive the expression for $\rho_{\boldsymbol{i}}$ and then show that the eigenfunctions are linear combinations of a finite number of the Hermite polynomials. We will also show the equivalence between the eigenvalue expansion (A.38) and the Hermite expansion (A.35), and show that $N_j^* = N_j$.

Let us define the summation index of an N-tuple of indices by the sum of its components, i.e., the summation index of $\boldsymbol{i} = (i_1, \ldots, i_N)^T$ is its absolute value $|\boldsymbol{i}|$. Then, one can show by the stars and bars argument that there are $\binom{N+K-1}{N}$ different combinations of the N-tuple of indices that correspond to a fixed summation index K.

For the case $|\boldsymbol{i}| = 0$, we choose the eigenfunction $\rho_{\boldsymbol{0}}$ to be the steady state distribution $\tilde{\Pi}_0$ in (A.29), and it is easy to see that the corresponding eigenvalue is zero, $\lambda_{\boldsymbol{0}} = 0$. For the case $|\boldsymbol{i}| = 1$, there are N distinct combinations where one of the indices is 1, and the rest are zeros. Suppose the j-th index equals to 1, $1 \le j \le N$, we

determine the eigenfunction of the form

$$\rho_{\boldsymbol{i}}(\boldsymbol{z}) = \boldsymbol{\eta}_j^T \boldsymbol{z} \rho_{\boldsymbol{0}}(\boldsymbol{z}) = \left(\sum_{s=1}^{N} \eta_{j,s} z_s \right) \rho_{\boldsymbol{0}}(\boldsymbol{z}), \quad |\boldsymbol{i}| = 1. \tag{A.39}$$

where $\boldsymbol{\eta}_j$ represents the j-th eigenvector corresponding to the j-th eigenvalue $\hat{\lambda}_j$ of the transformed Jacobian matrix $\tilde{\boldsymbol{J}}$ defined in (A.30), i.e.,

$$\tilde{\boldsymbol{J}} \boldsymbol{\eta}_j = \hat{\lambda}_j \boldsymbol{\eta}_j, \quad j = 1, 2, \ldots, N.$$

The following lemma shows that the eigenvalues $\hat{\lambda}_j$ of $\tilde{\boldsymbol{J}}$ are the eigenvalues $\lambda_{\boldsymbol{i}}$ of $\tilde{\mathscr{L}}_0$ with $|\boldsymbol{i} = 1|$, with the eigenfunction given by (A.39).

Proposition A.1. *The function* (A.39) *is an eigenfunction of operator $\tilde{\mathscr{L}}_0$ with eigenvalue $\lambda_{\boldsymbol{i}} = \hat{\lambda}_j$ if and only if $\boldsymbol{\eta}_j$ is an eigenvector of the matrix $\tilde{\boldsymbol{J}}$ with the eigenvalue μ_j.*

Proof. The proof for a similar argument is given by Lemma 5 in Liberzon & Brockett (2000), and we present here for completeness. From the previous section, the covariance matrix for the new variables is $\tilde{\boldsymbol{\Sigma}} = \frac{1}{2}\boldsymbol{I}$. By rearranging the steady state variance equation (A.30), we have the relation $\tilde{\boldsymbol{J}} = \boldsymbol{A} - \tilde{\boldsymbol{D}}$, where matrix $\boldsymbol{A} = \frac{1}{2}\left(\tilde{\boldsymbol{J}} - \tilde{\boldsymbol{J}}^T\right)$ is skew-symmetric. Then, using the fact that $\tilde{\mathscr{L}}_0 \rho_{\boldsymbol{0}}(\boldsymbol{z}) = 0$, for $|\boldsymbol{i}| = 1$ and $i_w = 1$, we have

$$\begin{aligned}
\tilde{\mathscr{L}}_0 \rho_{\boldsymbol{i}}(\boldsymbol{z}) &= \sum_{i,j=1}^{N} \tilde{D}_{i,j} h_{w,j} (-2 z_j) \rho_{\boldsymbol{0}}(\boldsymbol{z}) - \sum_{i,j=1}^{N} \tilde{J}_{i,j} z_j h_{w,i} \rho_{\boldsymbol{0}}(\boldsymbol{z}) \\
&= -\rho_{\boldsymbol{0}}(\boldsymbol{z}) \sum_{i,j=1}^{N} (\tilde{J}_{j,i} + 2\tilde{D}_{j,i}) h_{w,j} z_i = \rho_{\boldsymbol{0}}(\boldsymbol{z}) \sum_{i,j=1}^{N} (A_{i,j} - \tilde{D}_{i,j}) h_{w,j} z_i \\
&= \sum_{i,j=1}^{N} \tilde{J}_{i,j} h_{w,j} z_i \rho_{\boldsymbol{0}} = \hat{\lambda}_w \sum_{i=1}^{N} h_i z_i \rho_{\boldsymbol{0}},
\end{aligned}$$

where in the last step we have use the fact that $\boldsymbol{h}_w$ is the eigenvector of $\tilde{J}$ corresponding to the eigenvalue $\hat{\lambda}_w$. Thus, $\hat{\lambda}_w$ is an eigenvalue of $\tilde{\mathscr{L}}_0$ if and only if it is an

eigenvalue of matrix $\tilde{\boldsymbol{J}}$. □

For the more general cases $|\boldsymbol{i}| \geq 1$, we find that the eigenvalue of $\tilde{\mathscr{L}}_0$ is the sum of the eigenvalues of $\tilde{\boldsymbol{J}}$,

$$\lambda_{\boldsymbol{i}} = \sum_{j=1}^{N} i_j \hat{\lambda}_j, \tag{A.40}$$

and the corresponding eigenfunction is of the form

$$\rho_{\boldsymbol{i}}(\boldsymbol{z}) = \sum_{\substack{j_1,j_2,\ldots,j_N=0 \\ j_1+j_2+\cdots+j_N=|\boldsymbol{i}|}}^{|\boldsymbol{i}|} \ddot{T}^{|\boldsymbol{i}|}_{\boldsymbol{i},\boldsymbol{j}} u_{\boldsymbol{j}}(\boldsymbol{z}), \tag{A.41}$$

where $\ddot{\boldsymbol{T}}^s \in \mathbb{R}^{N^s \times N^s}$ is a transition matrix and the function $u_{\boldsymbol{j}}$ has the form

$$u_{\boldsymbol{j}}(\boldsymbol{z}) = \rho_{\boldsymbol{0}}(\boldsymbol{z}) \prod_{w=1}^{N} \left(\boldsymbol{\eta}_w^T \boldsymbol{z}\right)^{j_w}. \tag{A.42}$$

Such form of the eigenfunctions of the linear Fokker-Planck operators was presented in Liberzon & Brockett (2000), and we state it in the following proposition.

Proposition A.2. *The sums of the eigenvalues of matrix $\tilde{\boldsymbol{J}}$, as in* (A.40)*, are the eigenvalues of the Fokker-Planck operator $\tilde{\mathscr{L}}_0$. The corresponding eigenfunction $\rho_{\boldsymbol{i}}$ is a finite linear combinations of functions $u_{\boldsymbol{j}}$ with summation index $|\boldsymbol{j}| \leq |\boldsymbol{i}|$.*

Proof. We first show that operator $\tilde{\mathscr{L}}_0$ acting on the functions $u_{\boldsymbol{i}}$ results in a linear combinations of the functions of the same type. Let us denote the j-th unit vector in $\mathbb{R}^N$ by $\boldsymbol{e}_j$, $j = 1,2,\ldots,N$. Then, using Proposition A.1 and the fact that $\tilde{\mathscr{L}}_0 \rho_{\boldsymbol{0}} = 0$, one

can derive that

$$\begin{aligned}
\tilde{\mathscr{L}}_0 u_{\boldsymbol{i}}(\boldsymbol{z}) &= \sum_{i,j=1}^{N} \sum_{s_1,s_2=1}^{N} i_{s_1} i_{s_2} \tilde{D}_{i,j} \eta_{s_1,i} \eta_{s_2,j} \rho_{\boldsymbol{0}}(\boldsymbol{z}) \prod_{w=1}^{N} \left(\boldsymbol{\eta}_w^T \boldsymbol{z}\right)^{i_w - \mathbb{1}_{w=s_1} - \mathbb{1}_{w=s_2}} \mathbb{1}_{i_w - \mathbb{1}_{w=s_1} - \mathbb{1}_{w=s_2} \geq 0} \\
&\quad + \left(\sum_{w=1}^{N} i_w \hat{\lambda}_w \right) \rho_{\boldsymbol{0}}(\boldsymbol{z}) \prod_{w=1}^{N} \left(\boldsymbol{\eta}_w^T \boldsymbol{z}\right)^{i_w} \\
&= -\frac{1}{2} \sum_{i,j=1}^{N} \sum_{s=1}^{N} i_s^2 \left(\tilde{J}_{i,j} + \tilde{J}_{j,i}\right) \eta_{s,i} \eta_{s,j} \rho_{\boldsymbol{0}}(\boldsymbol{z}) \prod_{w=1}^{N} \left(\boldsymbol{\eta}_w^T \boldsymbol{z}\right)^{i_w - 2\cdot\mathbb{1}_{w=s}} \mathbb{1}_{i_w - 2\cdot\mathbb{1}_{w=s} \geq 0} \\
&\quad + \left(\sum_{w=1}^{N} i_w \hat{\lambda}_w \right) \rho_{\boldsymbol{0}}(\boldsymbol{z}) \prod_{w=1}^{N} \left(\boldsymbol{\eta}_w^T \boldsymbol{z}\right)^{i_w} \\
&= \left(\sum_{w=1}^{N} i_w \hat{\lambda}_w \right) u_{\boldsymbol{i}}(\boldsymbol{z}) - \sum_{s=1}^{N} i_s^2 \hat{\lambda}_s u_{\boldsymbol{i}-2\boldsymbol{e}_s}(\boldsymbol{z}).
\end{aligned} \tag{A.43}$$

The first equality above is adopted from the equation above Theorem 6 in Liberzon & Brockett (2000), and in the second equality, we have used the fact that $\tilde{\boldsymbol{D}} = -\frac{1}{2}(\tilde{\boldsymbol{J}} + \tilde{\boldsymbol{J}}^T)$ and $\boldsymbol{\eta}_{s_1}^T \tilde{\boldsymbol{J}} \boldsymbol{\eta}_{s_2} = 1$ if $s_1 = s_2$, or 0 otherwise. The recursive relation (A.43) implies that the sums of eigenvalues of matrix $\tilde{\boldsymbol{J}}$ are the eigenvalues of operator $\tilde{\mathscr{L}}_0$. The corresponding eigenfunction is a linear combination of the function $u_{\boldsymbol{i}}$ as given in (A.41). □

The values of $\ddot{\boldsymbol{T}}^{|\boldsymbol{i}|}$ can be recursively evaluated as follows: we first set $\rho_{\boldsymbol{i}} = u_{\boldsymbol{i}}$, and $\tilde{\mathscr{L}}_0 \rho_{\boldsymbol{i}}$ contains extra terms $u_{\boldsymbol{i}-2\boldsymbol{e}_s}$ as derived in (A.43). Then we add $\ddot{T}^{|\boldsymbol{i}|}_{\boldsymbol{i},\boldsymbol{i}-2\boldsymbol{e}_s} u_{\boldsymbol{i}-2\boldsymbol{e}_s}$ to $\rho_{\boldsymbol{i}}$ where the value of $\ddot{T}^{|\boldsymbol{i}|}_{\boldsymbol{i},\boldsymbol{i}-2\boldsymbol{e}_s}$ is chosen such that the coefficient in front of $u_{\boldsymbol{i}-2\boldsymbol{e}_s}$ in $\tilde{\mathscr{L}}_0 \rho_{\boldsymbol{i}}$ is $\lambda_{\boldsymbol{i}} \ddot{T}^{|\boldsymbol{i}|}_{\boldsymbol{i},\boldsymbol{i}-2\boldsymbol{e}_s}$. For example, we have

$$\begin{aligned}
\tilde{\mathscr{L}}_0 \left(u_{\boldsymbol{i}}(\boldsymbol{z}) - \sum_{s=1}^{N} \frac{i_s^2 \hat{\lambda}_s}{\lambda_{\boldsymbol{i}} - \lambda_{\boldsymbol{i}-2\boldsymbol{e}_s}} u_{\boldsymbol{i}-2\boldsymbol{e}_s} \right) &= \lambda_{\boldsymbol{i}} \left(u_{\boldsymbol{i}}(\boldsymbol{z}) - \sum_{s=1}^{N} \frac{i_s^2 \hat{\lambda}_s}{\lambda_{\boldsymbol{i}} - \lambda_{\boldsymbol{i}-2\boldsymbol{e}_s}} u_{\boldsymbol{i}-2\boldsymbol{e}_s} \right) \\
&\quad + \sum_{s_1,s_2=1}^{N} \frac{i_{s_1}^2 \hat{\lambda}_{s_1} \hat{\lambda}_{s_2} (\boldsymbol{i} - 2\boldsymbol{e}_{s_1})_{s_2}}{\lambda_{\boldsymbol{i}} - \lambda_{\boldsymbol{i}-2\boldsymbol{e}_s}} u_{\boldsymbol{i}-2\boldsymbol{e}_{s_1}-2\boldsymbol{e}_{s_2}},
\end{aligned} \tag{A.44}$$

and this suggest that, in the case $|\boldsymbol{j}| = |\boldsymbol{i}|$, we have $\ddot{T}^{|\boldsymbol{i}|}_{\boldsymbol{i},\boldsymbol{j}} = 1$ if $\boldsymbol{j} = \boldsymbol{i}$, or 0 otherwise; in the case $|\boldsymbol{j}| = |\boldsymbol{i}| - 1$, we have $\ddot{T}^{|\boldsymbol{i}|}_{\boldsymbol{i},\boldsymbol{j}} = 1$; in the case $|\boldsymbol{j}| = |\boldsymbol{i}| - 2$, we have $\ddot{T}^{|\boldsymbol{i}|}_{\boldsymbol{i},\boldsymbol{j}} =$

$i_s^2\hat{\lambda}_s/(\lambda_{\boldsymbol{i}}-\lambda_{\boldsymbol{j}})$ for $\boldsymbol{i}-\boldsymbol{j}=2\boldsymbol{e}_s$, or 0 otherwise; and in the case $|\boldsymbol{j}|=|\boldsymbol{i}|$, we have $\ddot{T}^{|\boldsymbol{i}|}_{\boldsymbol{i},\boldsymbol{j}}=0$. We continue the procedure to compute $\ddot{T}^{|\boldsymbol{i}|}_{\boldsymbol{i},\boldsymbol{j}}$ until $|\boldsymbol{j}|=0$ such that all the coefficients in front of the lower order terms $\tilde{\mathscr{L}}_0\rho_{\boldsymbol{i}}$ are $\lambda_{\boldsymbol{i}}\ddot{T}^{|\boldsymbol{i}|}_{\boldsymbol{i},\boldsymbol{i}-2\boldsymbol{e}_s}$. Then the resulting function $\rho_{\boldsymbol{i}}$ will be the required eigenfunction satisfying (A.38).

Using the result of Proposition A.2, the following lemma can be proved.

Lemma A.3. *Define a subspace $\mathscr{U}_{N^*}$ of $L^2(\mathbb{R}^N)$ spanned by the eigenfunctions $\{\rho_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\le N^*}$ of the Fokker-Planck operator $\tilde{\mathscr{L}}_0$. Then, functions $\{u_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\le N^*}$ given by* (A.42) *are also the basis functions of space $\mathscr{U}_{N^*}$. And there exists a transition matrix $\ddot{\boldsymbol{T}}^{N^*}\in\mathbb{R}^{N^{N^*}\times N^{N^*}}$ from $\{\rho_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\le N^*}$ to $\{u_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\le N^*}$.*

Proof. It is easy to see that functions $\{u_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\le N^*}$ are linearly independent (not not necessarily orthogonal). Then using the form of the eigenfunction $\{\rho_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\le N^*}$ constructed as in (A.44), we see that the functions $\{\rho_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\le N^*}$ are also linearly independent. This means the dimension of the space $\mathscr{U}_{N^*}$ spanned by $\{\rho_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\le N^*}$ equals to the space $\tilde{\mathscr{U}}_{N^*}$ spanned by $\{u_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\le N^*}$. Our construction (A.44) also implies that the eigenfunction $\rho_{\boldsymbol{i}}$ is a linear combination of functions $\{u_{\boldsymbol{j}}\}_{|\boldsymbol{j}|\le|\boldsymbol{i}|}$. Thus, we have $\mathscr{U}_{N^*}=\tilde{\mathscr{U}}_{N^*}$. The existence of the transition matrix $\ddot{\boldsymbol{T}}^{N^*}$ naturally follows. □

Since the transition matrix $\ddot{\boldsymbol{T}}^{N^*}$ is invertible, we can transform the eigenfunction expansion (A.38) into an expansion of $\{u_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\le N_j^*}$ given by

$$\tilde{\Pi}_{\mathrm{m},j}(\boldsymbol{z})=\sum_{s=0}^{N_j^*}\sum_{\substack{i_1,i_2,\ldots,i_N=0\\ i_1+i_2+\cdots+i_N=s}}^{s}\ddot{c}^{(j)}_{\boldsymbol{i}}u_{\boldsymbol{i}}(\boldsymbol{z}),\tag{A.45}$$

where the constant coefficients $\ddot{c}^{(j)}_{\boldsymbol{i}}$ can be computed via

$$\ddot{c}^{(j)}_{\boldsymbol{i}}=\sum_{s=0}^{N_j^*}\sum_{\substack{j_1,j_2,\ldots,j_N=0\\ j_1+j_2+\cdots+j_N=s}}^{s}\hat{c}^{(j)}_{\boldsymbol{j}}\ddot{T}^{N_j^*}_{\boldsymbol{j},\boldsymbol{i}}.$$

The following lemma will help us to connect the expansion (A.45) to the Hermite polynomial expansion given by (A.35).

Lemma A.4. *The Hermite polynomial functions,* $\left\{H_{\boldsymbol{i}}(\boldsymbol{z}, \frac{1}{2}\boldsymbol{I})\rho_{\boldsymbol{0}}(\boldsymbol{z})\right\}_{|\boldsymbol{i}|\leq N^*}$*, form a basis for the space* $\mathscr{U}_{N^*}$ *defined in Lemma A.3.*

Proof. By definition, the Hermite polynomials are orthogonal, and thus linearly independent. Thus, it is sufficient to prove the argument by showing that there exists a full-rank linear transition matrix $\check{\boldsymbol{T}}^{N^*} \in \mathbb{R}^{N^{N^*}\times N^{N^*}}$ from the functions $\left\{H_{\boldsymbol{i}}\rho_{\boldsymbol{0}}\right\}_{|\boldsymbol{i}|\leq N^*}$ to the functions $\{u_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\leq N^*}$, i.e.,

$$u_{\boldsymbol{i}}^{(j)}(\boldsymbol{z}) = \sum_{s=0}^{N^*} \sum_{\substack{j_1,j_2,\ldots,j_N=0 \\ j_1+j_2+\cdots+j_N=N_j}}^{s} \check{T}_{\boldsymbol{i},\boldsymbol{j}}^{N_j} H_{\boldsymbol{j}}(\boldsymbol{z}, \frac{1}{2}\boldsymbol{I})\rho_{\boldsymbol{0}}(\boldsymbol{z}), \quad |\boldsymbol{i}| \leq N^*. \tag{A.46}$$

To achieve this, we first introduce $\boldsymbol{i}$-th order monomials, denoted by $\omega_{\boldsymbol{i}}(\boldsymbol{z}) = z_1^{i_1}\cdots z_N^{i_N}$. Then, we can construct the transition matrix $\dot{\boldsymbol{T}}^{N^*}$ as

$$\check{\boldsymbol{T}}^{N^*} = \grave{\boldsymbol{T}}^{N^*}\left(\acute{\boldsymbol{T}}^{N^*}\right)^{-1},$$

where matrix $\grave{\boldsymbol{T}}^{N^*} \in \mathbb{R}^{N^{N^*}\times N^{N^*}}$ transforms the basis from $\{\omega_{\boldsymbol{i}}\rho_{\boldsymbol{0}}\}_{|\boldsymbol{i}|\leq N^*}$ to $\{u_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\leq N^*}$, and matrix $\acute{\boldsymbol{T}}^{N^*} \in \mathbb{R}^{N^{N^*}\times N^{N^*}}$ transforms the basis from $\{\omega_{\boldsymbol{i}}\rho_{\boldsymbol{0}}\}_{|\boldsymbol{i}|\leq N^*}$ to $\{H_{\boldsymbol{i}}\rho_{\boldsymbol{0}}\}_{|\boldsymbol{i}|\leq N^*}$.

The elements of the transition matrix $\grave{\boldsymbol{T}}^{N^*}$ can be implied from the form of functions $u_{\boldsymbol{i}}$ given in (A.42). Multiplying out the brackets, we see that functions $u_{\boldsymbol{i}}$ is are linear combinations of functions $\omega_{\boldsymbol{j}}\rho_{\boldsymbol{0}}$ with $|\boldsymbol{j}| \leq |\boldsymbol{i}|$. Then, using the fact that monomials are linearly independent, we have that $\{\omega_{\boldsymbol{i}}\rho_{\boldsymbol{0}}\}_{|\boldsymbol{i}|\leq N^*}$ forms a basis of space $\mathscr{U}^{N^*}$.

The elements of the transition matrix $\acute{\boldsymbol{T}}^{N^*}$ are given by the explicit expression of the Hermite function provided in (A.27). Then using the similar argument as above, one can show that $\{H_{\boldsymbol{i}}\rho_{\boldsymbol{0}}\}_{|\boldsymbol{i}|\leq N^*}$ also form a basis of the space spanned by functions $\{\omega_{\boldsymbol{i}}\rho_{\boldsymbol{0}}\}_{|\boldsymbol{i}|\leq N^*}$, i.e., $\mathscr{U}^{N^*}$. □

Using the results of Lemmas A.3 and A.4, we can now state the equivalence between the Hermite expansion (A.35) and the eigenfunction expansion (A.38).

Proposition A.5. *The expanded coefficient* $\tilde{\Pi}_{\mathrm{m},j}(\boldsymbol{z})$ *of* (A.20) *admits the Hermite expansion* (A.35) *with some finite value of* N_j*, if and only if it admits the eigenfunction expansion* (A.38) *with* $N_j^* = N_j$.

Proof. Combining Lemmas A.3 and A.4, we have that both $\{\rho_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\leq N_j}$ and $\{H_{\boldsymbol{i}}\rho_{\boldsymbol{0}}\}_{|\boldsymbol{i}|\leq N_j}$ are basis functions of the same subspace $\mathscr{U}^{N_j}$ of $L^2(\mathbb{R}^N)$. Thus, if $\tilde{\Pi}_{\mathrm{m},j} \in \mathscr{U}^{N_j}$, it can be written as a linear combinations of both types of the basis functions. □

Using the result of Proposition A.5, we can derive the transition matrix $\bar{\boldsymbol{T}}^{N_j} \in \mathbb{R}^{N^{N_j}\times N^{N_j}}$ is the transition matrix from the Hermite polynomial basis $\{H_{\boldsymbol{i}}\rho_{\boldsymbol{0}}\}_{|\boldsymbol{i}|\leq N_j}$ to the eigenfunction basis $\{\rho_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\leq N_j}$, i.e., satisfying

$$\rho_{\boldsymbol{i}}(\boldsymbol{z}) = \sum_{s=0}^{N_j} \sum_{\substack{j_1,j_2,\ldots,j_N=0 \\ j_1+j_2+\cdots+j_N=N_j}}^{s} \bar{T}^{N_j}_{\boldsymbol{i},\boldsymbol{j}} H_{\boldsymbol{i}}(\boldsymbol{z}, \frac{1}{2}\boldsymbol{I})\rho_{\boldsymbol{0}}(\boldsymbol{z}),$$

and its elements can be computed as

$$\bar{\boldsymbol{T}}^{N_j} = \ddot{\boldsymbol{T}}^{N_j}\check{\boldsymbol{T}}^{N_j}, \tag{A.47}$$

where matrix $\ddot{\boldsymbol{T}}^{N_j}$ is given in (A.41) and the transition matrix $\dot{\boldsymbol{T}}^{N_j}$ is given in (A.46). Subsequently, given the eigenfunction expansion (A.38) with coefficients $\left\{\hat{c}^{(j)}_{\boldsymbol{i}}\right\}_{|\boldsymbol{i}|<N_j}$, we can obtain the corresponding Hermite expansion (A.35) with coefficients $\left\{\tilde{c}^{(j)}_{\boldsymbol{i}}\right\}_{|\boldsymbol{i}|<N_j}$ via

$$\tilde{c}^{(j)}_{\boldsymbol{i}} = \sum_{s=0}^{N_j} \sum_{\substack{j_1,j_2,\ldots,j_N=0 \\ j_1+j_2+\cdots+j_N=N_j}}^{s} \hat{c}^{(j)}_{\boldsymbol{j}} \bar{T}^{N_j}_{\boldsymbol{j},\boldsymbol{i}}. \tag{A.48}$$

A.6 The equations for the expansion coefficients

In this section, we derive the coefficients, $\tilde{\Pi}_{\mathrm{m},j}$, of the Hermite expansion (A.35) using the eigenfunction approach.

We first multiply $\frac{1}{\boldsymbol{n}!}\int_{\mathbb{R}^N}\mathrm{d}\boldsymbol{z}H_{\boldsymbol{n}}(\boldsymbol{z},\frac{1}{2}\boldsymbol{I})$ to both sides of (A.32), and obtain

$$\frac{1}{\boldsymbol{n}!}\int_{\mathbb{R}^N}H_{\boldsymbol{n}}(\boldsymbol{z},\frac{1}{2}\boldsymbol{I})\tilde{\mathscr{L}}_0\tilde{\Pi}_{\mathrm{m},j}(\boldsymbol{z})\mathrm{d}\boldsymbol{z}=\frac{1}{\boldsymbol{n}!}\sum_{i=0}^{j-1}\int_{\mathbb{R}^N}H_{\boldsymbol{n}}(\boldsymbol{z})\tilde{\mathscr{L}}_{j-i}\tilde{\Pi}_{\mathrm{m},i}(\boldsymbol{z})\mathrm{d}\boldsymbol{z}. \tag{A.49}$$

For the right-hand-side of (A.49), we insert the eigenfunction expansion (A.38) which yields

$$\frac{1}{\boldsymbol{n}!}\int_{\mathbb{R}^N}H_{\boldsymbol{n}}(\boldsymbol{z},\frac{1}{2}\boldsymbol{I})\tilde{\mathscr{L}}_0\tilde{\Pi}_{\mathrm{m},j}(\boldsymbol{z})\mathrm{d}\boldsymbol{z}=\sum_{s=0}^{N_j}\sum_{\substack{i_1,i_2,\ldots,i_N=0\\ i_1+i_2+\cdots+i_N=s}}^{s}\hat{c}_{\boldsymbol{i}}^{(j)}\lambda_{\boldsymbol{i}}\frac{1}{\boldsymbol{n}!}\int_{\mathbb{R}^N}H_{\boldsymbol{n}}(\boldsymbol{z},\frac{1}{2}\boldsymbol{I})\rho_{\boldsymbol{i}}(\boldsymbol{z})\mathrm{d}\boldsymbol{z},$$

where we have used the fact that $\rho_{\boldsymbol{i}}$ is the eigenfunction corresponding to eigenvalue $\lambda_{\boldsymbol{i}}$ of the Fokker-Plank operator $\tilde{\mathscr{L}}_0$. Then, using the result from Proposition A.5, we transform the eigenfunctions of the above formula into the Hermite polynomials, and we have

$$\frac{1}{\boldsymbol{n}!}\int_{\mathbb{R}^N}H_{\boldsymbol{n}}(\boldsymbol{z},\frac{1}{2}\boldsymbol{I})\tilde{\mathscr{L}}_0\tilde{\Pi}_{\mathrm{m},j}(\boldsymbol{z})\mathrm{d}\boldsymbol{z}=\sum_{s=0}^{N_j}\sum_{\substack{i_1,i_2,\ldots,i_N=0\\ i_1+i_2+\cdots+i_N=s}}^{s}\bar{c}_{\boldsymbol{i}}^{(j)}\frac{1}{\boldsymbol{n}!}\int_{\mathbb{R}^N}H_{\boldsymbol{n}}(\boldsymbol{z},\frac{1}{2}\boldsymbol{I})H_{\boldsymbol{i}}(\boldsymbol{z},\frac{1}{2}\boldsymbol{I})\rho_{\boldsymbol{0}}(\boldsymbol{z})\mathrm{d}\boldsymbol{z},$$

where the constant coefficients $\left\{\bar{c}_{\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\le N_j}$ are linear combinations of the coefficients $\left\{\hat{c}_{\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\le N_j}$ as

$$\bar{c}_{\boldsymbol{i}}^{(j)}=\sum_{s=0}^{N_j}\sum_{\substack{j_1,j_2,\ldots,j_N=0\\ j_1+j_2+\cdots+j_N=N_j}}^{s}\lambda_{\boldsymbol{j}}\hat{c}_{\boldsymbol{j}}^{(j)}\bar{T}_{\boldsymbol{j},\boldsymbol{i}}^{N_j}. \tag{A.50}$$

Then, inserting the orthogonality condition (A.26) of the Hermite polynomials into

above formula yields

$$\frac{1}{\boldsymbol{n}!}\int_{\mathbb{R}^N} H_{\boldsymbol{n}}(\boldsymbol{z})\tilde{\mathscr{L}}_0\tilde{\Pi}_{\mathrm{m},j}(\boldsymbol{z})\mathrm{d}\boldsymbol{z} = \begin{cases} \bar{c}_{\boldsymbol{n}}^{(j)} & \text{if } |\boldsymbol{n}| \le N_j \\ 0 & \text{otherwise.} \end{cases} \tag{A.51}$$

For the right-hand-side of (A.49), we insert the Hermite expansion (A.35) and obtain

$$\begin{aligned} &\sum_{i=0}^{j-1}\frac{1}{\boldsymbol{n}!}\int_{\mathbb{R}^N} H_{\boldsymbol{n}}(\boldsymbol{z})\tilde{\mathscr{L}}_{j-i}\tilde{\Pi}_{\mathrm{m},i}(\boldsymbol{z})\mathrm{d}\boldsymbol{z} \\ &= \sum_{i=0}^{j-1}\sum_{s=0}^{N_i}\sum_{\substack{m_1,m_2,\ldots,m_N=0 \\ m_1+m_2+\cdots+m_N=s}}^{s} \tilde{c}_{\boldsymbol{m}}^{(i)} \sum_{w=0}^{\lceil (j-i)/2\rceil}\sum_{k=1}^{j-i-2(w-1)}\sum_{\substack{i_1,\ldots,i_N=0 \\ i_1+\cdots+i_N=k}}^{k}\sum_{\substack{j_1,\ldots,j_N=0 \\ j_1+\cdots+j_N=j-i-k-2(w-1)}}^{j-i-k-2(w-1)} \\ &\times \mathscr{F}_{k,j-i-k-2(w-1),w}^{\boldsymbol{i},\boldsymbol{j}}\mathscr{G}_{\boldsymbol{n},\boldsymbol{m}}^{\boldsymbol{i},\boldsymbol{j}}, \end{aligned} \tag{A.52}$$

where

$$\mathscr{G}_{\boldsymbol{n},\boldsymbol{m}}^{\boldsymbol{i},\boldsymbol{j}} = \frac{1}{\boldsymbol{n}!}\int_{\mathbb{R}^N} H_{\boldsymbol{n}}(\boldsymbol{z},\frac{1}{2}\boldsymbol{I})\left(\frac{\partial}{\partial \boldsymbol{z}}\right)^{\boldsymbol{i}}\boldsymbol{z}^{\boldsymbol{j}}H_{\boldsymbol{m}}(\boldsymbol{z},\frac{1}{2}\boldsymbol{I})\tilde{\Pi}_0(\boldsymbol{z},\frac{1}{2}\boldsymbol{I})\mathrm{d}\boldsymbol{z}. \tag{A.53}$$

Note that for notation simplicity, we use $b_{\boldsymbol{n}}^{(j)}$ to represent the right-hand-side of (A.52).

Combining (A.51) and (A.52), we have derived from (A.49) to obtain the values of the coefficients $\left\{\bar{c}_{\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\le N_j}$ to be

$$\bar{c}_{\boldsymbol{n}}^{(j)} = b_{\boldsymbol{n}}^{(j)}, \tag{A.54}$$

for $|\boldsymbol{n}| \le N_j$. Assuming for the moment that N_j is a finite number, (A.54) yields a finite set of equations. Then, inverting the transformation (A.50), we can derive the

coefficients $\left\{\hat{c}_{\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\leq N_j}$ of the eigenfunction expansion (A.38) satisfy

$$\hat{\boldsymbol{c}}^{(j)} = \left(\boldsymbol{\Lambda}^{N_j}\right)^{-1}\left(\left(\bar{\boldsymbol{T}}^{N_j}\right)^T\right)^{-1}\boldsymbol{b}^{(j)}, \tag{A.55}$$

where we have used the $\boldsymbol{\Lambda}^{N_j} \in \mathbb{R}^{N^{N_j}\times N^{N_j}}$ is a diagonal matrix with $\{\lambda_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\leq N_j}$ being its diagonal elements. We can also substitute (A.48) into (A.55), and derive the coefficient values of $\left\{\tilde{c}_{\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\leq N_j}$ in the Hermite expansion (A.35) as

$$\tilde{\boldsymbol{c}}^{(j)} = \left(\bar{\boldsymbol{T}}^{N_j}\right)^T\left(\boldsymbol{\Lambda}^{N_j}\right)^{-1}\left(\left(\bar{\boldsymbol{T}}^{N_j}\right)^T\right)^{-1}\boldsymbol{b}^{(j)}. \tag{A.56}$$

We can further substitute (A.36) into (A.56) and derive that the coefficients $\left\{c_{\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\leq N_j}$ of the original Hermite expansion (A.24) satisfy

$$\boldsymbol{c}^{(j)} = \left(\dot{\boldsymbol{T}}^{N_j}\right)^T\left(\bar{\boldsymbol{T}}^{N_j}\right)^T\left(\boldsymbol{\Lambda}^{N_j}\right)^{-1}\left(\left(\bar{\boldsymbol{T}}^{N_j}\right)^T\right)^{-1}\boldsymbol{b}^{(j)}. \tag{A.57}$$

It is important to note that the evaluations of the coefficient values in (A.55), (A.56) and (A.57) are possible only if

i) N_j are finite numbers for $j = 1,2,\ldots$, and

ii) the elements $b_{\boldsymbol{n}}^{(j)}$ can be evaluated.

We will prove these properties in the remaining part of the section.

We first derive an explicit expression for the integral $\mathscr{G}_{\boldsymbol{n},\boldsymbol{m}}^{\boldsymbol{i},\boldsymbol{j}}$ defined in (A.53). Owing to the separability of the Hermite function $H_{\boldsymbol{n}}\left(\boldsymbol{z},\frac{1}{2}\boldsymbol{I}\right)$, we can split the multiple integral as a product of single integrals as

$$\mathscr{G}_{\boldsymbol{n},\boldsymbol{m}}^{\boldsymbol{i},\boldsymbol{j}} = \frac{1}{\boldsymbol{n}!}\prod_{w=1}^{N}\tilde{\mathscr{G}}_{n_w,m_w}^{i_w,j_w},$$

and the single integral $\tilde{\mathcal{G}}_{n,m}^{i,j}$ is given by

$$\tilde{\mathcal{G}}_{n,m}^{i,j} = \int_{\mathbb{R}} H_n(z,1/2)(\partial/\partial z)^i z^j H_m(z,1/2)\tilde{\Pi}_0(z,1/2)\mathrm{d}z, \tag{A.58}$$

where $\tilde{\Pi}_0(z,1/2)$ denotes a centred Gaussian with variance 1/2, and $H_n(z,1/2)$ denotes the n-th univariate Hermite polynomial functions with respect to variance 1/2, also known as the physicists' Hermite polynomials. Using the properties of the Hermite polynomials, the following lemma can be easily proved.

Lemma A.6. *Consider $\tilde{\mathcal{G}}_{n,m}^{i,j}$ defined in* (A.58)*, its explicit expression is given by*

$$\tilde{\mathcal{G}}_{n,m}^{i,j} = \frac{j!}{2^j}\sqrt{\frac{m+n-i}{m!(n-i)!}}\sum_{s=\max(0,-h)}^{\min(m,n-i)}\binom{n-i}{s}\binom{m}{s}\frac{s!}{2^s(h+s)}, \quad \text{with} \quad h=\frac{j-m-n+i}{2}, \tag{A.59}$$

under the conditions that:

i) *the quantity $i+j-m-n$ is even,*

ii) $2\min(m,n-i) \geq m+n-i-j$.

Otherwise, we have $\tilde{\mathcal{G}}_{n,m}^{i,j}=0$.

Proof. Applying integration by parts to (A.58), we have

$$\tilde{\mathcal{G}}_{n,m}^{i,j} = \int_{\mathbb{R}} H_{n-i}(z,1/2) z^j H_m(z,1/2)\tilde{\Pi}_0(z,1/2)\mathrm{d}z.$$

According to Silverman et al. (1972), we have the explicit formula given by (A.59), with condition i). The condition ii) is then derived by requiring the upper bound of the summation is greater or equal to the lower bound. □

Lemma A.7. *The integral $\mathcal{G}_{\boldsymbol{n},\boldsymbol{m}}^{\boldsymbol{i},\boldsymbol{j}}$ is zero if $|\boldsymbol{n}| > |\boldsymbol{m}|+|\boldsymbol{i}|+|\boldsymbol{j}|$.*

Proof. The integral $\mathcal{G}_{\boldsymbol{n},\boldsymbol{m}}^{\boldsymbol{i},\boldsymbol{j}}$ is non-zero only if all single integrals $\tilde{\mathcal{G}}_{n_w,m_w}^{i_w,j_w}$, $w=1,2,\ldots,N$,

are non-zero. Using condition ii) of Lemma A.6, we have

$$n_w \leq m_w + i_w + j_w.$$

Summing up the above equality for all $1 \leq w \leq N$ yields the condition for non-trivial values of the integral $\mathscr{G}^{\boldsymbol{i},\boldsymbol{j}}_{\boldsymbol{n},\boldsymbol{m}}$, and the argument immediately follows. □

Now, we can show that the number of in the expansions, (A.35) and (A.38), is finite, and (A.54) yields a finite number of equations.

Proposition A.8. *The expansion coefficient,* $\tilde{\Pi}_{\mathrm{m},j}$*, in* (A.31) *can be represented as a sum of a finite number of Hermite functions of the form* (A.35)*. Moreover, the summation index satisfies* $N_j = 3j$*.*

Proof. To prove the number of the terms in the expansion is finite, it suffices to prove that there exists a finite value of N_j such that the coefficients $\tilde{c}^{(j)}_{\boldsymbol{i}} = 0$ for all $|\boldsymbol{i}| > N_j$. In fact, using Lemma A.7, one can easily derive that the right-hand-side of (A.52) is zero if $|\boldsymbol{n}| > N_{j-1} + 3$. Subsequently, the coefficients $\bar{c}^{(j)}_{\boldsymbol{i}}$ are also zero for all $|\boldsymbol{n}| > N_{j-1} + 3$, which implies that $N_j = N_{j-1} + 3$. Using the initial condition that $N_0 = 0$, we immediately have $N_j = 3j$. □

B

A perturbative analysis of the CFPE

Next, the same strategy is applied to solve an perturbative solution of the stationary CFPE (2.10). As before, we first apply the system size expansion to the CFPE and transform the distribution of populations into the distribution of the fluctuations around the deterministic mean. The distribution of fluctuations is expanded in the inverse powers of the square roots of volume V, and the expanded coefficients are sought using the Hermite expansion and the eigenvalue expansion.

B.1 System size expansion of the CFPE

We perform the system size expansion to the CFPE (2.10). As before, we transform model variables from the continuous molecular populations $\boldsymbol{x}$ to fluctuations $\boldsymbol{\epsilon}$ through $x_i/V = \phi_i + \sqrt{V}\epsilon_i,\ i = 1,2,\ldots,N$. Then, the CFPE distribution $P_{\mathrm{f}}(\boldsymbol{x},t)$ is transformed into a distribution of the fluctuations, written as $\Pi_{\mathrm{f}}(\boldsymbol{\epsilon},t)$, under the scal-

ing relationship

$$\Pi_{\mathrm{f}}(\boldsymbol{\epsilon}, t) = V^{N/2} P_{\mathrm{f}}(\boldsymbol{x}, t).$$

To fully expand the CFPE, one needs to transform the involved differential operators and propensity functions into the new variables. The operators involving derivatives comes from the truncated Taylor expansion (A.3) of the step operators, thus we can write the transformed operators as

$$\sum_{i=1}^{N} \nu_{j,i} \frac{\partial}{\partial x_i} = V^{-1/2} a_j^{(1)}, \quad \text{and} \quad \sum_{i,i'=1}^{N} \nu_{j,i} \nu_{j,i'} \frac{\partial^2}{\partial x_i \partial x_{i'}} = V^{-1} a_j^{(2)}, \tag{B.1}$$

for $j = 1, 2, \ldots, M$. The propensity function is expanded in the same way as we did in (A.7). Then, substituting (B.1) and (A.7) into (2.10) and rearranging the terms in the inverse orders of $\sqrt{V}$, we obtain the SSE expansion in the new variables as

$$\begin{aligned} \frac{\partial \Pi_{\mathrm{f}}(\boldsymbol{\epsilon}, t)}{\partial t} =& V^0 \sum_{j=1}^{N} \left(a_j^{(2)} b_j^{(0,0)} - a_j^{(1)} b_j^{(1,0)} \right) \Pi_{\mathrm{f}}(\boldsymbol{\epsilon}, t) \\ &+ V^{-1/2} \sum_{j=1}^{M} \left(a_j^{(2)} b_j^{(1,0)} - a_j^{(1)} b_j^{(2,0)} - a_j^{(1)} b_j^{(0,2)} \right) \Pi_{\mathrm{f}}(\boldsymbol{\epsilon}, t) \\ &+ V^{-1} \sum_{j=1}^{M} \left(a_j^{(2)} b_j^{(2,0)} + a_j^{(2)} b_j^{(0,2)} - a_j^{(1)} b_j^{(1,2)} \right) \Pi_{\mathrm{f}}(\boldsymbol{\epsilon}, t) + \mathcal{O}\left(V^{-3/2} \right), \end{aligned} \tag{B.2}$$

where the order $V^{1/2}$ terms cancels due to the fact that $\boldsymbol{\phi}$ are the solution of the deterministic reaction rate equations. We can equivalently write (B.2) as

$$\frac{\partial \Pi_{\mathrm{f}}(\boldsymbol{\epsilon}, t)}{\partial t} = \sum_{s=0}^{\infty} V^{-s/2} \mathscr{L}_{\mathrm{f},s} \Pi_{\mathrm{f}}(\boldsymbol{\epsilon}, t), \tag{B.3}$$

where operator $\mathscr{L}_{\mathrm{f},s}$ corresponds to the collection of the terms acting on Π_f in (B.2) that are proportional to $V^{-s/2}$, i.e.,

$$\begin{aligned}\mathscr{L}_{\mathrm{f},s} &= \sum_{j=1}^{M}\left(\sum_{w=0}^{\lceil s/2\rceil}\sum_{k=1}^{\min(s-2(w-1),2)}(-1)^k a_j^{(k)} b_j^{(s-k-2(w-1),2w)}\right)\\ &= \sum_{w=0}^{\lceil s/2\rceil}\sum_{k=1}^{\min(s-2(w-1),2)}\sum_{\substack{i_1,\ldots,i_N=0\\ i_1+\cdots+i_N=k}}^{k}\sum_{\substack{j_1,\ldots,j_N=0\\ j_1+\cdots+j_N=s-k-2(w-1)}}^{s-k-2(w-1)}\mathscr{D}^{\boldsymbol{i},\boldsymbol{j}}_{k,s-k-2(w-1),w}\left(\frac{\partial}{\partial\boldsymbol{\epsilon}}\right)^{\boldsymbol{i}}\boldsymbol{\epsilon}^{\boldsymbol{j}},\end{aligned}\tag{B.4}$$

where $\mathscr{D}^{\boldsymbol{i},\boldsymbol{j}}_{k,q,w}$ is defined in (A.9).

The CFPE in the expanded form (B.3) suggests one may expand the probability distribution Π_f in an inverse power series of square root of volume V, i.e.,

$$\Pi_\mathrm{f}(\boldsymbol{\epsilon},t) = \sum_{j=0}^{\infty}\Pi_{\mathrm{f},j}(\boldsymbol{\epsilon},t)V^{-j/2},\tag{B.5}$$

which is the CFPE counterpart of the expansion in (A.12). The subscript, $\cdot_\mathrm{f}$, is used here to distinguish from the expanded coefficients for the CME. The evolution of the expanded coefficients $\Pi_{\mathrm{f},j}$ can be derived by, first, substituting the expansion (B.5) into (B.3), which yields

$$\sum_{j=0}^{\infty}V^{-j/2}\frac{\partial\Pi_{\mathrm{f},j}(\boldsymbol{\epsilon},t))}{\partial t} = \sum_{j=0}^{\infty}V^{-j/2}\sum_{w=0}^{j}\mathscr{L}_{\mathrm{f},j-w}\Pi_{\mathrm{f},w}(\boldsymbol{\epsilon},t),$$

Then, collecting the terms that correspond to the same inverse orders of $\sqrt{V}$ gives the equation for the expanded coefficients as

$$\left(\frac{\partial}{\partial t}-\mathscr{L}_0\right)\Pi_{\mathrm{f},j}(\boldsymbol{\epsilon},t) = \sum_{i=0}^{j-1}\mathscr{L}_{\mathrm{f},j-i}\Pi_{\mathrm{f},i}(\boldsymbol{\epsilon},t),\tag{B.6}$$

which obeys the normalisation conditions

$$\int_{\mathbb{R}^N}\Pi_{\mathrm{f},0}(\boldsymbol{\epsilon},t)\mathrm{d}\boldsymbol{\epsilon}=1\quad\text{and}\quad\int_{\mathbb{R}^N}\Pi_{\mathrm{f},j}(\boldsymbol{\epsilon},t)\mathrm{d}\boldsymbol{\epsilon}=0,\quad j=1,2,3,\ldots.$$

B.2 Hermit expansion of the expanded coefficients

As before, we assume that the deterministic reaction rate equation only has one stable fixed point, which allows us to analysing the steady state distribution by setting the time derivatives in previous equations to zero. Let us define

$$\Pi_{\mathrm{f}}(\boldsymbol{\epsilon}) = \lim_{t\to\infty} \Pi_{\mathrm{f}}(\boldsymbol{\epsilon}, t) \quad \text{and} \quad \Pi_{\mathrm{f},j}(\boldsymbol{\epsilon}) = \lim_{t\to\infty} \Pi_{\mathrm{f},j}(\boldsymbol{\epsilon}, t),$$

and rewrite (B.5) as

$$\Pi_{\mathrm{f}}(\boldsymbol{\epsilon}) = \sum_{j=0}^{\infty} \Pi_{\mathrm{f},j}(\boldsymbol{\epsilon}) V^{-j/2}, \tag{B.7}$$

with normalisation conditions $\int_{\mathbb{R}^N} \Pi_{\mathrm{f},0}(\boldsymbol{\epsilon})\mathrm{d}\boldsymbol{\epsilon} = 1$ and $\int_{\mathbb{R}^N} \Pi_{\mathrm{f},j}(\boldsymbol{\epsilon})\mathrm{d}\boldsymbol{\epsilon} = 0$ with $j \geq 1$.

The expanded coefficients, $\Pi_{\mathrm{f},j}$, in (B.7) solves the following elliptic equation

$$\mathscr{L}_{\mathrm{f},0}\Pi_{\mathrm{f},j}(\boldsymbol{\epsilon}) = \sum_{i=0}^{j-1} \mathscr{L}_{j-i}\Pi_{\mathrm{f},i}(\boldsymbol{\epsilon}), \tag{B.8}$$

where operators $\mathscr{L}_{\mathrm{f},j}$ are given in (B.4). The system of partial differential equation (B.8) is solved by considering the multivariate Hermite expansion of the expanded coefficients:

$$\Pi_{\mathrm{f},j}(\boldsymbol{\epsilon}) = \sum_{s=0}^{M_j} \sum_{\substack{i_1,i_2,\ldots,i_N=0 \\ i_1+i_2+\cdots+i_N=s}}^{s} c_{\mathrm{f},\boldsymbol{i}}^{(j)} H_{\boldsymbol{i}}(\boldsymbol{\epsilon}, \boldsymbol{\Sigma}) \Pi_0(\boldsymbol{\epsilon}, \boldsymbol{\Sigma}), \tag{B.9}$$

where $c_{\mathrm{f},\boldsymbol{i}}^{(j)}$ are constant coefficients, M_j denotes its maximum summation index, and $H_{\boldsymbol{i}}(\boldsymbol{\epsilon}, \boldsymbol{\Sigma})$ represents the $\boldsymbol{i}$-th multivariate Hermite polynomial function with respect to $\boldsymbol{\Sigma}$ as defined in (A.25). In (B.9), we have $\Pi_0 = \Pi_{\mathrm{f},0}$, where $\Pi_{\mathrm{f},0}$ satisfies the stationary equation $\mathscr{L}_{\mathrm{f},0}\Pi_{\mathrm{f},0}(\boldsymbol{\epsilon}) = 0$, and operator $\mathscr{L}_{\mathrm{f},0} = \mathscr{L}_0$ is also the linear Fokker-Planck operator of the LNA. Thus, $\Pi_{\mathrm{f},0}$ is a centred Gaussian distribution with covariance

$\boldsymbol{\Sigma}$, i.e.,

$$\Pi_{\mathrm{f},0}(\boldsymbol{\epsilon}) = \Pi_{\mathrm{m},0}(\boldsymbol{\epsilon}) = \frac{1}{\sqrt{(2\pi)^N |\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}\boldsymbol{\epsilon}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\epsilon}\right), \tag{B.10}$$

where $\boldsymbol{\Sigma}$ satisfies the Lyapunov equation (A.23).

To evaluate the constant coefficients $\left\{c_{\mathrm{f},\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\le M_j}$, we first perform the transformation of coordinate as we did in Section § A.4. This is done by introducing the transition matrix $\boldsymbol{T} \in \mathbb{R}^{N\times N}$ and defining the new variables $\boldsymbol{z} = \boldsymbol{T}\boldsymbol{\epsilon}$ as in (A.28). Upon coordinate transformation, the coefficient $\Pi_{\mathrm{f},0}$ in (B.10) is transformed into a centred Gaussian distribution, denoted by $\tilde{\Pi}_{\mathrm{f},0}$, with covariance matrix being $\frac{1}{2}\boldsymbol{I}$. And the Hermite expansions for higher order coefficients are transformed to

$$\tilde{\Pi}_{\mathrm{f},j}(\boldsymbol{\epsilon}) = \sum_{s=0}^{M_j} \sum_{\substack{i_1,i_2,\ldots,i_N=0 \\ i_1+i_2+\cdots+i_N=s}}^{s} \tilde{c}_{\mathrm{f},\boldsymbol{i}}^{(j)} H_{\boldsymbol{i}}(\boldsymbol{\epsilon}, \frac{1}{2}\boldsymbol{I}) \Pi_0(\boldsymbol{\epsilon}, \frac{1}{2}\boldsymbol{I}), \tag{B.11}$$

and the coefficients $\left\{\tilde{c}_{\mathrm{f},\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\le M_j}$ can be linearly combined to get values of $\left\{c_{\mathrm{f},\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\le M_j}$ in (B.11) using

$$c_{\boldsymbol{i}}^{(j)} = \sum_{\substack{j_1,j_2,\ldots,j_N=0 \\ j_1+j_2+\cdots+j_N=M_j}}^{M_j} \tilde{c}_{\boldsymbol{j}}^{(j)} \dot{T}_{\boldsymbol{j},\boldsymbol{i}}^{M_j}, \tag{B.12}$$

where $\dot{\boldsymbol{T}}^{M_j} \in \mathbb{R}^{N^{M_j}\times N^{M_j}}$ is a transition matrix from the Hermite polynomial basis $\left\{H_{\boldsymbol{i}}(\boldsymbol{z}, \frac{1}{2}\boldsymbol{I})\right\}_{|\boldsymbol{i}|\le M_j}$ to the basis $\{H_{\boldsymbol{i}}(\boldsymbol{\epsilon}, \boldsymbol{\Sigma})\}_{|\boldsymbol{i}|\le M_j}$, and the elements of the transition matrix is defined in (A.37).

The elliptic equation (B.8) in the new variables are given by

$$\tilde{\mathscr{L}}_{\mathrm{f},0}\tilde{\Pi}_{\mathrm{f},j}(\boldsymbol{z}) = \sum_{i=0}^{j-1} \tilde{\mathscr{L}}_{j-i}\tilde{\Pi}_{\mathrm{f},i}(\boldsymbol{z}), \tag{B.13}$$

where the explicit expression of operator $\tilde{\mathscr{L}}_{\mathrm{f},s}$ is given by

$$\tilde{\mathscr{L}}_{\mathrm{f},s} = \sum_{w=0}^{\lceil s/2 \rceil} \sum_{k=1}^{\min(s-2(w-1),2)} \sum_{\substack{i_1,\ldots,i_N=0 \\ i_1+\cdots+i_N=k}}^{k} \sum_{\substack{j_1,\ldots,j_N=0 \\ j_1+\cdots+j_N=s-k-2(w-1)}}^{s-k-2(w-1)} \mathscr{F}^{\boldsymbol{i},\boldsymbol{j}}_{k,s-k-2(w-1),w} \left(\frac{\partial}{\partial \boldsymbol{z}}\right)^{\boldsymbol{i}} \boldsymbol{z}^{\boldsymbol{j}},$$

and the coefficients $\mathscr{F}^{\boldsymbol{i},\boldsymbol{j}}_{k,q,w}$ are same as in (A.33).

As we discussed in Section § A.5, we need the eigenfunction expansion for solving (B.8), and thus we write the eigenfunction expansion of $\tilde{\Pi}_{\mathrm{f},j}$ as

$$\tilde{\Pi}_{\mathrm{f},j}(\boldsymbol{z}) = \sum_{s=0}^{M_j} \sum_{\substack{i_1,i_2,\ldots,i_N=0 \\ i_1+i_2+\cdots+i_N=s}}^{s} \hat{c}^{(j)}_{\mathrm{f},\boldsymbol{i}} \rho_{\boldsymbol{i}}(\boldsymbol{z}), \tag{B.14}$$

where $\rho_{\boldsymbol{i}}$ is the $\boldsymbol{i}$-th eigenfunction of the linear Fokker-Planck operator $\tilde{\mathscr{L}}_{\mathrm{f},0}$, or equivalently $\tilde{\mathscr{L}}_0$, corresponding to the eigenvalue $\lambda_{\boldsymbol{i}} = \sum_{s=1}^{N} i_s \lambda_s$, and λ_s is the s-th eigenvalue of the transformed Jacobian matrix $\tilde{\boldsymbol{J}}$ in (A.30). The form of eigenfunctions $\rho_{\boldsymbol{i}}$ is given in Proposition A.2. We have also shown, in Proposition A.5, that the eigenfunctions $\{\rho_i\}_{|\boldsymbol{i}|\le M_j}$ in (B.14) form a basis for the same subspace $\mathscr{U}_{M_j}$, of which the Hermite polynomials $\{H_{\boldsymbol{i}}(\boldsymbol{z}, \frac{1}{2}\boldsymbol{I})\rho_{\boldsymbol{0}}(\boldsymbol{z})\}_{|\boldsymbol{i}|\le M_j}$ in (B.11) are also its basis functions. In (A.47), we derived the transition matrix $\bar{\boldsymbol{T}}^{M_j}$ from the basis $\{H_{\boldsymbol{i}}(\boldsymbol{z}, \frac{1}{2}\boldsymbol{I})\rho_{\boldsymbol{0}}(\boldsymbol{z})\}_{|\boldsymbol{i}|\le M_j}$ to the basis $\{\rho_i\}_{|\boldsymbol{i}|\le M_j}$, and we can transform the coefficients $\left\{\hat{c}^{(j)}_{\mathrm{f},\boldsymbol{i}}\right\}_{|\boldsymbol{i}|\le M_j}$ to the coefficients $\left\{\tilde{c}^{(j)}_{\mathrm{f},\boldsymbol{i}}\right\}_{|\boldsymbol{i}|\le M_j}$ using

$$\tilde{\boldsymbol{c}}^{(j)}_{\mathrm{f}} = \left(\bar{\boldsymbol{T}}^{M_j}\right)^T \hat{\boldsymbol{c}}^{(j)}_{\mathrm{f}}, \tag{B.15}$$

which is the CFPE counterpart of the formula (A.48).

B.3 The equation for the expansion coefficients

Next, we apply the same procedure as we did solving the expansion coefficients of the CME in Section § A.6 to solve its CFPE counterpart (B.13).

We insert the eigenfunction expansion (B.14) into the right-hand-side of (B.13), and then transform the result into a linear combination of the Hermite polynomials. For the left-hand-side of (B.14), we direct insert in the Hermite expansion (B.11). Then, multiplying $\frac{1}{\boldsymbol{n}!}\int_{\mathbb{R}^N} \mathrm{d}\boldsymbol{z} H_{\boldsymbol{n}}(\boldsymbol{z}, \frac{1}{2}\boldsymbol{I})$ to both sides of the resulting equation, and using the properties of the multivariate Hermite polynomials, we obtain, for $|\boldsymbol{n}| \leq M_j$,

$$\sum_{s=0}^{M_j} \sum_{\substack{j_1,j_2,\ldots,j_N=0 \\ j_1+j_2+\cdots+j_N=s}}^{s} \lambda_{\boldsymbol{j}} \hat{c}_{\mathrm{f},\boldsymbol{j}}^{(j)} \bar{T}_{\boldsymbol{j},\boldsymbol{n}}^{M_j} = b_{\mathrm{f},\boldsymbol{n}}^{(j)}, \tag{B.16}$$

where $b_{\mathrm{f},\boldsymbol{n}}^{(j)}$ denotes the valuation of the right-hand-side of the equation, with explicit expression given by

$$\begin{aligned} b_{\mathrm{f},\boldsymbol{n}}^{(j)} = \sum_{i=0}^{j-1} \sum_{s=0}^{M_i} \sum_{\substack{m_1,m_2,\ldots,m_N=0 \\ m_1+m_2+\cdots+m_N=s}}^{s} \tilde{c}_{\mathrm{f},\boldsymbol{m}}^{(i)} \sum_{w=0}^{\lceil (j-i)/2 \rceil} \sum_{k=1}^{\min(j-i-2(w-1),2)} \sum_{\substack{i_1,\ldots,i_N=0 \\ i_1+\cdots+i_N=k}}^{k} \sum_{\substack{j_1,\ldots,j_N=0 \\ j_1+\cdots+j_N=j-i-k-2(w-1)}}^{j-i-k-2(w-1)} \\ \times \mathscr{F}_{k,j-i-k-2(w-1),w}^{\boldsymbol{i},\boldsymbol{j}} \mathscr{G}_{\boldsymbol{n},\boldsymbol{m}}^{\boldsymbol{i},\boldsymbol{j}}, \end{aligned} \tag{B.17}$$

with $\mathscr{G}_{\boldsymbol{n},\boldsymbol{m}}^{\boldsymbol{i},\boldsymbol{j}}$ is defined in (A.53).

For a finite value of M_j, we can invert (B.16) and obtain the values for coefficients $\left\{\hat{c}_{\mathrm{f},\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\leq M_j}$ in the eigenfunction expansion (B.14) as

$$\hat{\boldsymbol{c}}_{\mathrm{f}}^{(j)} = \left(\boldsymbol{\Lambda}^{M_j}\right)^{-1} \left(\left(\bar{\boldsymbol{T}}^{M_j}\right)^T\right)^{-1} \boldsymbol{b}_{\mathrm{f}}^{(j)}, \tag{B.18}$$

where $\boldsymbol{\Lambda}^{M_j} \in \mathbb{R}^{N^{M_j} \times N^{M_j}}$ is a diagonal matrix with $\{\lambda_{\boldsymbol{i}}\}_{|\boldsymbol{i}|\leq M_j}$ being its diagonal elements. The values of the coefficients $\left\{\tilde{c}_{\mathrm{f},\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\leq M_j}$ in the transformed Hermite expan-

sion (B.11) can be obtained by substituting (B.15) into (B.18), which yields

$$\tilde{\boldsymbol{c}}_{\mathrm{f}}^{(j)} = \left(\bar{\boldsymbol{T}}^{M_j}\right)^T \left(\boldsymbol{\Lambda}^{M_j}\right)^{-1} \left(\left(\bar{\boldsymbol{T}}^{M_j}\right)^T\right)^{-1} \boldsymbol{b}_{\mathrm{f}}^{(j)}. \tag{B.19}$$

Finally, we derive the coefficients $\left\{c_{\mathrm{f},\boldsymbol{i}}^{(j)}\right\}_{|\boldsymbol{i}|\leq M_j}$ in the original Hermite expansion (B.9) through substituting (B.12) into (B.19), which gives

$$\boldsymbol{c}_{\mathrm{f}}^{(j)} = \left(\dot{\boldsymbol{T}}^{M_j}\right)^T \left(\bar{\boldsymbol{T}}^{M_j}\right)^T \left(\boldsymbol{\Lambda}^{M_j}\right)^{-1} \left(\left(\bar{\boldsymbol{T}}^{M_j}\right)^T\right)^{-1} \boldsymbol{b}_{\mathrm{f}}^{(j)}. \tag{B.20}$$

Note that, the above equation for the expansion coefficients are based on the assumption that the maximum summation M_j is a finite value. This is proved in the following proposition.

Proposition B.1. *The expansion coefficients,* $\tilde{\Pi}_{\mathrm{f},j}$*, can be represented as a sum of finite number of Hermite function of the form* (B.11)*. Moreover, the summation index satisfies* $M_j = 3j$*.*

Proof. To show that there exists a finite number M_j such that the coefficients $\tilde{c}_{\mathrm{f},\boldsymbol{i}}^{(j)} = 0$ for all $|\boldsymbol{i}| > M_j$, it is suffice to show that the right-hand-side, $b_{\mathrm{f},\boldsymbol{n}}^{(j)}$ is zero for all $|\boldsymbol{n}| > M_j$. Applying Lemma A.7 to (B.17), we have $\mathscr{G}_{\boldsymbol{n},\boldsymbol{m}}^{\boldsymbol{i},\boldsymbol{j}} = 0$ for all $|\boldsymbol{n}| > M_{j-1} + 3$, and thus we set $M_j = M_{j-1} + 3$. Using the initial condition that $M_0 = 0$, we have $M_j = 3j$. □

C

Index

D

Nomenclature

Symbols

$f_j(\cdot)$	The j-th macroscopic rate function
α_j	The i-th propensity function
$\boldsymbol{A}$	Bold font letter in upper case refers to the tensor matrix
$\boldsymbol{A}_i^{(r)}$	The canonical factors of tensor matrix $\boldsymbol{A}$ in the i-th direction in rank r
$\boldsymbol{D}(\cdot)$	The diffusion matrix
$\boldsymbol{D}^+$	The partial sums of the positive terms in the diffusion matrix
$\boldsymbol{D}^-$	The partial sums of the negative terms in the diffusion matrix
$\boldsymbol{e}_j$	The j-th unit vector
$\boldsymbol{\epsilon}$	The fluctuations around the deterministic mean concentration (Chapter 2)

$\boldsymbol{\eta}_j$ The j-th eigenvector of the transformed Jacobian matrix

$\boldsymbol{H}_j$ The diagonal tensor matrix corresponding to the nodal value of the j-th propensity function

$\boldsymbol{H}_j^{(d)}$ The d-th factor matrix of tensor matrix $\boldsymbol{H}_j$

$\boldsymbol{I}$ Identity (tensor) matrix

$\boldsymbol{J}(\cdot)$ The Jacobian matrix

$\boldsymbol{J}_{\text{dist}}$ The nodal evaluation of the distance function

$\boldsymbol{J}_{\text{flux}}(\cdot)$ The probability flux

$\boldsymbol{\mu}(\cdot)$ The drift vector

$\boldsymbol{\nu}$ The stoichiometry matrix

$\boldsymbol{\nu}_j^+$ The vector of the amounts of molecules of each chemical species that are created in on instance of the j-th reaction.

$\boldsymbol{\nu}_j^-$ The vector of the amounts of molecules of each chemical species that are consumed in on instance of the j-th reaction.

$\boldsymbol{\nu}_j$ The j-th state changing vectors

$\boldsymbol{P}_n^{2n}$ The prolongation matrix

$\boldsymbol{\phi}$ The vector of mean concentrations

$\boldsymbol{Q}_d$ The tensor matrix corresponding to the upwind difference scheme

$\boldsymbol{Q}_{\hat{d}}$ The tensor matrix corresponding to the downwind difference scheme

$\boldsymbol{\sigma}(\cdot)$ The diffusion matrix of the chemical Langevin equation

$\boldsymbol{\Sigma}(\cdot)$ The covariance matrix in the LNA

$\boldsymbol{T}$ The transition matrix (from correlated variables to un-correlated variables

$\boldsymbol{v}(\cdot)$ The effective drift (or flow) vector

$\boldsymbol{x}$ Bold font letter in lower case refers to the tensor

$\boldsymbol{x}_i^{(r)}$ The canonical factors of tensor $\boldsymbol{x}$ in the i-th direction in rank r

$\boldsymbol{x}_{(k)}$ The mode k unfolding of tensor $\boldsymbol{x}$

$\check{\boldsymbol{T}}^{N^*}$ The transition matrix from the Hermite functions to the eigenfunction basis

$\chi(\cdot)$ The blending function

$\Omega_{\boldsymbol{k}}$ Parameter space

$\Omega_{\boldsymbol{x}}$ State space

Ω_ϵ The ϵ-neighbourhood of the singular boundary

$\ddot{\boldsymbol{T}}^s$ The transition matrix between the polynomials and the eigenfunctions up to the s-th coefficients

$\Delta\Pi_j(\cdot)$ The difference between the j-th expanded coefficients of the CME and the CFPE distributions

$\delta_{i,j}$ The Kronecker symbol

Ω_{f} Ellipticity domain of the CFPE

Ω_{f} The domain of definition of the CFPE

Ω_{f}^c Degenerate region of the CFPE

$\dot{\boldsymbol{T}}^{N_j}$ The transition matrix between the Hermite polynomial basis up to the j-th expanded coefficients

$\hat{\lambda}_j$ The j-th eigenvalue of the j-th eivalue of the transformed Jacobian matrix

κ The conditioning number

$\lambda_{\boldsymbol{h}}$ Principle eigenvalue of $\boldsymbol{A}_{\boldsymbol{h}}$

$\lambda_{\boldsymbol{h},t}$ Principle eigenvalue of $\boldsymbol{A}_{\boldsymbol{h},t}$

$\lambda_{\boldsymbol{h},\hat{t}}$ Principle eigenvalue of $\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$

$\mathscr{E}$ The step operator

$\mathscr{I}_d$ The domain interval in the d-th direction

$\mathscr{L}_s$ The sub-operator that operates the expanded coefficients on the order $V^{-s/2}$ of the system size expansion

$\mathscr{L}_{\mathrm{f},s}$ The sub-operator that operates the expanded coefficients on the order $V^{-s/2}$ of the system size expansion of the CFPE

$\mathscr{P}(\cdot)$ The unit-rate Poisson process

$\mathscr{S}_R$ The set of the chemical reactions

$\mathscr{S}_S$ The set of the chemical species

$\mathscr{S}_{\boldsymbol{\nu}}$ The set of the stoichiometric vectors

$\mathscr{U}_{N^*}$ The subspace of $L^2(\mathbb{R}^N)$ spanned by the N^* eigenfunctions

$\mathrm{mat}(\cdot)$ Matricisation

$\mathrm{vec}(\boldsymbol{x})$ The vectorisation of tensor $\boldsymbol{x}$

$\mathsf{diag}(\cdot)$ Diagonal matrix

$\mathscr{L}_{\mathrm{f}}$ The CFPE operator

$\mathscr{L}_{\mathrm{f}}^{k_d}$ The d-th sub-non-parametric CFPE operator

$\boldsymbol{A}_{\boldsymbol{h}}$ The finite difference approximation of the CFPE operator

$\boldsymbol{A}_{\boldsymbol{h},t}$ The finite difference approximation of the CFPE operator in tensor format

$\boldsymbol{A}_{\boldsymbol{h},\hat{t}}$ Compressed CFPE operator in the QTT format

$\mathscr{L}_{\mathrm{f}}^{\boldsymbol{k}}$ The parametric CFPE operator

$\mathscr{L}_{\mathrm{m}}$ The CME operator

$\oplus$ The direct sum

$\otimes$ Tensor product

$P_{\mathrm{cf}}(\cdot)$ The solution of the hybrid CME-CFPE equation

$P_{\mathrm{f}}(\cdot)$ The time-dependent solution of the chemical Fokker-Planck equation

$P_{\Omega}(\cdot)$ The principle eigenfunction of the CFPE operator in a bounded domain with homogeneous Dirichlet boundary condition

$\tilde{P}_{\mathrm{fp}}(\cdot)$ The solution of the concentration form of the CFPE

$\boldsymbol{p}_{\boldsymbol{h}}$ The finite difference solution of the CFPE

$\boldsymbol{p}_{\boldsymbol{h},t}$ The finite difference solution of the CFPE in tensor format

$\hat{\boldsymbol{p}}_{\boldsymbol{h},\hat{t}}$ The ε-approximation of the solution $\boldsymbol{p}_{\boldsymbol{h},\hat{t}}$

$\pi_{\mathrm{f}}(\cdot)$ The stationary solution of the chemical Fokker-Planck equation

$\Pi_{\mathrm{m},j}(\cdot)$ The expanded coefficients of the transformed CME distribution

$\Pi_{\mathrm{f},j}(\cdot)$ The expanded coefficients of the transformed CFPE distribution

$P_{\mathrm{km}}(\cdot)$ The solution of the Kramers-Moyal equation

$P_{\mathrm{m}}(\cdot)$ The time-dependent solution of the chemical master equation

$\tilde{P}_{\mathrm{km}}(\cdot)$ The solution of the concentration form of the Kramers-Moyal expansion

$\pi_{\mathrm{m}}(\cdot)$ The stationary solution of the chemical master equation

$\Pi_{\mathrm{m}}(\cdot)$ The distribution of fluctuations around the deterministic concentration

$\rho_{\boldsymbol{i}}(\cdot)$ The $\boldsymbol{i}$-th eigenfunction of the LNA operator

σ_{shift} The shift value in the inverse power method

Σ Singular boundary of the CFPE

Θ Observable

$\tilde{\alpha}_j$ The non-parametric part of the propensity function

$\times_{\ell}^{k}$ The mode-$\binom{k}{\ell}$ product

$\times_k$ The mode-k product

V System volume

$d(\cdot,\cdot)$ The distance function

$H_{\boldsymbol{i}}(\cdot,\cdot)$ The multivariate Hermite expansion

J_{dist} The distance (measure) function

J_{energy} The objective (energy) function

K The number of uncertain parameters

k_j The i-th reaction rate

$L(\cdot)$ The likelihood function

M The number of chemical reactions

M_j The maximum summation index in the j-th expanded coefficients of the transformed CFPE distribution

M_{cle} The number of Brownian drivers in the chemical Langevin equation

N The number of reacting molecular species

$N_{\boldsymbol{x}}$ The number of species with small molecular numbers

$N_{\boldsymbol{y}}$ The number of species with large molecular numbers

N_{gene} The copy number of genes

N_j The maximum summation index in the j-th expanded coefficients of the transformed CME distribution

S_i The i-th chemical species

w The nodal value of a function on a grid

$\left[\cdot\right]$ The Stirling number of the first kind

Abbreviation

ALS Alternating least squares

AMEN Alternating minimal energy method

CFPE The chemical Fokker-Planck equation

CLE The chemical Langevin equation

CME The chemical master equation

CP Canonical polyadic decomposition

DMRG Density matrix renormalisation group

GRN Gene regulation network

HOSVD The higher order singular value decomposition

HT Hierarchical Tucker decomposition

LNA Linear noise approximation

MG Multi-grid

MPO Matrix product operator

MPS Matrix product states

NRHM Next reaction hybrid method

ODE Ordinary differential equation

QSA Quasi-stationary approximation

QSSA Quasi-steady-state approximation

QTT Quantized tensor train decomposition

SDE Stochastic differential equation

SPD Symmetric positive definite

SSA Stochastic simulation algorithm

SSE System size expansion

SVD Singular value decomposition

TT Tensor train

URN Uniform random number

ZC Zero crossing